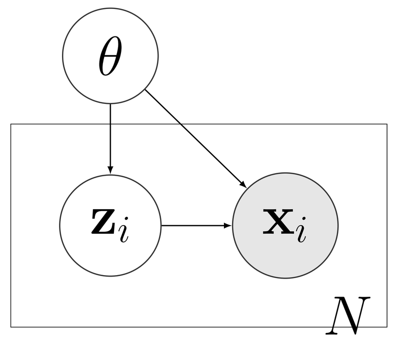
Auto-Encoding Variational Bayes

Paweł F. P. Budzianowski, Thomas F. W. Nicholson, William C. Tebbutt

UNIVERSITY OF CAMBRIDGE

Problem Definition

Perform approximate inference in model with local latent variables z_i whilst learning point estimates for the MAP solution for the global parameters θ having observed x_i.



SGVB

Stochastic Gradient Variational Bayes provides a method to find a deterministic approximation to an intractable posterior distribution by finding parameters ϕ such that $\mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \,||\, p_\theta\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right)\right)$ is minimised for all i. This is achieved by, for each observation, maximising a lower bound

$$\mathcal{L}\left(\phi; \mathbf{x}_i\right) = \mathbb{E}_{q_\phi(\mathbf{z}_i \,|\, \mathbf{x}_i)}\left[\log p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_i\right)\right] - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \,||\, p\left(\mathbf{z}_i\right)\right).$$

The expectation term in this lower bound cannot typically be computed exactly.

$$\tilde{\mathcal{L}}^B\left(\theta, \phi; \mathbf{x}_i\right) = \frac{1}{L}\sum_{l=1}^{L}\left(\log p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_{i,l}\right) - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \,||\, p_\theta\left(\mathbf{z}_i\right)\right)\right),$$

where reparameterising $\mathbf{z} = g_\phi(\mathbf{x}, \epsilon)$ with $\epsilon \sim p(\epsilon)$ yields a differentiable Monte Carlo approximation.

Variational Autoencoder

The Variational Autoencoder is a generative latent variable model for data in which $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{x}_i \sim p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_i\right)$, where this conditional is parameterised by an multi-layer perceptron (MLP).

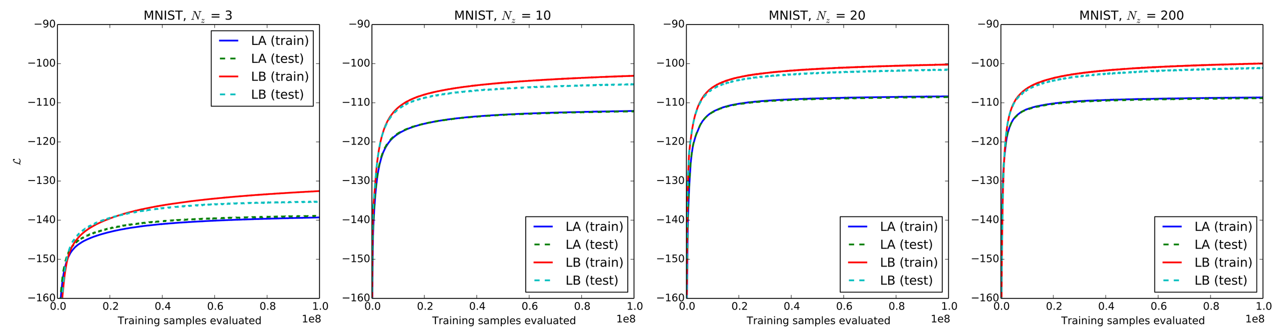
An MLP recognition model $q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right)$ is used to provide fast approximate posterior inference in $\mathbf{z}_i \mid \mathbf{x}_i$.

The MLPs used in the recognition model q_ϕ and conditional distribution $p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_i\right)$ are often compared to the encoder and decoder networks in traditional autoencoders respectively.

