Shape your Space: A Gaussian Mixture Regularization Approach to Deterministic Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Abstract

Variational Autoencoders (VAEs) are powerful probabilistic models to learn representations of complex data distributions. One important limitation of VAEs is the strong prior assumption that latent representations learned by the model follow a simple uni-modal Gaussian distribution. Further, the variational training procedure poses considerable practical challenges. Recently proposed regularized autoencoders offer a deterministic autoencoding framework, that simplifies the original VAE objective and is significantly easier to train. Since these models only provide weak control over the learned latent distribution, they require an ex-post density estimation step to generate samples comparable to those of VAEs. In this paper, we propose a simple and end-to-end trainable deterministic autoencoding framework, that efficiently shapes the latent space of the model during training and utilizes the capacity of expressive multi-modal latent distributions. The proposed training procedure provides direct evidence if the latent distribution adequately captures complex aspects of the encoded data. We show in experiments the expressiveness and sample quality of our model in various challenging continuous and discrete domains.

1 Introduction

Variational autoencoders (VAEs) constitute one of the popular generative learning frameworks widely used for applications such as image understanding and generation, sentence modeling, and optimizing discrete data and graph-based structures [6, 19, 27, 32, 39]. The VAE framework elegantly combines autoencoders with variational inference [20]. The encoder of the model maps the input data into a lower-dimensional latent space according to a given inference model. The decoder provides a mapping from the latent space back to the original input space. Both are jointly optimized by maximizing a lower bound on the model evidence, regularizing the latent space towards a fixed prior distribution, usually a uni-modal Gaussian. By sampling from the latent space prior, we can utilize the decoder network to efficiently generate new samples from the training distribution. Due to the variational formulation, optimizing the VAE training objective poses significant practical challenges. Further, the over simplistic prior assumption often leads to an unsatisfying trade-off between the quality of reconstructed samples and the prior regularization [2]. Recent works have shown that choosing more flexible priors helps to improve the generative performance of VAEs [35].

Since the initial introduction of VAEs, various novel training objectives have been proposed. One line of work focuses on different regularization techniques derived from alternative probabilistic metrics to shape the latent space of the model during training, e.g. using the Wasserstein distance [34]. In contrast to the KL-divergence, the Wasserstein distance measure induces a metric on probability distributions. Practically, this facilitates smoother convergence even for initially non-overlapping

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

distributions. Further, it overcomes the *over-regularization effect* in VAEs. To be precise, it prevents the undesired behaviour of multiple data points being mapped to the same latent representation by the encoder. Since closed-form solutions for metrics like the Wasserstein distance can only be derived for very few prior distributions, these approaches rely on numerical approximations during training.

Recent work by Ghosh et al. proposed to reinterpret deterministic autoencoders as variational models, even when trained with a deterministic loss [10]. During training, this approach maximizes the negative log-marginal likelihood of the latent samples under a Gaussian normal distribution as a regularization in addition to minimizing the reconstruction loss. The results of the work shows that this regularization alone does not suffice to generate high quality samples using the Gaussian prior. Gosh et al. propose to use a multi-modal Gaussian mixture model (GMM) to fit arbitrary learned latent spaces. While this approach leads to good sampling efficiency and generalization if the post-hoc fit is reasonable, sampling quality can suffer significantly if the learned latent space can not be modeled well by a GMM.

In this work, we propose a deterministic training scheme for autoencoders that is applicable to expressive priors and overcomes the necessity of a post-hoc density estimation step for deterministic training. To be precise, we derive a deterministic regularization loss from the distance metric used in the non-parametric Kolmogorov-Smirnov (KS) test for equality of probability distributions. The resulting training objective can be derived in closed form for a class of expressive multi-modal prior distributions and provides a strong signal to efficiently shape the latent space of the model during training. We chose our experiments to evaluate the proposed approach in terms of sampling quality and expressiveness. In the first line of experiments, we compare the quality of newly generated and reconstructed samples from our model with those from a variety of other VAE variants. In the second line, we investigate our method's capability to model discrete and complex structured inputs such as arithmetic expressions and molecules. In these domains, VAEs have recently been proposed as a tool for dimensionality reduction in optimization. Applying our regularization scheme effectively utilizes multi-modal prior distributions in this context and significantly improves optimization performance.

