MATH5472 Final Project: Don't Blame the ELBO! A Linear Perspective on Posterior Collapse

LI, Hongru 20837693

December 2022

1 Introduction and Preliminaries

1.1 Posterior Collapse

In the Bayesian Inference Model, such as **Variational Autoencoder (VAE)** [Kingma and Welling (2013)], Bayesian Network etc., in addition to preventing the model from gradient vanishing and exploding, we should also pay attention to the **Posterior Collapse** problem. The Posterior Collapse problem in VAE model is always caused by the vanishing of KL divergence (KL-Vanishing), and this problem is also called KL-vanishing.

In VAE model , the posterior distribution $p(z|x)$ is always intractable, so we also introduce a parameterized neural network $q_\theta(z|x)$ to approximate the true posterior distribution, the optimized parameter is obtained by maximizing the **Evidence Lower Bound (ELBO)**:

$$\begin{aligned}\log p(x) &= \log \int_z p(x)q(z \mid x)dz = \mathbb{E}_{q(z|x)}[\log p(x)] \\ &\geq \underbrace{\mathbb{E}_{q(z|x)}[\log p(x,z) - \log q(z \mid x)]}_{\text{ELBO}} \\ &= \underbrace{\mathbb{E}_{q(z|x)}[\log p(x \mid z)]}_{\text{Reconstruct term } L_{\text{Rec}}} - \underbrace{D_{KL}(q(z \mid x)\|p(z))}_{\text{KL term } L_{KL}}\end{aligned} \tag{1}$$

In the training process of VAE, the KL term tends to be zero, and then the decoder of the VAE will ignore the posterior distribution $q_\theta(z|x)$ and only samples from the prior Gaussian distribution $\mathcal{N}(0, I)$, so that VAE The network fails. A lot of literature have announced the this may be caused by the strong constrict of the KL term, i.e., $L_{KL} = 0$, which makes the poterior distribution $q(z|x)$ deteriorate to the prior Gaussian distribution $p(z)$, thus the KL divergence vanishes and cannot provide information for the decoder. Another perspective is that the high-capacity decoders are blamed for the posterior collapse, i.e., the network does not need the posterior distribution and can generate $\hat{x}$ by sampling from the noise $\mathcal{N}(0, I)$. At this time, the posterior is not needed, and the VAE network is also invalid.

1.2 Linear AE, Linear VAE, PCA, pPCA

Principal Component Analysis (PCA) and **AutoEncoder (AE)** are two representative methods in unsupervised learning. PCA and Autoencoder are both non-probability

method, and they both have a corresponding probability form called **probability PCA (pPCA)** and **Variational Autoencoders (VAE)**. The method can be summarized in Table 1, 2.

Table 1: PCA and Autoencoder

	PCA	Autoencoder
Encoder	$z_i = W^T(x_i + b) \ \in \mathcal{R}^k$	$z_i = \sigma(W^T(x_i + b)) \ \in \mathcal{R}^k$
Decoder	$\hat{x}_i = Wz_i - b \ \in \mathcal{R}^p$	$\hat{x}_i = \sigma(\hat{W}^T(x_i + \hat{b})) \ \in \mathcal{R}^k$
Reconstruction Error	$\sum_{i=1}^n \|\|x_i - \hat{x}_i\|\|_p^p$	$\sum_{i=1}^n \|\|x_i - \hat{x}_i\|\|_p^p$

Table 2: pPCA and VAE

	pPCA	Variational Autoencoder
Latent Marginal Dist	$p(z) = \mathcal{N}(0, I)$	$p(z) = \mathcal{N}(0, I)$
Observation Conditional Dist	$p_\theta(x\|z) = N(x\|f(z;\theta), \sigma^2 I)$	$p_\theta(x\|z) = N(x\|f(z;\theta), \sigma^2 I)$
Deterministic Function	$f(z;\theta) = Wz + \mu$	$f(z;\theta) = \sigma(Wz + \mu)$
Reconstruction	$\hat{x} = Wz + \mu + \epsilon, \epsilon \sim N(0, \sigma^2 I)$	$\hat{x} = f(z;\theta) + \epsilon, \epsilon \sim N(0, \sigma^2 I)$

From Table 1, we can find that PCA actually equals to **Linear Autoencoder**, i.e., when σ is a affine transformation. We always get the global optimal analytic solution when we deal with PCA while we usually use back-probpagation to get local optimal numerical solution when we deal with Linear AE.

