

# SUPPLEMENTARY MATERIAL

**Anonymous authors**

Paper under double-blind review

## A FURTHER EXPERIMENTS

### A.1 CONVOLUTIONAL MODELS FOR MNIST AND OMNIGLOT

Table 1: Test ELBO on static MNIST and OMNIGLOT for convolutional models

Model	MNIST	OMNIGLOT
VAE (L=2)	-82.41	-97.65
VampPrior (L=2)	-81.09	-97.56
HVAE (L=4), $\mathcal{L}_{5000}$	-81.58	-96.08

### A.2 SINGLE LAYER VAE

Table 2: Active units on a 1-layer for HVAE versus a vanilla VAE on dynamic MNIST. The training was stopped after 1M steps. Models have a single layer of 64 latent variables.

Model	V. ELBO	Active Units
VAE	-89.7	18
HVAE	-85.1	31

### A.3 TIMING COMPARISON

Table 3: Average time per epoch on a 4 stochastic layer MLP model on MNIST

Model	Average Time/Epoch (seconds)
VAE	4.2
IWAE	4.4
HVAE	4.6

### A.4 VARIABLE SELECTION OR POSTERIOR COLLAPSE?

In this section we present an experiment to investigate whether the lower number of active units with standard VAE training can be interpreted as variable selection aimed at reducing model complexity.

In the table 4 we show models trained with successively smaller latent dimensions on MNIST. All models have 4 stochastic layer with two layer of 200 units in each stochastic layer. The latent dimensions in all layers are the same in each model and are chosen from  $\{40, 30, 20, 10, 5\}$ . All models are standard VAE’s trained using KL annealing for 1M steps.

It can be seen in the table that the number of active units in the top two layers are very similar for all models despite the fact that the validation ELBO varies vastly and is significantly poor for lower dimension models. This suggests that posterior collapse in higher layers in VAE’s is due to loss of essential information at the higher levels rather than pruning to reduce complexity.

Table 4: Varying latent dimensions on a 4 stochastic layer MLP model on MNIST

Model	V. ELBO $\mathcal{L}_{100}$	KLD	Top KLD	Active units
40-40-40-40	-84.73	28.07	1.13	1,6,15,40
30-30-30-30	-88.7	26.5	0.69	1,6,13,30
20-20-20-20	-85.4	23.9	1.38	1,5,12,20
10-10-10-10	-90.47	18.38	1.30	2,3,9,10
5-5-5-5	-104.36	12.59	0.636	1,3,5,5

#### A.5 COMPARISON OF GRADIENT VARIANCE

We empirically show that variance is reduced by smoothing in figure 1. We train a 4 layer VAE and periodically measure the gradient variance relative to the mean output of the first encoder layer over 100 samples both with and without OU smoothing. We use  $\rho = 0.8$  for the smoothing.

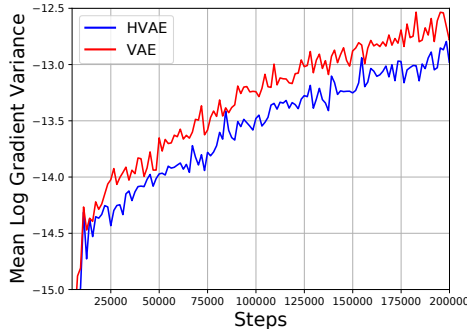


Figure 1: Comparison of gradient variance

#### A.6 OUTPUT VARIANCE

It is known that for many common output distributions the maximum likelihood problem for deep latent variable models is ill-posed [1], leading to unbounded likelihoods. With Gaussian output models one way to mitigate the problem is to constrain the output variance (see proposition 2 in [1] for a justification). However, it is of interest to see whether in particular practical instances VAE models suffer from the problem of unbounded likelihood even with unconstrained variance.

To investigate this for our method we perform an experiment with a 4 stochastic layer ResNet decoder VAE on CIFAR-10. We replace the decoder output with the simpler Gaussian decoder and measure the minimum variance across the output dimensions. We also give the training and validation ELBOs for reference. The model is trained for 150,000 steps.