2 Related Work

Since the introduction of VAEs many follow up works were proposed to overcome the practical and theoretical limitations of the framework, e.g. [2, 34, 35], and make them applicable to specific applications such as clustering [5, 31] or anomaly detection [40]. We first review some seminal examples of VAE models with different priors and probability metrics for latent regularization. Since our proposed regularization term structures the latent space to a Gaussian mixture model, we also compare it to prior work on deep clustering. Last, we discuss VAEs in the context of black-box optimization approaches such as Bayesian Optimization.

VAEs In the standard VAE, the prior distribution is commonly assumed to be a Gaussian normal distribution. This might lead to simplified representations learned by the model which are unable to represent the rich semantics in the data distribution. Following this argumentation, several methods were proposed to introduce more flexible and expressive priors to the VAE formulation. Casale et al. for instance employ a Gaussian process-based prior to account for correlations between the data samples. In [12], a Bayesian non-parametric prior is used with a hierarchical non-parametric variational autoencoder for video representation learning. Chen et al. use an auto-regressive prior to achieve improved generative performance on image datasets. Sebag et al. propose to replace the standard spherical Gaussian prior with a more general version with an arbitrary covariance matrix and learn the respective correlations by optimizing the evidence lower bound of the model. Although the proposed methods offer competitive performance, they often employ complex architectures [3] to achieve desired performance.

In another line of work, multi-modal priors were utilized in VAE models. Zong et al. propose to use a GMM prior in autoendocers for unsupervised anomaly detection. However, this requires training an additional network for estimating the parameters of the GMM. Lee et al. address unsupervised meta-learning using a GMM prior in VAEs to efficiently shape the latent space, they employ an extension of the evidence lower bound to complex variational inference schemes. Jakub et al. propose to replace the GMM prior assumption by a coupling of the posterior and prior of the model. Adversarial autoencoders [25] propose to improve the generative performance of VAEs by incorporating adversarial learning into the VAE framework and offer competitive performance in image generation. However,

this comes with increased computational complexity and training stability issues. To account for the over regularization effect of the KL divergence term in the standard VAE framework, [34] proposed to minimize the Wasserstein distance between the representations learned by the model and the target prior. Recently introduced regularized autoencoders [10], question the variational framework adopted by the VAEs and propose a deterministic approach to achieve comparable or better image generation performance. The authors propose to use negative log likelihood for regularization which is not strong enough to derive a sampling procedure from it. Our approach elegantly combines the idea of new training objectives with the extension to multi-modal priors without increasing training complexity or compromising sampling quality. We derive a strong training signal from a stronger metric than the Wasserstein distance, which can be derived in closed form for multimodal priors. This ensures reliable regularization of the latent space, improving sampling quality.

Deep Clustering Deep Clustering approaches benefit from well structured latent spaces. Thus, several methods directly employ Gaussian mixture VAEs for data encoding [5, 31] or establish a GMM-like latent space structure through k-means models in the latent space. For example [37] train an autoencoder and apply a KL-divergence loss for better k-means clustering while [7], use the autoencoder reconstruction loss along with relative cluster entropy. Similar approaches have been proposed in [7, 13, 15, 18, 38, 33]. Caron et al. iteratively group points using k-means during optimization. In the context of clustering, generative models have been considered for example in [9, 28]. While we are not considering the clustering task in this paper, we hypothesise that the proposed regularization can be beneficial in this context since it implicitly optimizes for mode assignments.

Structural VAEs and optimization High-dimensional optimization problems in structured discrete input domains are ubiquitous. VAEs have been used in this context to learn low dimensional continuous representations of high dimensional structured data like molecules or arithmetic expressions. Recent work proposes to use such representations to perform efficient optimization by running Bayesian optimization in the latent space of VAEs. In this case, prior knowledge of the structure of the latent space is crucial to allow for an efficient exploration and generation of valid samples. Yet, as discussed above, VAEs can suffer from simplistic prior assumptions in the original objective. Thus, sampling from the latent space of vanilla models can result in invalid samples, reducing the sampling efficiency of BO [14]. Kusner et al. overcome this issue if data follows a specific grammar. Lu et al. propose a VAE that directly works on parse trees from context-free grammars to represent discrete data. Yet, those only work with unimodal priors which limits the generalization capabilities. Our approach can be readily used to extend these models to better encode structural data and improve BO performance.

