

Fixed $\sigma^2$	ELBO	$\sigma^2$ -tuned ELBO	Tuned $\sigma^2$	Posterior collapse (%)	KL Divergence
30	$-1850.0 \pm 29.0$	$-1374.9 \pm 199.0$	4.451	91.78	$10.9 \pm 6.69$
10	$-1450.3 \pm 4.17$	$-1098.2 \pm 28.3$	2.797	89.08	$28.8 \pm 1.39$
3	$-1114.9 \pm 1.05$	$-1018.8 \pm 0.99$	1.361	59.50	$58.5 \pm 1.39$
1	$-1022.1 \pm 5.42$	$-1018.3 \pm 5.28$	1.140	8.28	$125.4 \pm 4.19$
0.3	$-1816.7 \pm 270.6$	$-1104.6 \pm 6.23$	1.28	1.5	$179.3 \pm 85.9$
0.1	$-3697.3 \pm 493.3$	$-1190.8 \pm 37.4$	0.968	1.9	$368.8 \pm 94.6$

Table 1: Evaluation of deep Gaussian VAEs (averaged over 5 trials) on real-valued MNIST. Collapse percent gives the percentage of latent dimensions which are within 0.01 KL of the prior for at least 99% of the encoder inputs. Note that these values differ from those in the current manuscript as we have adopted the procedure from Papamakarios et al. 2017 for ease of comparison (this is ultimately non-linear preprocessing and a constant shift to all values). All results in the paper are now consistent with this.

1 We are grateful for your comments and suggestions. Our paper provides a thorough, novel, theoretical analysis of the  
2 linear VAE and builds a clear picture of posterior collapse in this model. We showed that the linear VAE can be trained  
3 with ELBO without spurious local maxima and is fully identifiable. Empirically, we explored the extent to which the  
4 linear model can explain observations of the non-linear case. Per reviewer feedback, we have added additional empirical  
5 results (a subset of which are included here) which aim to better explore this relationship.

6 **Experiments and theory (R1/R3/R4)** We agree that the experiments and theory could have been better aligned. To  
7 correct this we now directly measure posterior collapse statistics in Table 1 (subset included top). The analysis of the  
8 linear model predicts that larger  $\sigma^2$  will lead to more posterior collapse and worse ELBO which is generally the case.  
9 However, the non-linear model has some unexplained behaviours — e.g. the best model has some posterior collapse.

10 **Claims for non-linear case (R1/R4)** We agree that our choice of language could be improved with regards to these  
11 claims and have softened language in these areas of the paper. To help transition between the theory and experiments,  
12 per R1’s suggestion, we have clarified that we study the non-linear VAE empirically (and not theoretically) in Section 5.

13 **Linear decoder with non-linear encoder? (R1)** With a linear decoder and non-linear encoder, Lemma 1 still holds,  
14 and the optimal variational distribution is the same as the true posterior has not changed. However, Corollary 1 and  
15 Theorem 1 no longer hold in general. Even a deep linear encoder will not have a unique global maximum and new  
16 stationary points (possibly maxima) may be introduced to ELBO in general. We have added experiments exploring  
17 different encoders with linear decoders (see Figure 1). We do not expect the linear encoder to be out-performed and  
18 indeed the empirical results support this. Also note that the references you provided do not use Gaussian observation  
19 models and so are much harder to analyze (see Appendix C.1 for an example).

20 **Related work (R1)** We have included the references and added a long-form related work section to the appendix.

21 **Significance (R1):** We acknowledge that the theoretical results apply only to the linear case but argue this is significant  
22 nonetheless. We give a novel interpretation of posterior collapse which is theoretically grounded and opposes existing  
23 folk-wisdom (that the KL term is responsible). With our results, linear VAEs provide a simple, well-understood, test-bed  
24 for analyzing new VAE training strategies. Finally, we proved that linear VAEs are without local maxima and are fully  
25 identifiable (unlike regularized linear autoencoders which only identify the orthogonal subspace [Kunin et al. 2019]).

26 **Solutions to posterior collapse (R3)** We have identified that  
27 the linear case does not need a novel solution: learning the  
28 observation noise is sufficient for finding the global maximum.  
29 Experimentally, we identified that the non-linear case has chal-  
30 lenges in learning  $\sigma^2$  which we hypothesize leads to worse  
31 models. Posterior collapse is a widely studied, challenging  
32 problem which is poorly understood. We believe that the results  
33 in our paper will guide researchers’ search for solutions.

34 **Fixing  $\sigma^2$  in the wild (R4)** We believe this stems from a mis-  
35 understanding of VAEs as a regularized autoencoder rather than a  
36 probabilistic model and note it was also observed by Dai &  
37 Wipf, 2019. We can remove this text if preferred.

38 **Intuitions from Table 1 (R4)** Yes, we have identified that initialization and other pre-processing are critical for training  
39 the non-linear VAEs (which the linear VAE theory does not predict).

40 **Evidence from two datasets (R4)** We will soften this statement. However, the results around fixed  $\sigma^2$  were very  
41 consistent across all experiments: larger  $\sigma^2$  learned a less rich representation.

42 **Additional reference:** George Papamakarios, Theo Pavlakou, Iain Murray. *Masked Autoregressive Flow for Density*  
43 *Estimation, 2017.*

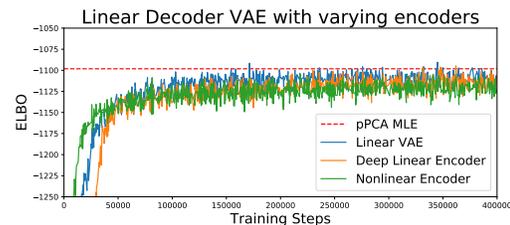


Figure 1: VAEs with linear decoders trained on real-valued MNIST. Final average ELBO on training set are (ordered by legend): -1098.2, -1108.7, -1112.1, -1119.6.