The results are shown in figure 2. Here we see that the minimum output variance over the course of training remains bounded away from 0 (at around 0.001) and that the train and validation ELBO are similar. This implies that the likelihood values produced by the model remain bounded.

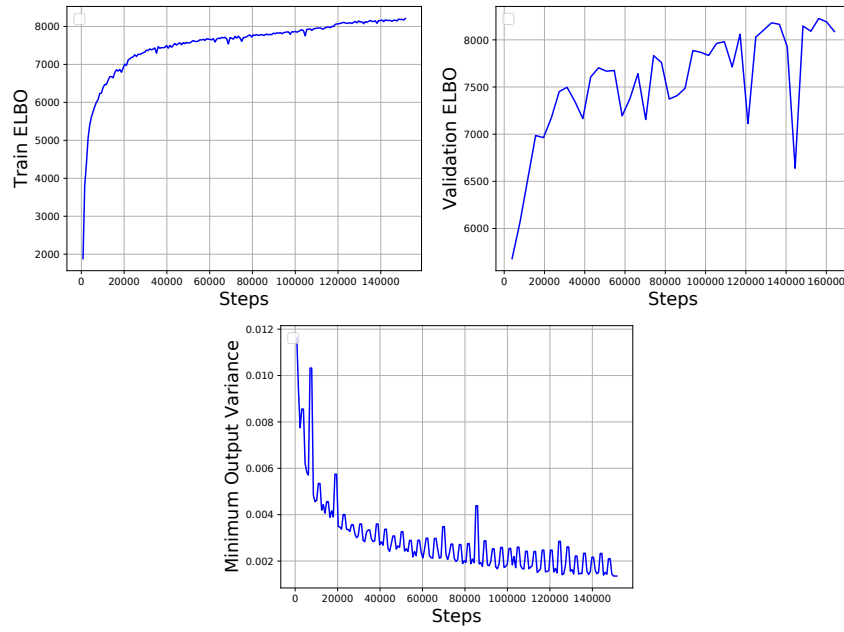


Figure 2: Investigating unbounded likelihoods

#### A.7 FURTHER EMPIRICAL EVIDENCE FOR PHASE TRANSITIONS

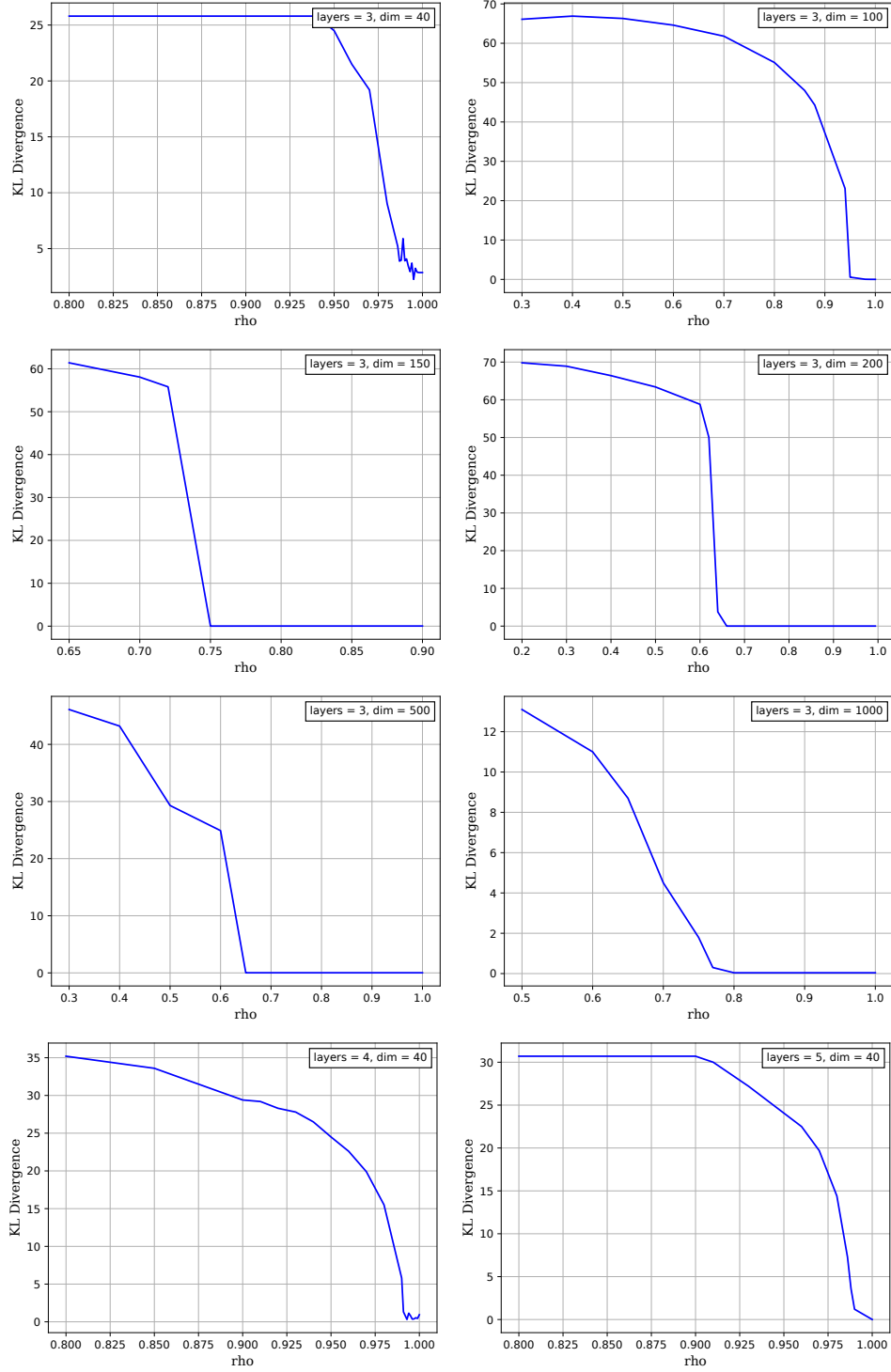