3 Method

In this work, we introduce a novel loss function to regularize the latent representation learned by deterministic autoencoders towards a given prior distribution. The definition of our loss builds on the non-parametric statistical Kolmogorov-Smirnov (KS) test for equality of one-dimensional probability distributions. We propose a multivariate variant of the distance measure used in the Kolmogorov-Smirnov test, that allows for gradient based optimization and can easily be applied to expressive multi-modal prior distributions. For ease of exposition, we start with introducing our regularization loss for unimodal Gaussian priors in section 3.1. In Section 3.2, we extend the formulations to expressive multi-modal Gaussian mixture models. Finally, in Section 3.4, we give an explicit way to estimate the weighting parameters of our loss.

3.1 Uni-Modal Latent Regularization

The KS test can be used to determine whether a collection of N one-dimensional samples is distributed according to a given reference distribution. It compares the cumulative distribution function (CDF) of the reference distribution with the empirical CDF $\bar{F}^{(N)}$ of the samples. It is often applied to the class of one-dimensional Gaussian distributions, which has important analytical properties.

For spherical Gaussians, the one-dimensional KS test quantifies a distance between the empirical distribution function of the data and the cumulative distribution function Φ of the univariate Gaussian as $\sup_{z\in\mathbb{R}} |\bar{F}^{(N)}(z) - \Phi(z)|$. Extending this *KS distance* to higher dimension is particularly challenging, since it requires matching *joint* CDFs[8, 11, 30]. Especially in higher dimensions this is

not feasible [24]. Alternative multi-variate normality tests, like the Mardia test [26] and the BEHP test [1] suffer from slow convergence rates.

To derive a suitable regularization loss from the KS distance, we propose to overcome this issue by taking into consideration the *marginal* CDFs and correlations in the prior distribution separately. Given d-dimensional latent samples $\mathbf{z}_1, \dots, \mathbf{z}_N$, the empirical marginal CDF in dimension j is given by

$$\bar{F}_j^{(N)}(z) = \frac{1}{n} \sum_{n=1}^{N} \mathbb{1}_{\mathbf{z}_{n,j} \leq z}. \tag{1}$$

We aim to regularize the latent space of our models by comparing the empirical marginal CDFs with the one-dimensional CDFs of the marginal distributions of the prior. To strengthen the training signal of our regularization scheme and make it suitable for gradient-based optimization, we replace the supremum in the original KS distance by a smoother MSE loss, that compares the distances between those functions at latent representations. For a uni-modal Gaussian prior with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, this results in

$$\mathcal{L}_{\mathrm{KS}}(\mathbf{z}_1, \dots, \mathbf{z}_N) = \frac{1}{d} \sum_{j=1}^{d} \mathrm{MSE}\left(\bar{F}_j^{(N)}(\mathbf{z}_j), \Phi(\bar{\mathbf{z}}_j)\right), \qquad \bar{\mathbf{z}}_j = \frac{\mathbf{z}_j - \boldsymbol{\mu}_j}{[\boldsymbol{\Sigma}]_{j,j}}. \tag{2}$$

Here, $\bar{F}_j^{(N)}(\mathbf{z}_j)$ denotes the vector with entries $\bar{F}_j^{(N)}(\mathbf{z}_{ij})$ and $\Phi(\bar{\mathbf{z}}_j)$ is defined accordingly. This loss will be minimized, if the empirical marginal CDFs of the latent samples match those of the multi-modal Gaussian prior. Using the above loss alone will not account for correlations between different dimensions in latent space. In the case of a spherical Gaussian prior with identity covariance matrix for example, samples with perfectly correlated Gaussian components $\mathbf{z}_{ij} = \mathbf{z}_{i'j}$, will also minimize this objective, see Figure 1. To overcome this problem, we equip our loss with an additional term, that matches covariances between different latent distributions explicitly. Along these lines, the second loss term is defined to be

$$\mathcal{L}_{\mathrm{CV}}(\mathbf{z}_1, \dots, \mathbf{z}_N) = \frac{1}{d^2} \sum_{l,j=1}^{d} \left([\bar{\boldsymbol{\Sigma}}]_{l,j} - [\boldsymbol{\Sigma}]_{l,j}\right)^2. \tag{3}$$

where $\bar{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of the latent representations. Compared to the negative log marginal regularization proposed in [10], our loss will actually enforce the latent representations to be spread across the entire support of the Gaussian prior.