From Tabel 2, we can find the difference between pPCA and VAE is whether $f(\cdot)$ is a affine transformation, but this differece make a great impact on the solution of the two methods. In pPCA, all the four distributions we deal with $(p(x), p(z), p(z|x), p(x|z))$ are Gaussian Distribution. But in VAE, we only hold this when σ is a affine transformation.

2 Overview of the Paper Lucas et al. (2019)

2.1 Main Contribution

In this paper, the authors show that the posterior collapse may be caused by the log marginal likelihood $\log p(x)$ itself rather than the KL term in ELBO or a high-capacity decoder which is the traditional perspective blame by a series experienments on pPCA and the linear VAE. And they also show that the ELBO objective for the linear VAE dose not introduce any additional local maxima relative to log marginal likelihood $\log p(x)$. Their contributions can be summarized as following:

- They prove that the ELBO objective optimized by linear VAE with diagonal covariance can exactly recover the global optimum of log-likelihood gained by pPCA.
- They prove that using ELBO to train linear VAE dose not introduce any spurious local maxima relative to the log-likelihood itself.
- They find that the observation noise greatly affects the degree of posterior collapse of Deep Gaussian VAE model.

- They give a formal definition for the posterior collapse which makes it possible to Quantitatively test the posterior collapse phenonmenon.

2.2 System model, Rough Analysis and Main Conclusion

Probability PCA and Stationary Point Analysis The Probability PCA model is a full Gaussian model, so we are able to hold the closed form for all the marginal distribution and the conditional distribution as following:

$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I}) \tag{2}$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) \tag{3}$$

$$p(\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right) \tag{4}$$

$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right) \tag{5}$$

$$\mathbf{M} = \mathbf{W}^W\mathbf{W} + \sigma^2\mathbf{I} \tag{6}$$

The MLE solution for W and σ^2 is as following:

$$\sigma^2_{\text{MLE}} = \frac{1}{n-k}\sum_{j=k+1}^{n}\lambda_j \tag{7}$$

$$\mathbf{W}_{\text{MLE}} = \mathbf{U}_k\left(\boldsymbol{\Lambda}_k - \sigma^2_{\text{MLE}}\mathbf{I}\right)^{1/2}\mathbf{R} \tag{8}$$

where the column of $\mathbf{U}_k$ are the subset of the eigenvector of the data covariance matrix $\mathbf{S}$, $\boldsymbol{\Lambda}_k$ is a diagonal matrix composed of the corresponding eigenvalue, σ_{MLE} is corresponding average of the missing eigenvalue. In Tipping and Bishop (1999), the authors prove that when the subset of the eigenvector is selected as the eigenvectors corresponding to the first largest eigenvalues, the logarithmic likelihood function can reach the maximum value, and all other solutions are saddle points. And also when σ^2 is fixed, $\mathbf{W}_{MLE}$ is still a stationary point, but the larger the σ^2 is, the more stable the stationary $\mathbf{W}_{MLE}$ is.

Linear VAE The Linear VAE model formed in the paper is as following:

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{z}) &= \mathcal{N}\left(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right) \\ q(\mathbf{z} \mid \mathbf{x}) &= \mathcal{N}(\mathbf{V}(\mathbf{x} - \boldsymbol{\mu}), \mathbf{D}) \end{aligned} \tag{9}$$

Here $\mathbf{D}$ is a diagonal covariance matrix, and the global optimum of pPCA will be recovered when $\mathbf{R} = \mathbf{I}$, in this case, the conrresponding distribution will have the same form, so the maxima of ELBO equals to the maxima of loglikelihood of pPCA, and at the maxima, we have $W_{VAE} = W_{pPCA}$.

