An Explanatory Analysis of the Geometry of Latent Variables Learned by Variational Auto-Encoders

Alexandra Peşte
Romanian Institute of Science and Technology
Max Planck Institute for Mathematics in the Sciences
peste@rist.ro

Luigi Malagò
Romanian Institute of Science and Technology
malago@rist.ro

Septimia Sârbu
Romanian Institute of Science and Technology
sarbu@rist.ro

Abstract

Variational AutoEncoders are generative models, consisting of two cascading networks: the recognition network and the generative network. Under the framework of variational inference, the original training algorithms of VAEs optimize a lower-bound on the log-likelihood, derived using the Kullback-Leibler divergence. More recent literature focused on improving the log-likelihood using alternative bounds, such as the ones derived from the Rényi divergence and their reformulations in terms of importance sampling. A thorough description of the influence of such bounds on the quality of the latent representation is lacking. Defining what makes a given latent representation better than another is not trivial. Learning adequate such descriptions represents one of the main determinants of the performance of VAEs. Representations in the latent space are reportedly distributed in a coherent way and the sub-manifold of observations appear to be mapped into an affine space. However, the explicit choice of the prior over the latent space remains the only known element in the construction of the geometry of this space. By means of an explanatory analysis, in our work-in-progress paper, we investigate the factors that shape the geometry of the latent space of VAEs. We evaluate the impact of different structural parameters of the model and that of the cost function optimized during training.

1 Introduction

Variational Autoencoders (VAEs) [7, 12] combine the advantages of neural networks with those of variational inference, to create powerful models for complex tasks, such as high-dimensional probability density estimation and image generation. The neural network in the first stage of VAE encodes the input data into a lower dimensional latent space, forming the representational part of VAE. Using the weights of the neural network optimized during the training phase, the decoder of the final stage of VAE generates the pixels of the output image, from the latent samples drawn from a zero mean, unit covariance matrix Gaussian distribution. These two neural networks are connected and their set of parameters are optimized under a Bayesian probabilistic framework. Due to the complexity of the modeling task and of the structure of VAE, the probability distributions

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

involved in the optimization process cannot be computed in closed form. As a solution, variational inference is used to approximate the intractable posterior distribution. One of the most common approach is the mean field approximation, where the posterior distribution is assumed to factorize into independent components [1], [14]. This approach imposes severe constraints on the type of the learned distribution. These authors argue that this model is not accurate enough to capture all the complexities of the input data, and, thus to generate high quality images.

In the optimization of the parameters of the neural networks of VAE, the main target is maximizing the value of the log-likelihood. Since this distribution is unknown and analytically and computationally intractable, a lower bound on the log-likelihood is optimized instead. Extensive research has been carried out in the effort to improve the estimation of this variational bound. However, in terms of extending the descriptive power of the encoder, to improve the optimization of VAE, the geometrical aspects of the latent space have not been in focus. The first improvement of the variational lower bound consists of incorporating an importance sampling scheme into the standard Kullback-Leibler variational lower bound, to obtain a tighter bound [1]. These results are refined in [11], by deriving estimators for the gradients needed to optimize VAE with such lower bounds. The authors of [9] obtain a new lower bound on the log-likelihood, by performing the derivations of the variational lower bound with the Rényi α-divergence. The above mentioned two approaches are combined to produce the Rényi importance sampling lower bound [15]. They obtained a slightly better value of the log-likelihood bound, on the MNIST benchmark dataset, than [11]. Using the χ-divergence, with several orders, the first upper bound on the log-likelihood is derived in [2]. The authors train the VAE algorithm, by optimizing the standard variational lower bound and the newly introduced upper bound simultaneously. To construct a power expectation propagation algorithm, which is a variational type of inference algorithms, Amari's α-divergence is proposed in [4], to measure the difference between the target intractable distribution and the approximating one. The variational lower bound is enriched with a regularization term, which takes into account extra information about the data [3]. The authors investigate this modification of the optimization objective on the learned latent space.

So far, the above mentioned studies have been focused on deriving tighter lower and upper bounds on the log-likelihood and creating VAE algorithms that improve its value, as a performance measure. The question of the meaning and optimization of the hyperparameters of the generalized divergences remains unresolved [9], [15], [4], [2]. Little to no connection of these generalized divergences has been made to the learning capacity of the latent model. Using the current probabilistic framework, the latent model is not powerful enough to exploit the neural networks to their maximum capacity. Thus, our goal is to develop a theoretical and computational framework, to characterize the geometry of the latent space, to learn richer representations for the input data. We aim at investigating the influence of the hyper-parameters of the generalized divergences on the improvement of the learning capacity of the encoder and compare it with the result given by the Kullback-Leibler divergence.