3.2 Multi-Modal Latent Regularization

One strong advantage of our approach is that it can be readily applied to expressive multi-modal prior distributions. While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modelling real data sets. In contrast, a linear combination of Gaussians can give rise to very complex densities while still allowing for closed form computations of important quantities, like CDFs and covariances. A d-dimensional K-modal Gaussian mixture model is a weighted superposition of K Gaussian distributions in $\mathbb{R}^d$, that are often referred to as the modes of the model. For $k \leq K$, let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ be the mean and covariance matrix of the k-th mode in the model. Further, let $p_k > 0$ be the weight of the k-th mode. Then, the marginal CDFs of a GMM model can be computed from the CDFs of univariate Gaussians as follows

$$F_{\mathrm{GMM},j}(z) = \sum_{k=1}^{K} p_k \Phi\left(\frac{z - \boldsymbol{\mu}_{k,j}}{[\boldsymbol{\Sigma}]_{j,j}}\right), \tag{4}$$

i.e. the marginal CDFs in the GMM are weighted sums of CDFs of one-dimensional Gaussians.

$$\boldsymbol{\Sigma}_{\mathrm{GMM}} = \sum_{k=1}^{K} p_k \boldsymbol{\Sigma}_k + \sum_{k=1}^{K} p_k \left(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}\right)^T, \qquad \bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{k=1}^{K} \mu_k. \tag{5}$$

Extending our proposed regularization scheme to multimodal GMMs is straight forward. Our first loss term is defined as

$$\mathcal{L}_{\mathrm{KS},K}(\mathbf{z}_1, \dots, \mathbf{z}_N) = \frac{1}{d} \sum_{j=1}^{d} \mathrm{MSE}\left(\bar{F}_j^{(N)}(\mathbf{z}_j), F_{\mathrm{GMM},j}(\mathbf{z}_j\right). \tag{6}$$

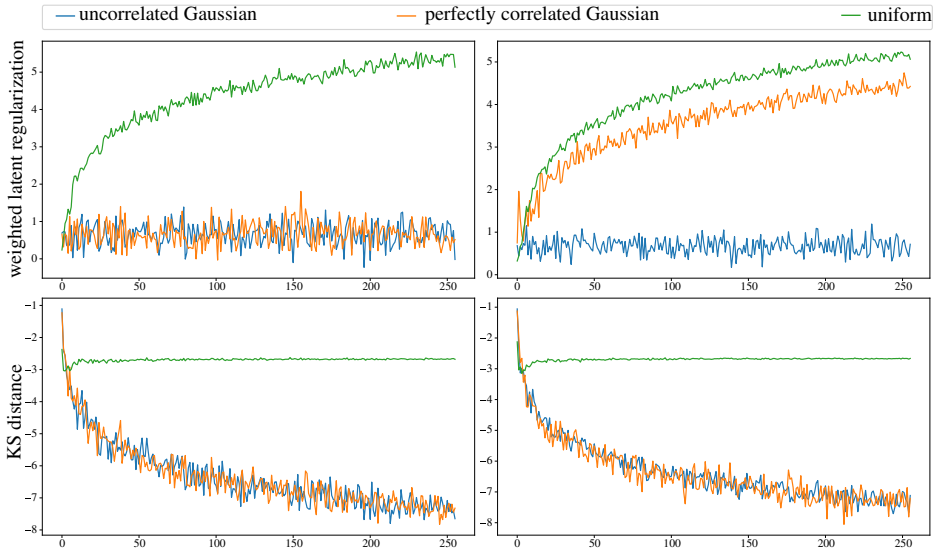


Figure 1: Uni-modal latent regularization in one and two dimensions for varying numbers of samples (x-axis) from different distributions: In two dimensions (right), the simplistic KS distance can not differentiate the target prior (blue) from other probability distributions. By contrast, our proposed regularization scheme successfully matches correlations across different dimensions.

Similarly, the second loss term is defined to be

$$\mathcal{L}_{\mathrm{CV},K}(\mathbf{z}_N, \dots, \mathbf{z}_N) = \frac{1}{d^2} \sum_{l,j=1}^{d} \left([\bar{\mathbf{\Sigma}}]_{lj} - [\mathbf{\Sigma}_{\mathrm{GMM}}]_{l,j} \right)^2 . \tag{7}$$