Existing literatures always blame the KL divergence for the posterior collapse, but here the authors prove that the ELBO objective for a linear VAE dose not introduce any

addtional local maxima relative to the log-likelihood.(Actually, I just understand a little of the proof of Theorem 1).

Formal Definition of Posterior Collapse Here the authors gives a formal definition for the posterior collapse – if $\mathbb{P}_{\mathbf{x}\sim p}\left[KL\left(q\left(z_i \mid \mathbf{x}\right) \| p\left(z_i\right)\right) < \epsilon\right] \geq 1-\delta$ holds, we say that dimension i is (ϵ, δ)-collapse.

3 Reproduction of the Main Result

In this section, we will reproduce the two part of the main result shown in the original paper, and the code is avaliable in https://github.com/hhhhhscott/MATH5472Final. Additionally, to reproduce the same result shown here, please just set the random seed for each python package as $seed = 0$.

3.1 Linear VAEs

Instead of using 1000 randomly chosen MNIST images, in the following simulation we will use the first 1000 images in MNIST dataset just for making it easy to test whether the loglikelihood or ELBO is calculated correctly.

Linear VAE and pPCA The first simulation is to verify the relationship of pPCA and Linear VAE. We set the latent dimension as 200, and use the MLE solution for σ^2_{MLE} and W_{MLE} in pPCA to calculate the global optimal likelihood. Then we use the same latent dimension setting to maximize the ELBO objective function of Linear VAE. The result is shown in Figure 1, and this prove that Linear VAE can recover the global optimum of pPCA. Additionally, we check the decoder weight W gained in Linear VAE and MLE solution W_{MLE} gained in pPCA. As shown in Figure 2 and 3, we found that although linear VAE recover the global optimum of pPCA, but the weight of the decoder W_{VAE} did not recover that gain in the pPCA MLE solution. In Figure 3, the maximum of the gram matrix $W^T_{pPCA}W_{VAE}$ did not appear in the first column and also the "pseudo" eigenvalue donot seem to be in order(Although I have sorted the column of W_{VAE} by the order of the true eigenvalue).

Effect of Stochastic ELBO Estimation The second simulation is about the effect of the Stochastic ELBO Estimation("Figure 3" in the original paper). We use the same experiment setting as that in the original paper. The result is shown in Figure 4, we found the high-variance noise caused by reparameterization and Monte Carlo sampling claimed in the original paper make almost no sense to global optimization. The convergence speed of both analytic ELBO and stochastic ELBO seems to be same, and they both recover the exact loglikelihood.