Noisy KL-divergence estimate

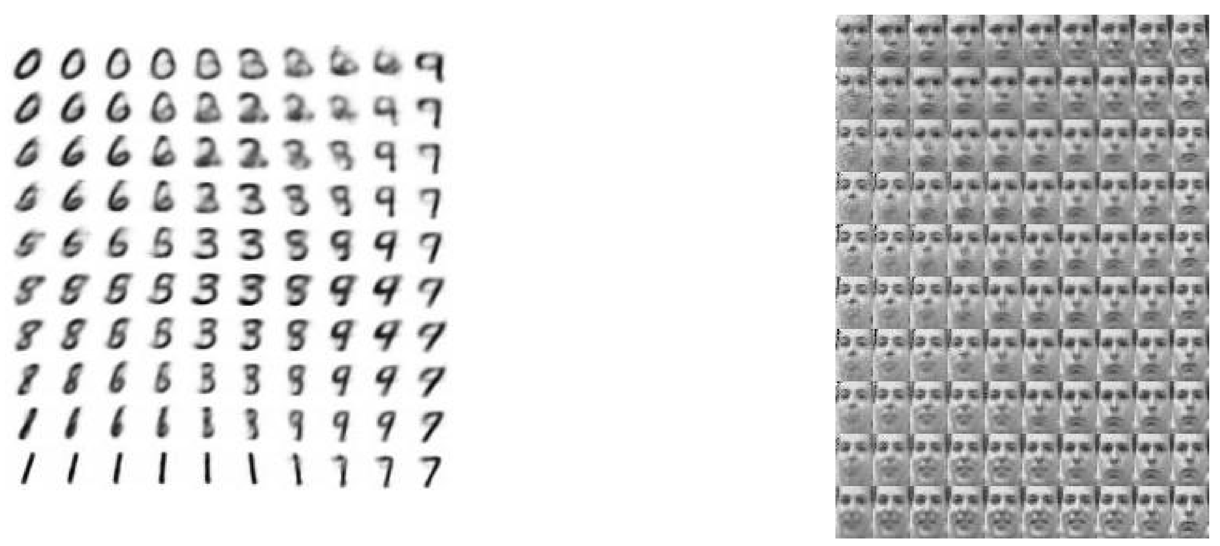
In the case of the non-Gaussian distributions, it is often impossible to obtain closed-form expression for the KL-divergence term which also requires estimation by sampling. This yields more generic estimator of the form:

$$\widetilde{\mathcal{L}}^A(\theta, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L}\left(\log p_\theta(\mathbf{x}^{(i)}, \mathbf{z}^{(i,l)}) - \log q_\phi(\mathbf{z}^{(i,l)}|\mathbf{x}^{(i)})\right).$$