3.3 Latent Regularization for Autoencoders

Our loss comprises of two terms: a reconstruction loss and an explicit regularization loss, that enforces the latent representations of the encoded data to match a predefined multi-modal prior distribution. The reconstruction loss $\mathcal{L}_{\mathrm{REC}}(\mathbf{x}'_1, \dots, \mathbf{x}'_N)$ of our model equals the mean squared error between inputs $\mathbf{x}_i$ and their respective reconstructions $\mathbf{x}'_i$. Given positive weights λ_{KS} and λ_{CV}, our final loss is given by

$$\mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \lambda_{\mathrm{REC}} \mathcal{L}_{\mathrm{REC}}(\mathbf{x}'_1, \dots, \mathbf{x}'_N) + \lambda_{\mathrm{KS}} \mathcal{L}_{\mathrm{KS},K}(\mathbf{z}_1, \dots, \mathbf{z}_N) + \lambda_{\mathrm{CV}} \mathcal{L}_{\mathrm{CV},K}(\mathbf{z}_1, \dots, \mathbf{z}_N). \tag{8}$$

Formally, the weights $\lambda_{\mathrm{KS}}, \lambda_{\mathrm{CV}}$ and λ_{REC} are hyperparameters of the model. Nevertheless, we propose an explicit way to set λ_{KS} and λ_{CV} and a simple heuristic to estimate λ_{REC} to avoid necessity of extensive optimization of these weights.

3.4 Loss Weight Estimation

A key challenge is to weight two regularization losses appropriately because in general they can live on a very different scale. Especially if modes of the GMM prior are far spread, the covariance $\mathcal{L}_{\mathrm{CV},K}$ loss will dominate the marginal CDF $\mathcal{L}_{\mathrm{KS},K}$ loss by far. Nevertheless, given a target GMM prior, the dimension of the latent space and the batch size n used during training, there is a concise way to fix those hyperparameters beforehand. To be precise, for $m = 1, \dots, M$ samples $\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_n^{(m)}$ from the prior GMM, we propose to set

$$\lambda_{\mathrm{KS}}^{-1} = \frac{1}{M} \mathcal{L}_{\mathrm{KS}} \left(\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_N^{(m)} \right), \qquad \lambda_{\mathrm{KS}}^{-1} = \frac{1}{M} \mathcal{L}_{\mathrm{CV}} \left(\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_N^{(m)} \right). \tag{9}$$

Formally, we can not overcome the necessity of tuning the weight of the reconstruction loss, which has significant impact on performance of the model. Nevertheless, a reasonable approximation to it

can be obtained by training a autoencoder model and using the inverse of the best obtained loss for λ_{REC}. Using this scaling, all loss terms in our regularization loss will ultimately converge to one if the target prior is matched successfully.

4 Experiments

The objective of our experiments is to investigate the potential of the proposed model when compared to other VAE variants in generating new samples, analyse the effect of the defined prior to effectively cluster the latent space and to shape the latent space efficiently in highly structured domains such as discrete spaces. We provide all the experimental settings and hyperparameters used in the Appendix. All experiments were run on a GPU cluster, with single GPU per individual experiments. Since the cluster is part of a carbon-neutral framework, these experiments did not contribute to climate change.

4.1 Image generation

We consider four dataset, MNIST [22], FASHIONMNIST [36], SVHN [29] and CELEBA [23] to evaluate the proposed method in image generation experiments. The qualitative analysis of the generated samples for MNIST, SVHN amd CELEBA images are shown in Figure 2 along with the reconstructed samples and interpolated samples in the latent space of the trained model. In order to access the quantitative performance of the quality of the generated images we evaluate the Fréchet Inception Distance (FID) [16] for each dataset and is tabulated in Table 1. For baseline comparison we evaluate the following models, vanilla variational autoencoder (VAE [20]), Wasserstein autoencoder (WAE) [34] with MMD loss, 2stage VAEs (2s-VAE) [4], constant variance-VAE (CV-VAE) [10] and regularized autoencoders (RAEs) [10]. We consider the following evaluation metrics: 1. Sampling FID (Samp.)- score of the generated random samples (please note that this is evaluated by generating random samples from the prior distribution of the respective models/by fitting a Gaussian distribution to models trained without any prior assumptions), 2. reconstruction FID (Rec.)- measured by computing FID between the test samples and their corresponding reconstruction by the model and 3. interpolation FID (Inter.) - between the interpolated samples in the latent space and test samples. As pointed out by [10], fitting an ex-post density estimator, $q_\delta(z)$ on the learned embedding after training of VAEs further improves the generation quality. Hence we also report the FID values by fitting a GMM on top of the learned latent space of the trained model (GMM column in Table 1, not evaluated for 2s-VAE since they perform similar kind of ex-post density estimation using another VAE).