2 Experimental Analysis

2.1 Latent Variables and Information Content

In this section we start our analysis over the MNIST dataset [8], by evaluating the number of latent variables which carry some information content to the decoder, using a VAE trained by optimizing a Rényi variational lower-bound. In order to determine this number and the role of α, we apply a PCA decomposition to the mean vectors associated to the approximate posterior distribution over the test set, which in the rest of the analysis we fix to be a Gaussian distribution with diagonal covariance matrix. We design both the encoder, which outputs the parameters of the Gaussian distribution, and the decoder, which generates the Bernoulli parameters for each pixels of the reconstructed image, as feed-forward neural networks with two hidden layers, each containing 200 nodes. The model is trained using the Adam optimizer algorithm. We run experiments using the ReLU activation function. Each experiment is conducted for 1000 epochs and the results are averaged over 3 runs. The dimension k of the latent space is equal to 30.

We define a component to carry information about the data if its eigenvalue is larger than $\epsilon = 0.01$, which implies that, the mean value of the component does not vary among the points in the dataset. The PCA decompositions of the mean vectors, and a summary of the number of latent components carrying information about the test set, as a function of the hyper-parameter α, are represented in Fig. 1 and 2 (left), respectively. Notice that, even if this is not evident from the PCA decomposition

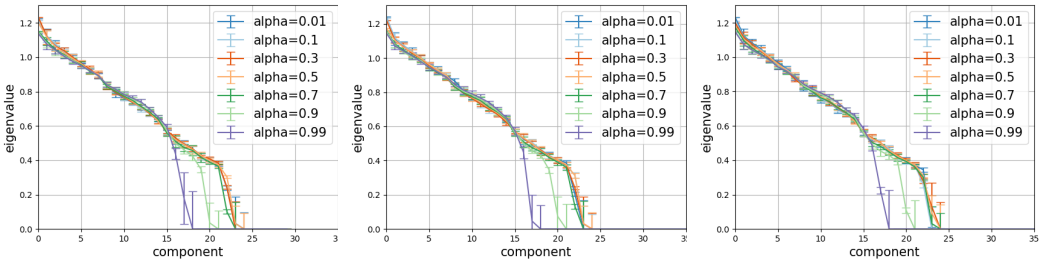

Figure 1: PCA eigenvalues for different values of α and dimensions of the latent space k, results averaged over 10 runs. The eigenvalues associated to the PCA decomposition are applied to the matrix of the approximate posterior means, for all points in the data set. For this experiment, VAE is trained on MNIST, to maximize a lower-bound derived from the Rényi divergence, using different values of α. (Left) $k = 30$; (Center) $k = 40$; (Right) $k = 50$.

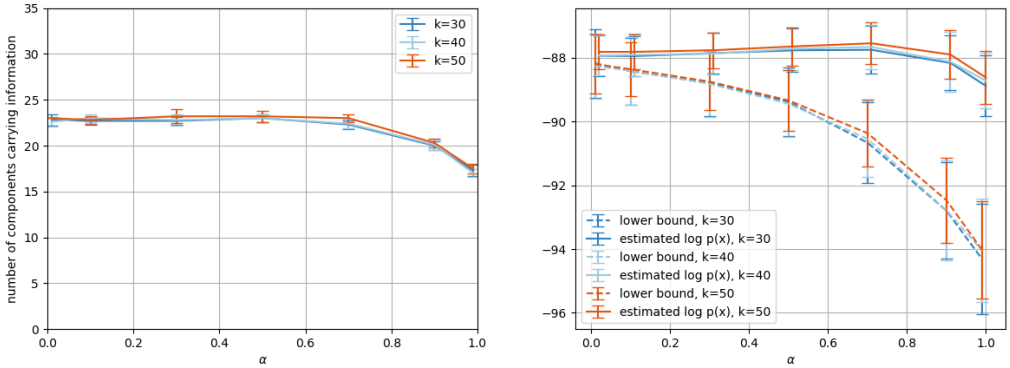

Figure 2: (Left) Average number (10 runs) of latent components in MNIST carrying information about data: VAE trained with the Rényi α-divergence, for different dimensions of the latent space k. (Right) Averaged results (10 runs) for the variational Rényi lower-bound (10 samples) and the importance sampling estimate (2000 samples) of the log-likelihood, on MNIST test data, as a function of α, for different dimensions of the latent space k.