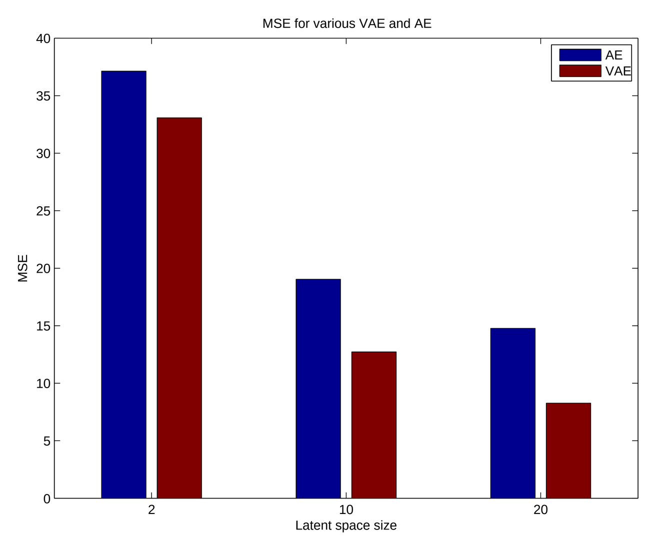
Visualisation of learned manifolds

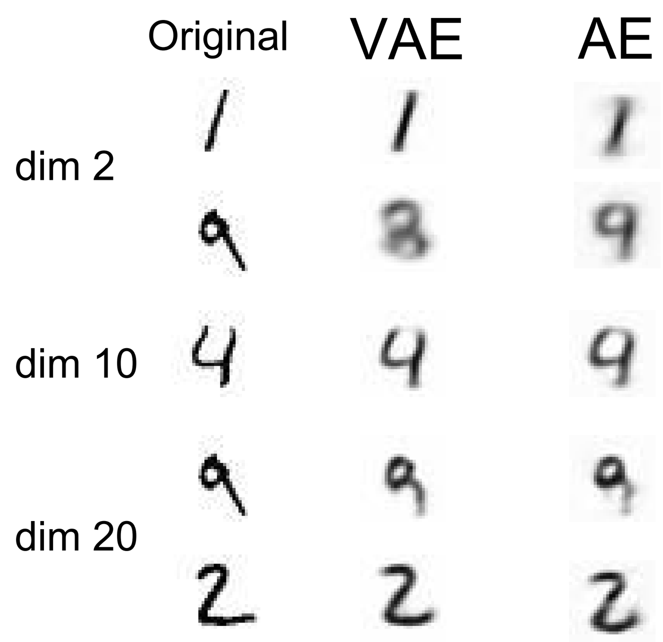
The linearly spaced grid of coordinates over the unit square is mapped through the inverse CDF of the Gaussian to obtain the value of $\mathbf{z}$ which can be used to sample from $p_\theta(\mathbf{x}|\mathbf{z})$ with the estimated parameters $\boldsymbol{\theta}$.



Bayesian: is it really all that?

Comparing reconstruction error to vanilla auto-encoder, we see stronger performance from VAEB.





Full Variational Bayes

Possible to perform full VB on parameters:

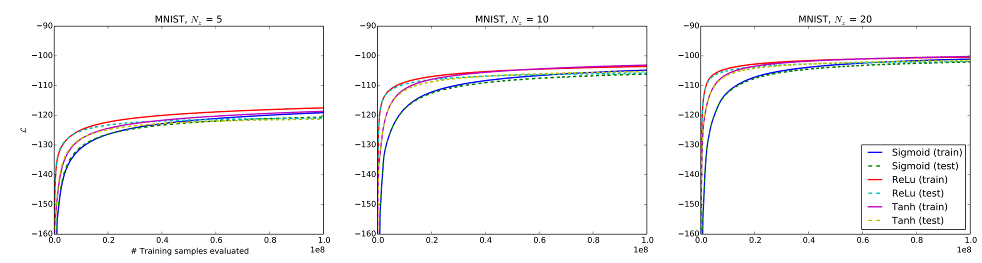
$$\mathcal{L}(\phi; \mathbf{X}) = \int q_\phi(\theta)(log\ p_\theta(X) + log\ p_\alpha(\theta) - log\ q_\phi(\theta))d\theta$$

A differentiable Monte Carlo estimate to perform SGVB, yielding a distribution over parameters.
Implementation showed a decrease of variational lower bound, but no evidence of learning, possibly due to strict Gaussian assumptions of variational approximate posteriors.

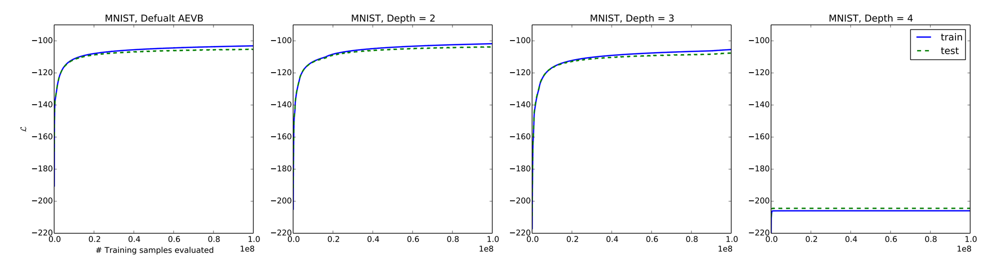
Architecture experiments

We examined various changes to the original architecture of the auto-encoder to test the robustness and flexibility of the model which lead to improvement in terms of optimising the lower bound and computational efficiency.

- Different activation functions.



- Increasing the depth of the encoder.



Future works

I. Scheduled training of VAEB [2].
II. Direct parameterization of differentiable transform [3].
III. Different priors over latent space.

References

1. Kingma, D. P., and Welling M., "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
2. Geras, K. J., and Sutton C., "Scheduled denoising autoencoders." arXiv preprint arXiv:1406.3269 (2014).
3. Tran, D, Ranganath, R. and Blei, M. "Variational Gaussian Process" arXiv preprint arXiv:1511.06499 (2015).