As shown in Table 1, our proposed method achieves better FIDs (Samp.) on all datasets considered, when compared to WAE, CV-VAE, 2s-VAE and RAE sampled by fitting a single Gaussian in the latent space. We also improved generation quality as argued above by fitting a mixture of Gaussians in the latent space and achieves better FIDs in MNIST, FASHION MNIST and CELEBA images, whereas for SVHN, WAEs achieved the overall best score. It is also important to note that the proposed method performs comparably or even better without employing the ex-post density estimation. The proposed method also achieves better reconstruction quality than the other VAEs except for SVHN images where RAEs performs better score. The interpolation FID indicates the overall structure of the learned latent space and the obtained FID values shows that the proposed method shapes the latent space better than the other approaches except for the CELEBA images where RAEs performs slightly better than our method. For fair comparison we use the same architecture and experimental settings in all the baseline evaluations. Please refer to the appendix for more details on the experimental settings.

4.2 Unsupervised Image Clustering

In this section we evaluate the performance of our method to naturally cluster the data points in the learned latent space in two dataset, MNIST and FASHION-MNIST. The Gaussian mixture model prior with k components in our method could be considered as k different classes/clusters to which the data points are mapped by the encoder. We train the model with latent space dimension of 10 for both dataset and visualize the random samples generated from each Gaussian component of our prior as shown in Figure 3. From the figure it can be seen that visually similar images are clustered in to same clusters indicating that natural clustering happens with the multi-modal GMM prior in our method.

Figure 2: Qualitative analysis on image generation across datasets, MNIST, SVHN and CELEBA in terms of FID (lower is better). Row 1 shows the randomly generated samples; row 2 shows the reconstructed samples by the decoder on test dataset after training, first row in each sections corresponds to the ground truth and the second one its corresponding reconstruction; row 3 shows randomly interpolated samples in the learned latent space of our model.

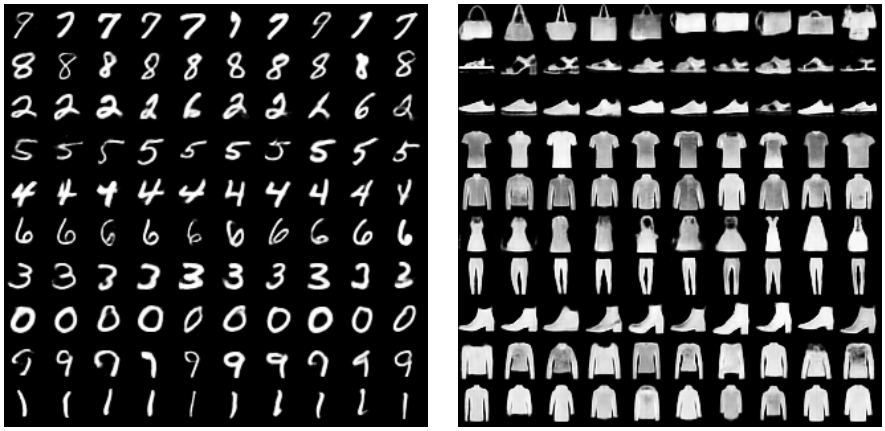
4.3 Modelling Discrete Data Structures

In this section we investigate the ability of our model to generate complex discrete data structures such as arithmetic expressions and molecules. The objective of this experiments is to analyze the model performance on shaping the latent space of such structured discrete spaces effectively. The learned latent space of the model is traversed to generate new samples with the desired properties by performing Bayesian Optimization (BO). We perform experiments in two sequence optimization problems similar to [21].

Table 1: Quantitative evaluation results across datasets. Samp. refers to the FID of the generated samples from the prior distribution or by fitting a Gaussian to the learned models trained without prior, GMM refers to the FID computed by fitting GMM on the learned model, Rec. refers to the reconstruction FID on test samples and Inter. refers to the Interpolation FID.

