Don't blame the ELBO! A Linear VAE Perspective on Posterior Collapse

James Lucas*† (jameslucas@google.com) George Tucker* Roger Grosse† Mohammad Norouzi*

†University of Toronto & Vector Institute *Google Brain

Quick Summary

We study the cause of posterior collapse in VAEs. Hypothesis: optimization plays a critical role.

- Analyze the loss landscape of Linear VAEs
- Present a thorough empirical analysis of posterior collapse in nonlinear VAEs

Linear VAE

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

We optimize the Linear VAE using the ELBO,

$$
\mathcal{L} := \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z})] - D_{KL}(q(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z})) \tag{3}
$$

For this model, we can write down the ELBO in closed form and compute analytic (deterministic) gradients.

Probabilistic PCA (pPCA)

The pPCA model [1] is:

$$
\begin{aligned}
p(\mathbf{z}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
p(\mathbf{x} \mid \mathbf{z}) &= \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I})
\end{aligned} \tag{1}
$$

Has an analytic maximum likelihood estimate and true posterior.

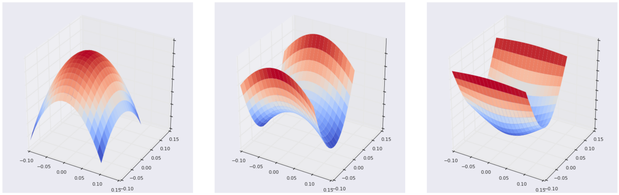
Tipping and Bishop [1] characterize the stationary points of the pPCA log-likelihood.

Probabilistic PCA Stationary Points

$$
\sigma^2_{\text{MLE}} = \frac{1}{n-k} \sum_{j=k+1}^{n} \lambda_j, \qquad \mathbf{W}_{\text{MLE}} = \mathbf{U}_k (\boldsymbol{\Lambda}_k - \sigma^2_{MLE} \mathbf{I})^{1/2} \mathbf{R}. \tag{4}
$$

Other stationary points have zero-columns in $\mathbf{W}$ corresponding to posterior collapse.

Escaping posterior collapse: As σ^2 is decreased, the conditioning of the objective at these points improves.

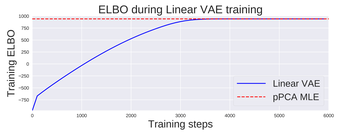


(a) σ^2 large: local maxima (b) σ^2 medium: saddle point (c) σ^2 small: local minima

VAE Analysis

Training Linear VAE with ELBO →

$W^T W =$



← Learns orthogonal weights

An expected result: The Linear VAE recovers pPCA.

Lemma 1. *The global maximum of the ELBO objective ((3)) for the linear VAE ((2)) is a global maximum of the marginal log-likelihood of pPCA.*

A less obvious result: The Linear VAE is fully identifiable (and computes an SVD).

Lemma 2. *The global optimum to the VAE solution has the scaled principal components as the columns of the decoder network.*

A surprising result: ELBO does not hurt Linear VAE optimization

Theorem 3. *The ELBO objective does not introduce any additional local maxima to the pPCA model.*

References

[1] Tipping and Bishop (1999), Probabilistic principal component analysis

[2] Alemi et al. (2017), Fixing a broken ELBO

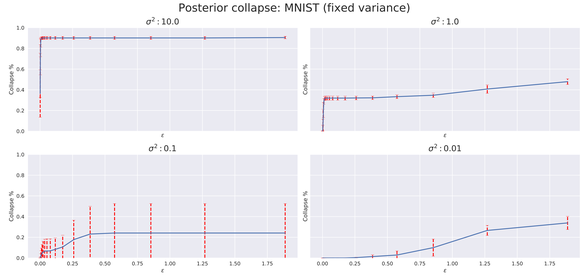
[3] He et al. (2019) Lagging inference networks and posterior collapse in variational autoencoders

[4] Dai et al. (2017), Hidden talents of the variational autoencoder

[5] Rolinek et al. (2018) Variational autoencoders pursue PCA directions (by accident)

Nonlinear VAE Experiments

Observation noise also controls posterior collapse in non-linear VAEs. (Gaussian MNIST VAE)



Correctly tuning σ^2 enables better representations to be learned with reduced posterior collapse.

	Model		ELBO	σ^2-tuned ELBO	Tuned σ^2	Posterior collapse (%)	KL Divergence
	Init σ^2	Final σ^2					
MNIST	10.0		-1450.3 ± 4.2	-1098.2 ± 28.3	1.797	89.88	28.8 ± 1.4
	1.0		-1022.1 ± 5.4	-1018.3 ± 5.3	1.145	27.38	125.4 ± 4.2
	0.1		-3697.3 ± 493.3	-1190.8 ± 37.4	0.968	3.25	368.7 ± 94.6
	0.01		-38612.5 ± 1189.8	-2090.8 ± 975.1	0.877	0.0	695.9 ± 118.1
	0.001		-504259.1 ± 49149.8	-1744.7 ± 48.4	0.81	0.0	756.2 ± 12.6
	10.0	1.32	-1022.2 ± 4.5	-1022.3 ± 4.6	1.318	73.75	73.8 ± 9.8
	1.0	1.183	-1011.1 ± 2.7	-1011.1 ± 2.8	1.182	47.88	106.3 ± 2.5
	0.1	1.194	-1025.4 ± 8.6	-1025.4 ± 8.6	1.195	29.25	116.1 ± 11.4
	0.01	1.194	-1030.6 ± 3.5	-1030.5 ± 3.5	1.191	23.0	121.9 ± 7.7
	0.001	1.208	-1038.7 ± 5.6	-1038.8 ± 5.6	1.209	27.0	124.9 ± 1.6
CELEBA 64	10.0		-73328.4 ± 0.49	-55186.7 ± 35.1	0.204	80.56	56.12 ± 0.42
	1.0		-59841.8 ± 30.1	-51294.8 ± 333.7	0.102	2.52	213.4 ± 6.3
	0.1		-50760.3 ± 353.4	-50698.5 ± 393.9	0.0883	32.72	483.8 ± 36.2
	0.01		-82478.7 ± 1823.3	-51373.9 ± 213.3	0.0817	0.0	1624.2 ± 8.78
	0.001		-531924.5 ± 17177.6	-57381.5 ± 512.6	0.0296	0.0	2680.2 ± 41.45
	10.0	0.0962	-51109.5 ± 408.2	-51109.5 ± 408.3	0.0963	53.32	364.5 ± 26.4
	1.0	0.0875	-50631.2 ± 163.4	-50631.0 ± 163.3	0.0875	54.76	462.2 ± 20.0
	0.1	0.0863	-50646.9 ± 269.0	-50645.9 ± 267.5	0.0869	28.84	520.9 ± 11.7
	0.01	0.0911	-51285.0 ± 708.1	-51284.8 ± 708.1	0.0963	5.64	557.0 ± 50.5
	0.001	0.104	-51695.1 ± 322.4	-51694.8 ± 322.7	0.0974	0.0	537.5 ± 46.2

The linear VAE is largely predictive of the nonlinear high capacity VAEs.

Project website (with code):

https://sites.google.com/view/dont-blame-the-elbo

Paper: arXiv/1911.02469