A. Expanding on experiments

A.1. Convolutional Models for MNIST and OMNIGLOT

Table 5. 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
OU-VAE (L=4), $\mathcal{L}_{5000}$	-81.58	-96.08

A.2. Posterior collapse on static MNIST

Table 6. Posterior collapse on static MNIST. The table shows top layer KL divergence and active units (top-to-bottom) for various method on the same 4 stochastic layer model with 40 units each, '+KL' indicates KL annealing. All models were trained for 1M steps with the same hyperparameters.

Model	V. ELBO ($\mathcal{L}_{100}$)	Top KLD	Active Units
VAE	-86.4	0.82	1,3,9,17
IWAE (5 samples)	-84.9	1.005	1,3,9,24
VAE+KL	-84.5	1.2	2,3,11,37
VAE+Freebits	-87	1.3	1,1,8,17
OU-VAE ($\rho = 0.85$)	-85.2	20.6	40,40,40,24
OU-VAE ($\rho = 0.9$)	-85.5	19.5	40,40,40,40
OU-VAE+KL ($\rho = 0.95$)	-84.2	23.5	38,34,17,40

A.3. Single Layer VAE

Table 7. 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.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 8 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 8. 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. Timing Comparison

Table 9. Average time per epoch on a 4 stochastic layer MLP model on MNIST

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

A.6. Memory Comparison

Table 10. Comparing maximum memory usage for various models. We report the memory usage 4 and 6-layer models on CIFAR-10, using 5 deterministic layers per stochastic layer with 100 units per layer and 5 OU or IWAE samples. The memory usage is the maximum amount of used memory with a batch size of 64. For LVAE and BIVA we use code from the BIVA (Maaløe et al., 2019) PyTorch repository.

Model	Layers	Memory	Parameters (M)
VAE	4	2.12G	12.4
VAE	6	2.45G	16.8
IWAE	4	8.59G	12.4
IWAE	6	10.5G	16.8
LVAE	15	6.4G	59.5
BIVA	15	10.3G	103
OU-VAE	4	3.74G	12.4
OU-VAE	6	4.6G	16.8

A.7. Comparison of Gradient Variance

We empirically show that variance is reduced by smoothing in figure 4. 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.