in Fig 1, the eigenvectors tend to align with the main axis, as a consequence of the independence assumption between latent variables in the Gaussian distribution. Finally, in Fig. 2 (right) we evaluate the impact of the α parameter on the quality of the variational lower-bound and the estimated log-likelihood using importance sampling, to correlate performance with the number of latent component carrying information about the data. Our analysis show that as α goes to 0, the number of learned components saturates, with different values which depends on the activation function, as well as the lower-bound gets closer to the true value of the log-likelihood. However, the importance sampling estimation of the log-likelihood becomes more noisy, and thus a larger number of samples is required to obtain reliable estimates. Other experiments we don't show here confirm that the number of components remains constant even for $k \in \{40, 50\}$.

2.2 Linear Separability in the Latent Space

In this section we study the geometry of the latent space learned by VAE during training, though the evaluation of the performance of independent logistic regression linear classifiers trained at different stages, similarly to what has been done for instance in [10]. The purpose of the analysis is to determine the impact of the size of the latent space over the classification accuracy, and possible advantages in performing the classification by adding among the features not only the entries of the mean vector, but also the diagonal of the covariance matrix.

Results appear in Fig. 3. Previous work on MNIST [8] showed how linear classifiers trained on the input space obtained accuracies over the test set in the range 0.88 and 0.92. Our experiments shows performance in range 0.93-0.94, which supports the fact that VAEs are capable of learning a compact representation in the latent space by performing an automatic feature extraction, which can be employed efficiently in classification using linear methods. What appears to be interesting from our experiments, rather than the marginally but consistent better performance associated to a larger number of variables in the latent space, is the improvement in accuracy given by adding the information from the covariance matrix in classification. This supports the intuition that the covariance matrix encodes information about the similarity of the images itself with neighborhood images in the latent space, which is relevant feature in classification. The impact of the use of Rényi divergence for different α in this setting appears to be less relevant.

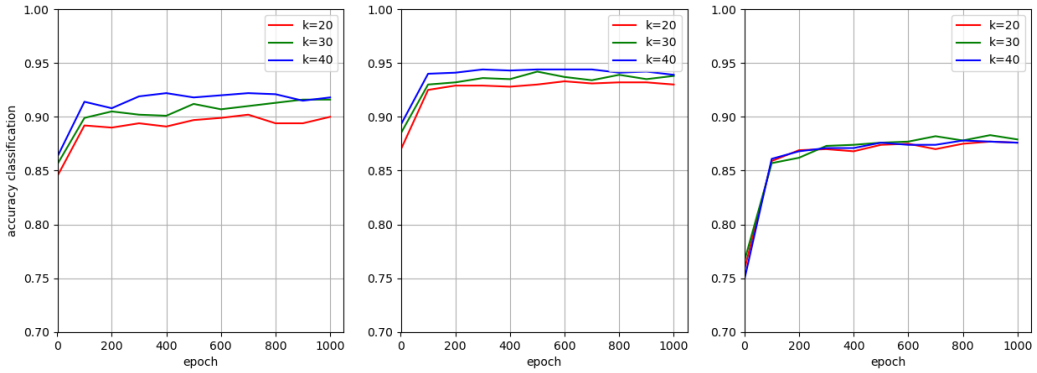

Figure 3: Classification performance over the test set of a logistic regression classifier trained at different epochs during the training of VAE for MNIST. The classification features correspond to mean vectors (left); mean vectors and diagonal entries of the covariance matrix (center); sampled latent variables (right). VAE trained with KL-divergence for different numbers of latent variables k.

2.3 Image Transitions through Geodesics

The recognition network of a VAE maps points in the dataset to the parameters of a probability distribution over the space of the latent variables. In the case of the original VAE algorithm [7, 12] the generative network implements a mapping between the input space and the family of Gaussian distributions with diagonal covariance matrix, from which the latent variables are then sampled. The representations of similar data points are mapped close to each other in the latent space. More in general the space of the latent variables seems to be characterized by an affine geometry. Indeed, in case of images, the convex hull between two latent representations, once decoded, gives rise to a smooth deformation of one image into the other, cf. Fig. 4 in Appendix A from [7], where this behavior has been visualized for the MNIST [8] and Frey Face [5] datasets.