Dataset	MNIST				FASHION MNIST			
	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.
VAE	27.27	20.52	21.59	21.05	50.50	36.22	33.33	44.12
WAE	20.20	12.90	14.07	16.19	39.66	28.01	24.84	35.01
CV-VAE	32.12	28.62	29.61	30.76	57.57	38.28	35.10	47.73
2sVAE	26.99	–	23.77	22.13	46.47	–	31.93	41.06
RAE	17.72	14.15	14.69	15.57	47.26	29.59	24.54	34.77
Ours	**13.11**	**12.82**	**8.99**	**12.82**	**33.70**	**26.62**	**19.56**	**29.17**

Dataset	SVHN				CELEBA			
	Samp.	GMM	Rec.	Inter.	Samp.	GMM	Rec.	Inter.
VAE	61.01	58.23	59.13	50.29	68.01	61.63	52.55	58.39
WAE	58.08	**34.87**	29.62	27.16	58.91	49.17	41.14	47.08
CV-VAE	51.01	54.19	48.53	47.65	57.61	52.72	45.32	50.87
2sVAE	45.84	–	44.27	40.23	53.12	–	44.78	47.64
RAE	42.35	35.12	**31.04**	27.30	52.33	48.23	41.61	**46.58**
Ours	**37.42**	36.46	31.27	**24.87**	**49.79**	**44.79**	**39.48**	47.13



MNIST FASHIOMNIST

Figure 3: Clustering performance on MNIST and FASHION-MNIST images with a 10 component GMM prior. Each row in the figure shows randomly generated images from different Gaussian components of the GMM prior. Similar looking images are mapped into same clusters.

Arithmetic Expression Given a dataset of 50,000 univariate (functions of x) arithmetic expressions following a formal grammar [21], the task is to find the expression that best fits a target dataset. This is done by optimizing the log(1+MSE) (lower the better) values of the generated expression with respect to the target points. The target data points are chosen similar to that of [21].

Chemical Design Given a dataset of drug molecules, the ZINC250k molecule dataset [17], the objective is to generate new drug like molecules. The drug likeliness of a molecule is quantified by the water octanol partition co-efficent (higher the better).

Results We extend the architecture and experimental settings of [21] to include our proposed losses during training. For baseline comparison, we consider Grammar VAE (GVAE) [21], Character VAE (CVAE) [14], Grammar constant variance VAE (GCVVAE) [10] and Grammar based RAE (GRAE) [10] frameworks. The three best scores found by our method for arithmetic expressions and the molecule experiments are reported in Table 2. Our model performs comparatively better than the considered baselines and achieves the best first score for both arithmetic and molecule experiments. In addition to the optimization performance, it is also important to consider the validity of the new samples generated by the models. A well-structured latent space should yield valid sampling following the defined grammar/rules of the used dataset. Our model achieves better validation and average scores as shown in Table 3 except for GCVVAE which achieves a better average score in the arithmetic expression task. All reported values are evaluated by averaging across 5 BO trials.

Table 2: Best scores found by each method for arithmetic expression and molecule experiments. Baseline values reported from [10].

Method	Expressions			Molecules		
	1st(↓)	2nd(↓)	3rd(↓)	1st(↑)	2nd(↑)	3rd(↑)
GVAE	0.10	0.46	0.52	3.13	3.10	2.37
CVAE	0.45	0.48	0.61	2.75	0.82	0.63
GCVVAE	0.39	0.40	0.43	3.22	2.83	2.63
GRAE	0.39	**0.39**	0.43	3.74	3.52	**3.14**
Ours	**0.03**	0.40	**0.41**	**4.15**	**3.84**	3.12

Table 3: Fraction of valid samples and their corresponding average scores for arithmetic expression and molecule experiments for each method. Baseline values reported from [10].

Method	Expressions		Molecules	
	Frac. valid (↑)	Avg. score (↓)	Frac. valid (↑)	Avg. score (↓)
GVAE	0.99 ± 0.01	3.26 ± 0.20	0.28 ± 0.04	-7.89 ± 1.90
CVAE	0.82 ± 0.07	4.74 ± 0.25	0.16 ± 0.04	-25.64 ± 6.35
GCVVAE	0.99 ± 0.01	**2.85 ± 0.08**	0.76 ± 0.06	-6.40 ± 0.80
GRAE	**1.00 ± 0.00**	3.22 ± 0.03	0.72 ± 0.09	-5.62 ± 0.71
Ours	**1.00 ± 0.00**	3.32 ± 0.04	**0.72 ± 0.03**	**-5.08 ± 1.30**

5 Discussion of Limitations

One limitation of our work is the necessity to chose the prior distribution in advance. We expect that fixing a suitable number of modes for the GMM is important to provide better sampling quality. Also, by considering marginal CDFs, we simplified the original distance metric from the KS test. While reducing computational complexity during training, this comes at the cost of an additional loss term. Further, our proposed addition of the KS distance is not suitable for matching higher order moments of the latent representations to the target prior, which at least from a conceptual point of view can lead to a mismatch of the prior. Further, our loss only facilitates matching empirical marginal CDFs of latent representations to the marginal CDFs of the prior evaluated at latent vector. As a consequence, our regularization loss might be a less stable training signal for small batch sizes in high dimensions.

