Auto-Encoding Variational Bayes

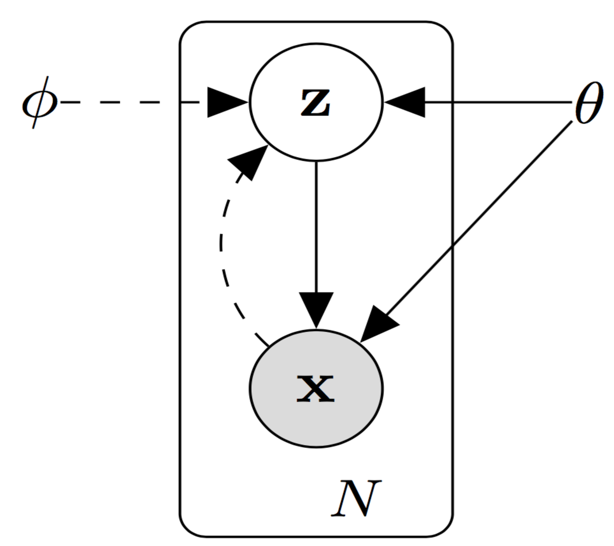
Griffiths Ryan-Rhys, Havasi Marton, Sihui Wang

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

March 16, 2017

Problem Setting

We have an i.i.d. dataset with latent variables per datapoint and would like to perform maximum likelihood (ML) or maximum a posteriori (MAP) inference on the parameters, and variational inference on the latent variables z given observations x. We wish to find an algorithm, using a recognition model $q_\phi(z|x)$ to approximate the intractable true posterior $p_\theta(z|x)$, that works efficiently for a large dataset even when the marginal likelihood is intractable.



The Variational Bound

The log-likelihood can be expressed in terms of a regularization term plus a reconstruction term. The regularization term (KL divergence) depends on how good $q_\phi(z|x)$ can approximate $p_\phi(z|x)$. We will tune ϕ and θ in order to maximize the log-likelihood.

$$\begin{aligned}\log p_\theta(x) &= \int q_\phi(z|x)\log p_\theta(x)dz \\ &= \int q_\phi(z|x)\log p_\theta(x)\frac{p_\theta(z|x)q_\phi(x|z)}{p_\theta(z|x)q_\phi(x|z)}dz \\ &= \int q_\phi(z|x)\log\frac{q_\phi(x|z)}{p_\theta(z|x)}dz + \int q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(x|z)}dz \\ &= D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}\end{aligned}$$

$$\mathcal{L} = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}(\log_\theta(x|z))$$

The Reparameterization Trick

An alternative method for generating samples from $q_\theta(z|x)$:

$$z = g_\phi(\epsilon, x), \quad \epsilon \sim p(\epsilon)$$

In the case of a Gaussian distribution, z can be constructed in the following way:

$$z = \mu + \sigma\epsilon$$

SGVB Estimator

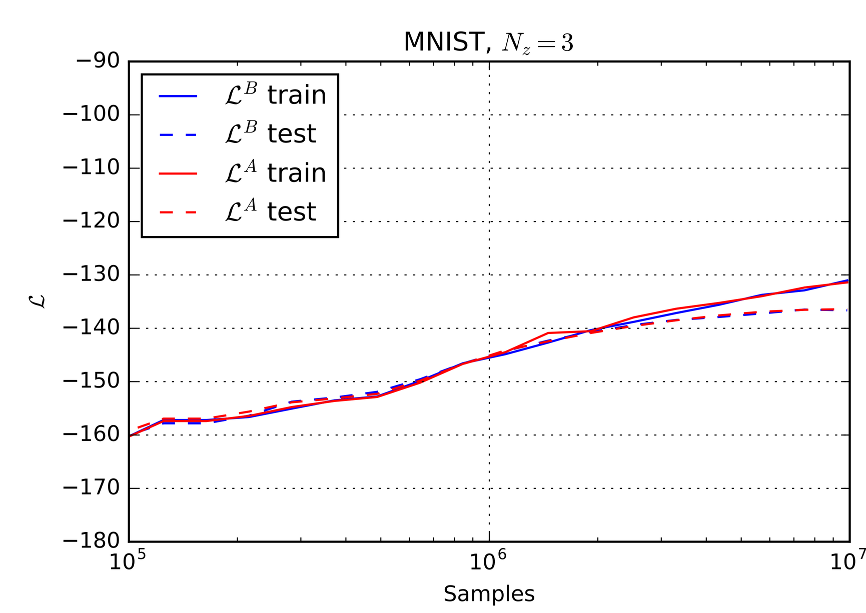
Two practical estimators of the lower bound:

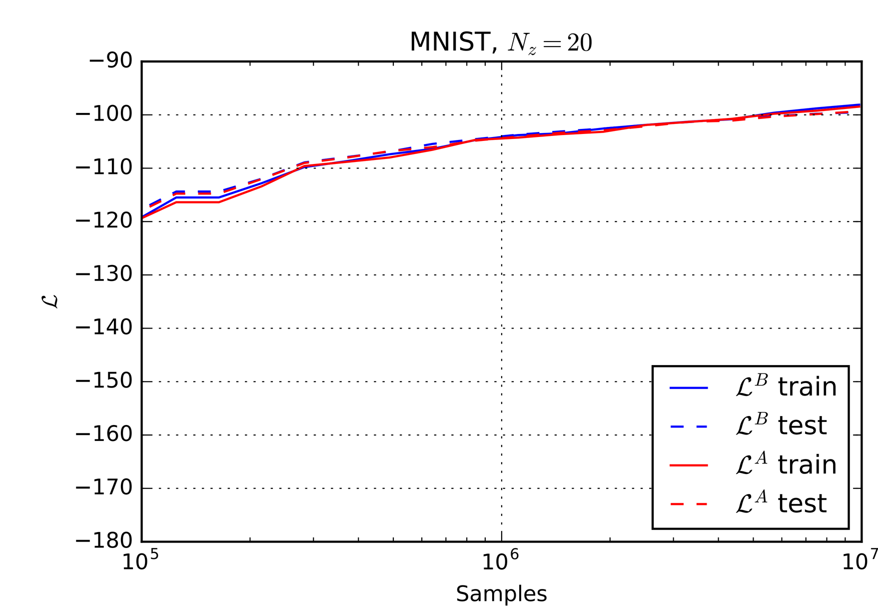
$$\tilde{\mathcal{L}}^A(\theta, \phi; x) = \frac{1}{L}\sum_{l=1}^{L}(\log p_\theta(x, z^l) - \log q_\phi(z^l|x))$$