In order to investigate the geometry of the latent space, instead of computing the convex hull between representations, one alternative approach consists in computing the shortest path between two Gaussian distributions parameterized by mean vectors and covariance matrix, by explicitly taking into account the geometry of this manifold. Indeed, we known from Information Geometry [6] that the manifold of Gaussian distributions, and more in general statistical models from the exponential family, admits a Riemannian geometry given by the Fisher-Rao metric tensor [13]. By taking into account such geometry, the shortest path between two Gaussian distributions corresponds to the geodesic curve connecting the two points on the manifold. Differently from the Euclidean geometry, where distances are computed by the convex hull between the two points, along the Fisher-Rao geodesic we have an increase of variance, as represented in Fig. 4, which translates in Fig. 5 (forth row) to a more variability in images sampled along the curve. The approach we propose here to compute a smooth transition between two given images, consists in moving along the geodesic and decode images without sampling. Due to the shape of geodesics, for images representing different digits, characterized by similar covariances and equally distant from the hyperplane separating classes, this results in a slightly sharper transition from one image to the other, due to the difference between performing uniform steps along the geodesic and along the convex hull between mean vectors.

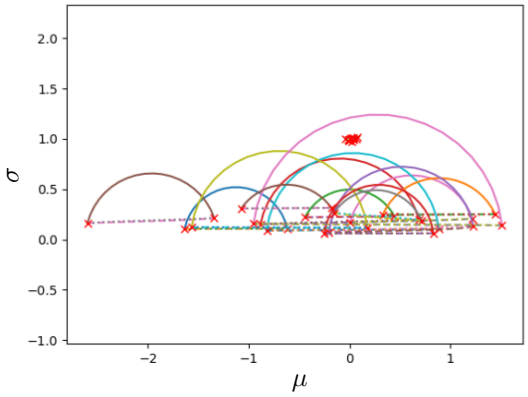

Figure 4: Geodesic curves for Gaussian distributions in the (μ, σ) space (solid line) vs convex hulls (dotted lines).

Figure 5: Different interpolations between two MNIST images, represented in the first and in the last column: convex hull in the input space (first row); convex hull between mean vectors, with no sampling (second row); convex hull in the space of the latent variables after sampling of the two images (third row); computation of the geodesics between two Gaussian distributions with sampling along the curve (forth raw) and without sampling (fifth row). In red we highlighted the corresponding transition images for the Euclidean and the Riemannian geodetic. The latter geodetic shows a slightly sharper transition.

3 Conclusions

In this work-in-progress paper we have performed a preliminary experimental analysis of the nature and geometry of the latent space generated by VAE over, the MNIST dataset, trained with different variational bounds. The purpose of this analysis is twofold, indeed a better understanding of the geometry of latent space can contribute to shed lights on the choice of which statistical models to use over the latent variables, as well as the design of optimization algorithms which explicitly take into account the geometry of the space itself.

References

[1] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

[2] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. M. Blei. The χ-divergence for approximate inference, published in NIPS 2017 under the title "Variational inference via χ upper bound minimization". In *Neural Information Processing Systems (NIPS 2017)*, 2017.

[3] G. Hadjeres, F. Nielsen, and F. Pachet. GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures. In *arXiv:1707.04588*, 2017.

[4] J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner. Black-box α-divergence minimization. In 33^{rd} *International Conference on Machine Learning (ICML 2016)*, 2016.

[5] http://www.cs.nyu.edu/ roweis/data.html.

[6] S. ichi Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191. Oxford University Press, 1993.

[7] D. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR 2014)*, 2014.

[8] Y. LeCun. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, issue 11, 1998.

[9] Y. Li and R. E. Turner. Rényi divergence variational inference. In *Neural Information Processing Systems (NIPS 2016)*, 2016.

[10] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. In *arXiv:1511.05644*, 2016.

[11] A. Mnih and D. J. Rezende. Variational inference for Monte Carlo objectives. In 33^{rd} *International Conference on Machine Learning (ICML 2016)*, 2016.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *31^{st} International conference on machine learning (ICML 2014)*, 2014.

[13] L. T. Skovgaard. A Riemannain geometry of the multivariate normal model. Technical Report 167, Department of Statistics, Stanford University, 1981.

[14] C. K. Sønderby, T. Raiko, L. Maale, S. K. Snderby, and O. Winther. Ladder variational autoencoders. In *Neural Information Processing Systems (NIPS 2016)*, 2016.

[15] S. Webb and Y. W. Teh. A tighter Monte Carlo objective with Rényi α-divergence. In *Bayesian deep learning workshop at NIPS 2016*, 2016.