Figure 3: The top layer KL divergence vs.  $\rho$  on OMNIGLOT with MLP models for 3, 4 and 5 stochastic layer models. The latent dimensions are indicated in the inset. It can be seen that the critical threshold is lower for larger latent dimensions.

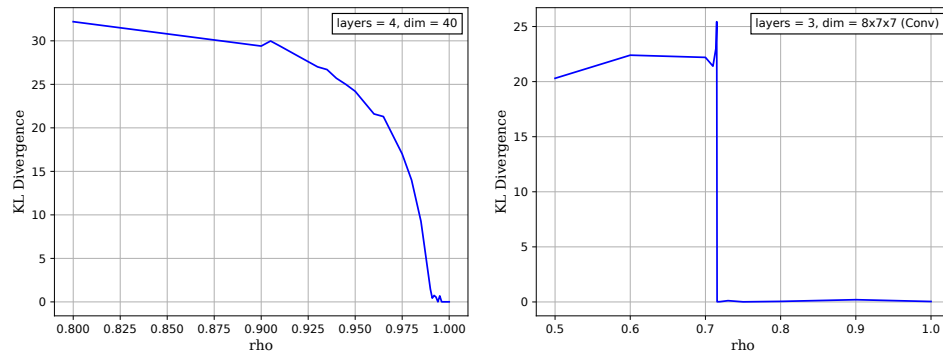


Figure 4: The top layer KL divergence vs.  $\rho$  on MNIST with an MLP model (left) and a convolutional model (right). It can be seen that the convolutional architecture has an especially steep transition.

## REFERENCES

- [1] Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models.