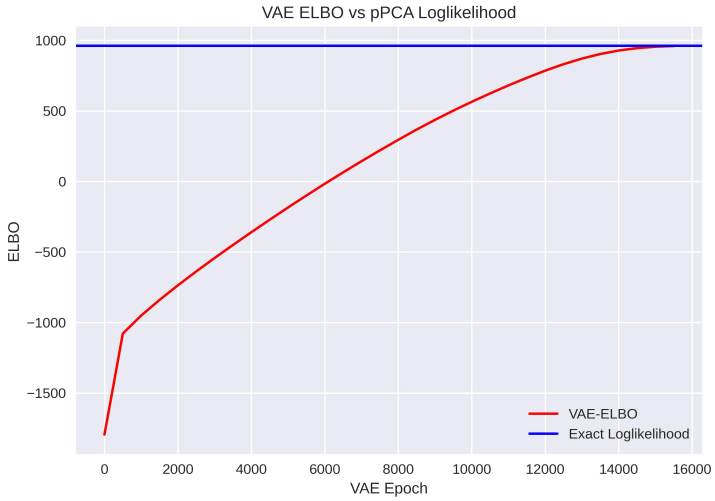

Figure 1: The ELBO of linear VAE vs the exact loglikelihood from pPCA

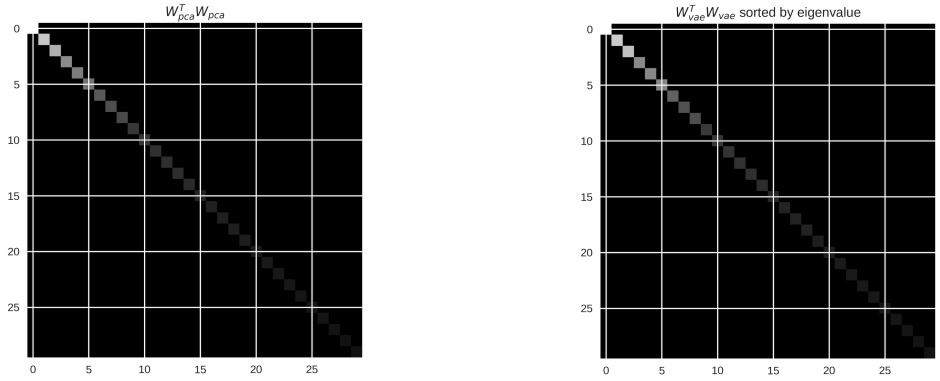

Figure 2: The gram matrix $W_{pPCA}^T W_{pPCA}, W_{VAE}^T W_{VAE}$

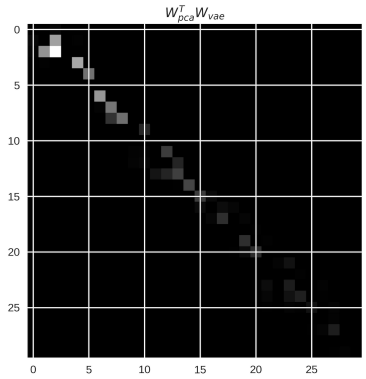

Figure 3: The gram matrix $W_{pPCA}^T W_{VAE}$

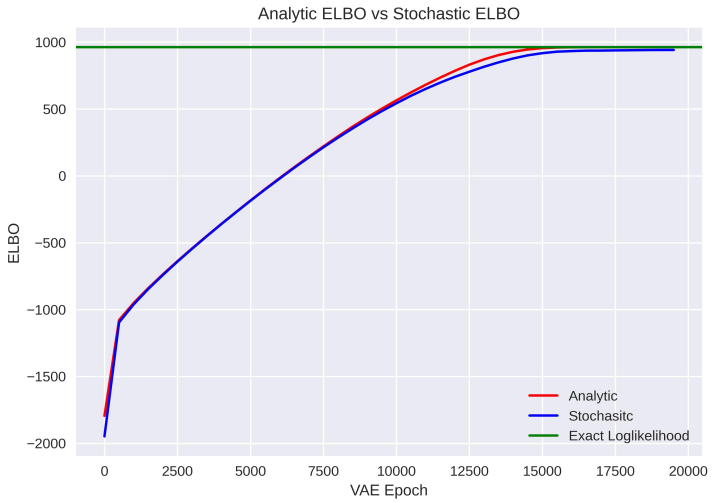

Figure 4: Analytic ELBO vs Stochastic ELBO

Various Encoder with Linear Decoder The third simulation is to reproduce the "Figure 4" in the original paper. Following the preprocess in Papamakarios et al. (2017) and Dinh et al. (2016)(this is also the preprocess in the paper, see Appendix E), we compare different encoder with same decoder. The result is shown in Figure 5, and we can find that neither of these Encoder can recover the exact log-likelihood.