6 Conclusion

Recent studies have illustrated the effectiveness of flexible expressive priors in VAEs to learn more meaningful latent representation. Following recent work that highlights the potential of deterministic alternatives to the variational formulation in VAEs, we propose a simple deterministic autoencoding framework with more powerful regularizers to accommodate for expressive multi-modal priors. In particular, we derive our novel deterministic regularization scheme from a strong metric on probability distributions. The proposed approach can be readily applied to effectively shape the latent space of existing autoencoding frameworks towards multi-modal Gaussian priors. Our experimental evaluations show that the proposed training objective yields comparable sampling quality to those of variational autoencoders and achieves better performance in modelling discrete complex data structures.

References

[1] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988.

[2] V. Berger and M. Sebag. Variational Auto-Encoder: not all failures are equal. preprint, February 2020.

[3] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *ArXiv*, abs/1611.02731, 2017.

[4] B. Dai and D. Wipf. Diagnosing and enhancing vae models. *ArXiv*, abs/1903.05789, 2019.

[5] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016.

[6] Z. Ding, Y. Xu, W. Xu, Y. Yang, M. Welling, and Z. Tu. Guided variational autoencoder for disentanglement learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7917–7926, 2020.

[7] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision*, pages 5736–5745, 2017.

[8] G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225:155–170, 1987.

[9] K. Ghasedi, X. Wang, C. Deng, and H. Huang. Balanced self-paced learning for generative adversarial clustering network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4391–4400, 2019.

[10] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. J. Black, and B. Schölkopf. From variational to deterministic autoencoders. *ArXiv*, abs/1903.12436, 2020.

[11] E. Gosset. A three-dimensional extended Kolmogorov-Smirnov test as a useful tool in astronomy. *Astronomy and Astrophysics.*, 188(1):258–264, 1987.

[12] P. Goyal, Z. Hu, X. Liang, C. Wang, E. P. Xing, and C. Mellon. Nonparametric variational auto-encoders for hierarchical representation learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5104–5112, 2017.

[13] X. Guo, X. Liu, E. Zhu, and J. Yin. Deep clustering with convolutional autoencoders. In *Neural Information Processing*, 2017.

[14] R. Gómez-Bombarelli, D. Duvenaud, J. Hernández-Lobato, J. Aguilera-Iparraguirre, T. Hirzel, R. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4, 2016.

[15] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35, March 2016.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[17] J. Irwin, T. Sterling, Michael M. Mysinger, Erin S. Bolstad, and R. Coleman. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52:1757 – 1768, 2012.

[18] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, pages 24–33, 2017.

[19] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *ICML*, 2018.

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[21] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1945–1954, 2017.

[22] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[24] R. Lopes, I. Reid, and P. Hobson. The two-dimensional kolmogorov-smirnov test. *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*, 2007.

[25] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. *ArXiv*, abs/1511.05644, 2015.

[26] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 12 1970.

[27] Y. Miao, L. Yu, and P. Blunsom. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1727–1736, 2016.

[28] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4610–4617, 2019.

[29] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 2011.

[30] J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.

[31] V. Prasad, D. Das, and B. Bhowmick. Variational clustering: Leveraging variational autoencoders for image clustering. *CoRR*, abs/2005.04613, 2020.

[32] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[33] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu. Learning deep representations for graph clustering. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

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

[35] J. M. Tomczak and M. Welling. Vae with a vampprior. In *AISTATS*, 2018.

[36] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[37] J. Xie, R.Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

[38] B. Yang, X. Fu, N. D Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3861–3870. JMLR. org, 2017.

[39] M. Zhang, S. Jiang, Z. Cui, R. Garnett, and Y. Chen. D-vae: A variational autoencoder for directed acyclic graphs. In *NeurIPS*, 2019.

[40] B. Zong, Q. Song, M. Renqiang M., W. Cheng, C. Lumezanu, D. k. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please refer to section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Since the proposed method focus on a fundamental learning problem rather than a specific application we believe that there are no direct societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Please refer to section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] Please refer to section 3
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please refer to the supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]