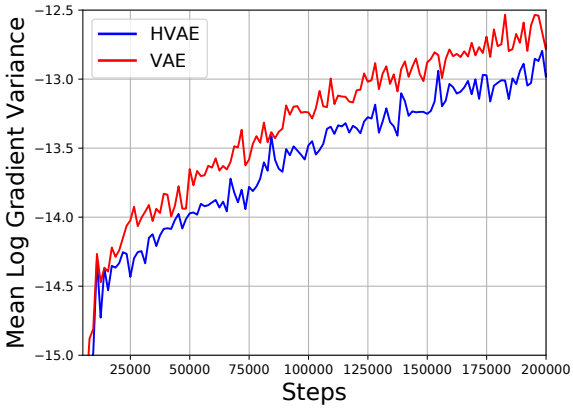

Figure 4. Comparison of gradient variance

A.8. Further Empirical Evidence for Phase Transitions

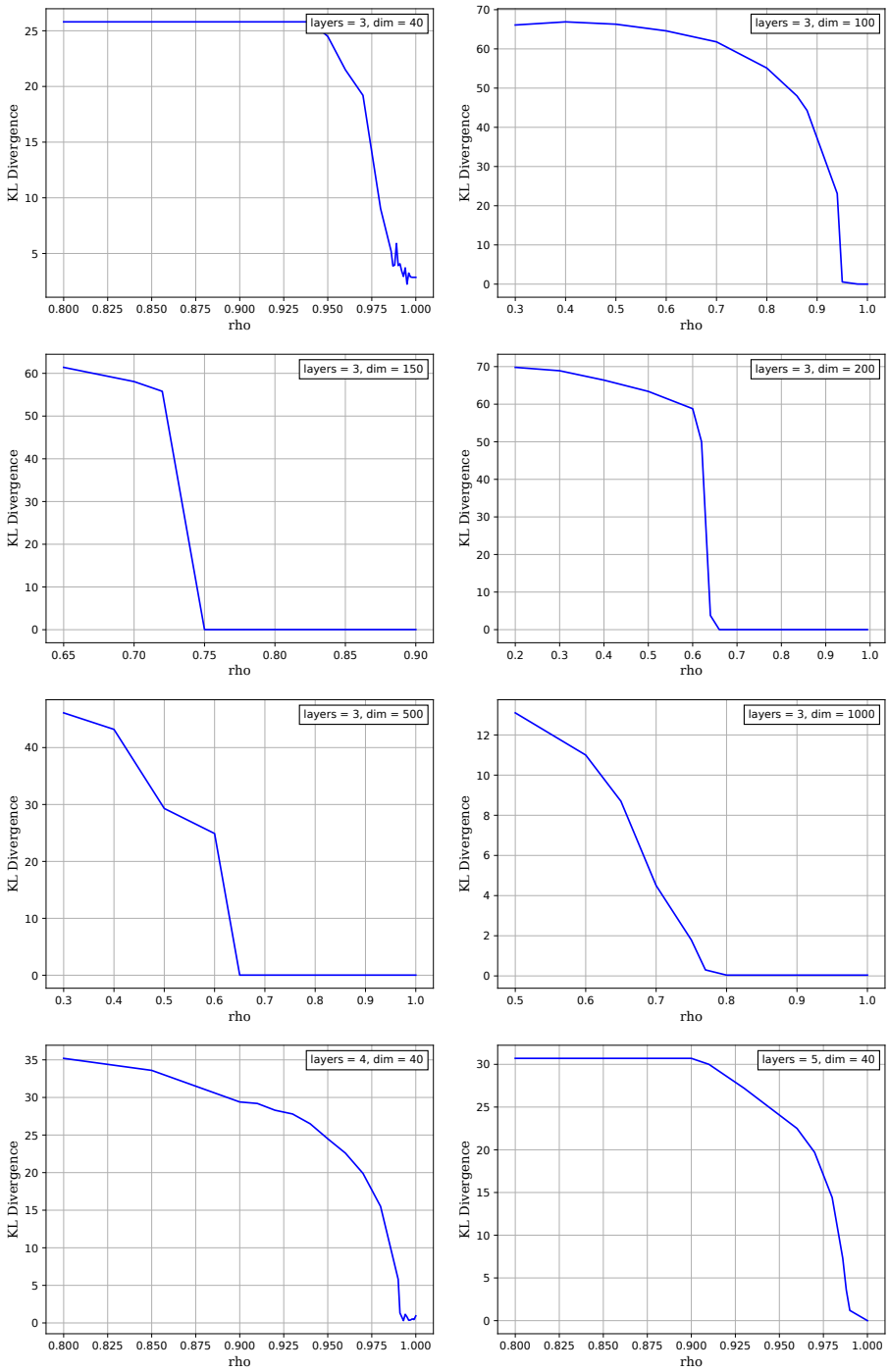

Figure 5. The top layer KL divergence *vs.* ρ on OMNIGLOT with MLP models for 3, 4 and 5 stochastic layer models. The latent dimensions are indicated in the inset. It can be seen that the critical threshold is lower for larger latent dimensions.

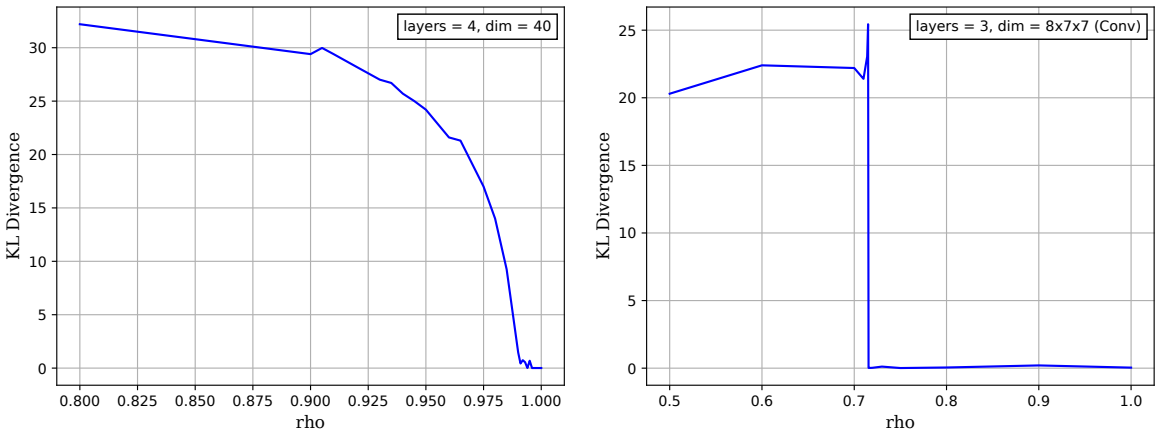

Figure 6. The top layer KL divergence *vs.* ρ on MNIST with an MLP model (left) and a convolutional model (right). It can be seen that the convolutional architecture has an especially steep transition.