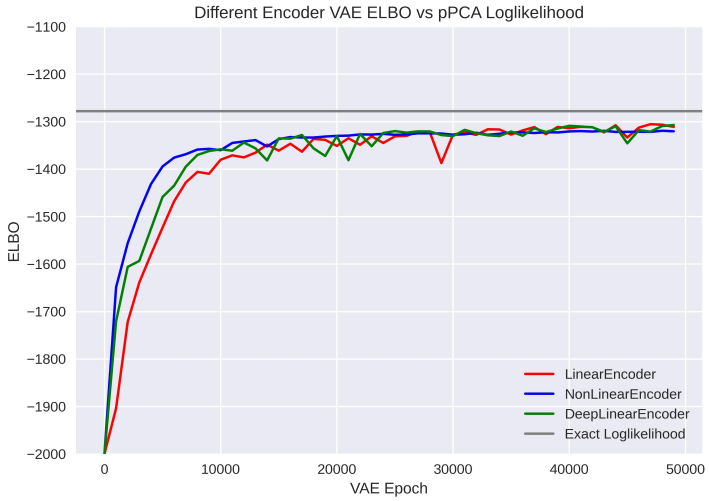

Figure 5: The ELBO of different encoder with linear decoder and the exact loglikelihood

Marginal log-likelihood of pPCA and ELBO of linear VAE with different latent dimension The fourth simulation is to reproduce the "Figure 2" in the original paper. As shown in Figure 6, with the same experienment setting, both three models

perfectly recover the result shown in the original paper.

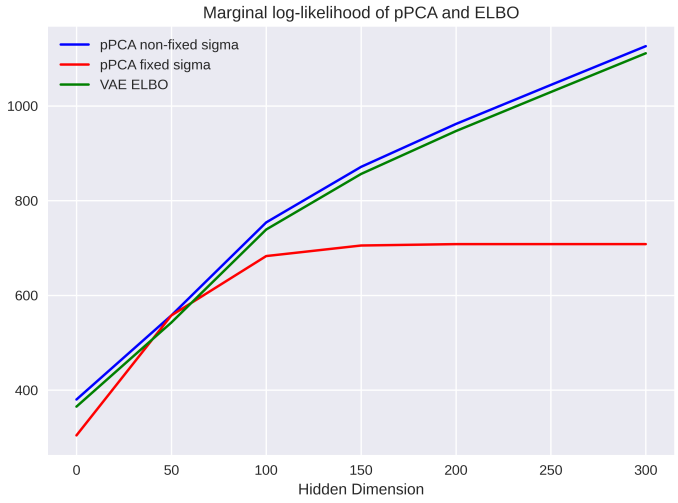

Figure 6: The ELBO of linear VAE and the logLikelihood gained by pPCA with non-fixed σ^2/fixed $\sigma^2(\sigma^2 = \sigma^2_{50})$ in different latent dimension

3.2 Posterior Collapse in Deep Non-Linear VAE

Posterior Collapse in Deep VAE and how different σ^2 influences ELBO in Deep Non-Linear VAE The final simulation is to reproduce the "Figure 5" and "Figure 6" in the original paper. We test the posterior collapse phenomenon on MNIST dataset. Under different σ^2 setting(including whether σ^2 is learnable and the different initial value), we obtain the result as shown in Figure 7 (learnable, different initial value σ^2) and Figure 8(fixed, different initial value σ^2). Due to the page limitation, please find the detail of the ELBO result and the posterior collapse result in *.ipynb* file.

4 Conclusion

According to theoritcal and pratical analysis, the authors prove that maximizing the ELBO objective of linear VAE dose not introduce additional local maxima relative to the loglikelihood objective. And this may can partly explain what have happened in the posterior Collapse of Deep VAE model.

References

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

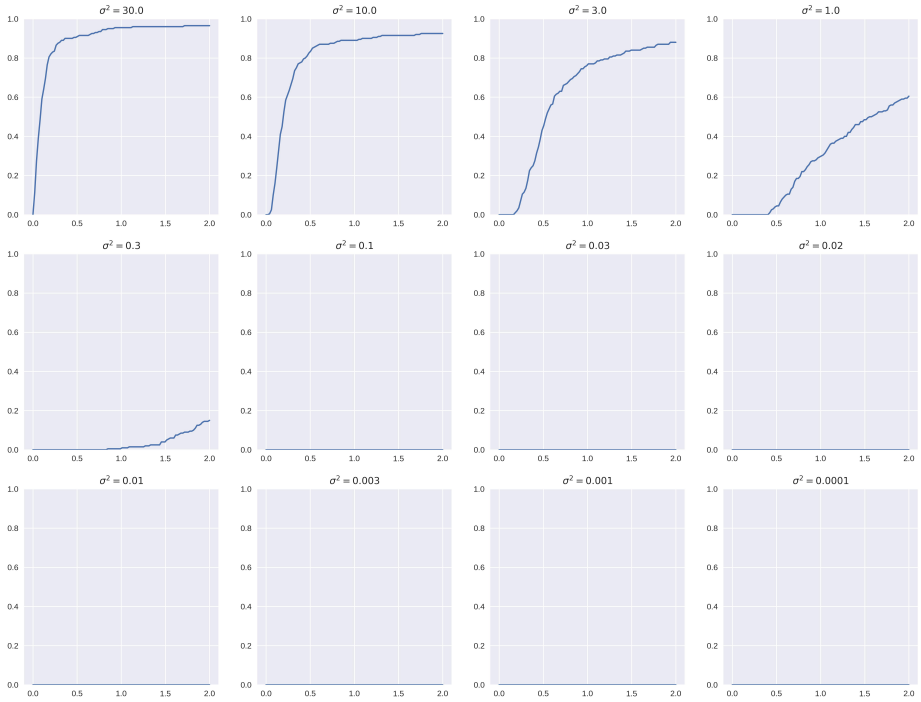

Figure 7: The posterior collapse

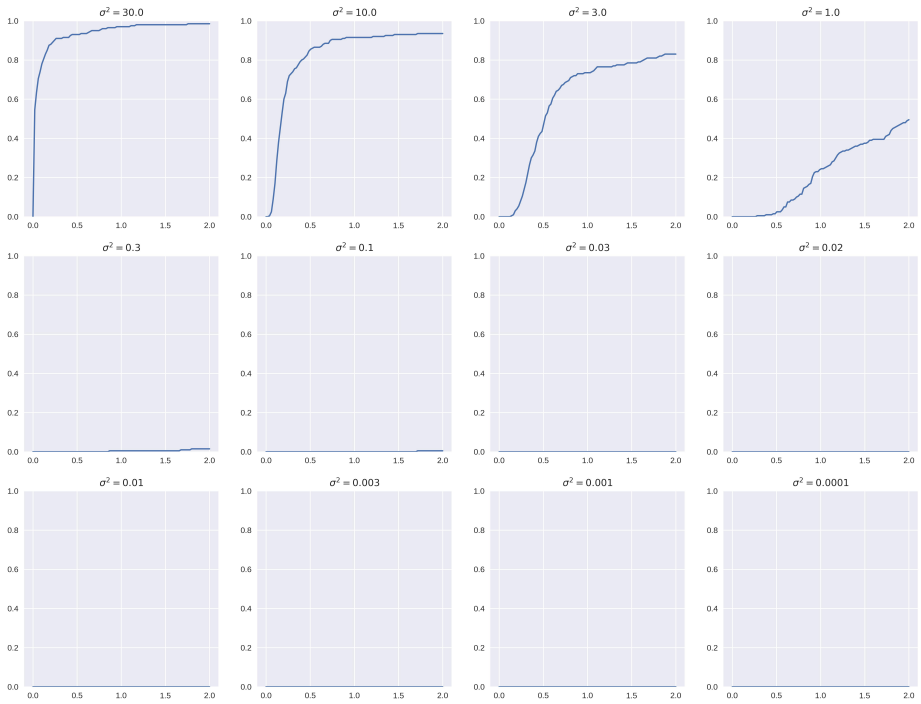

Figure 8: The posterior collapse

Kingma, D. P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lucas, J., G. Tucker, R. B. Grosse, and M. Norouzi (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems 32*.

Papamakarios, G., T. Pavlakou, and I. Murray (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems 30*.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.