$$\tilde{\mathcal{L}}^B(\theta, \phi; x) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L}(\log_\theta(x|z^L))$$

When the KL-divergence can be integrated analytically, we use $\mathcal{L}^B$ which typically generates less variance than $\mathcal{L}^A$

Example: Variational Autoencoder
A neural network is used for the probabilistic encoder and the prior over the latent variables is Gaussian.





Visualization of Learned Manifolds

Project high dimensional data to a 2 dimensional manifold.

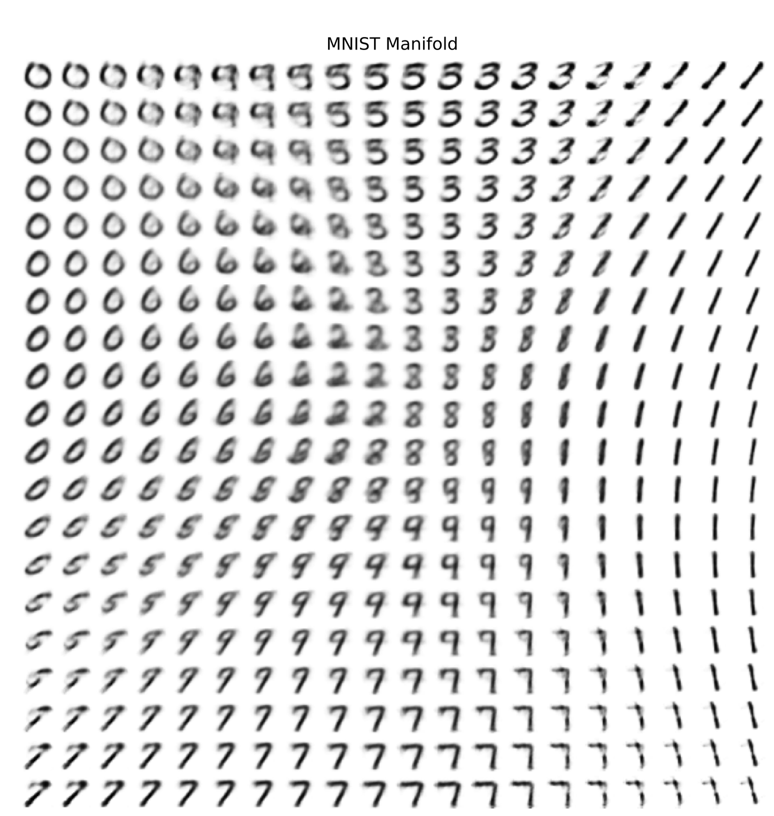
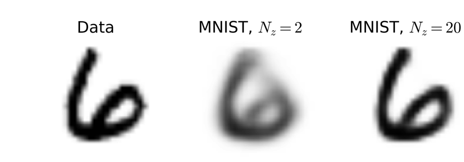


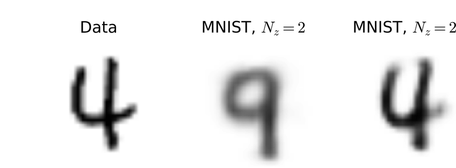
Figure 1: MNIST: $28 \times 28 \to 2$

Figure 2: Frey Face: $20 \times 28 \to 2$

Reconstruction of Images

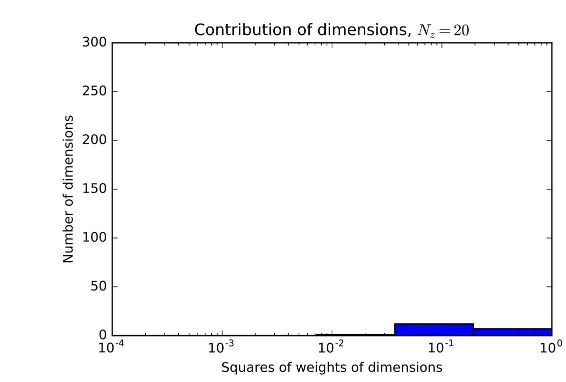
Reconstruction of MNIST with 2 and 20 dimensional latent space.

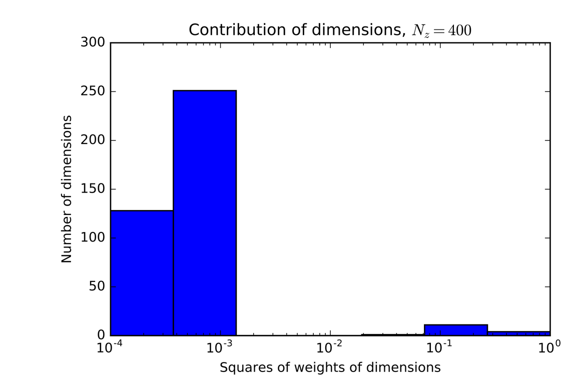




Regularization Effect

The KL-divergence term can be interpreted as regularizing ϕ,encouraging the approximate posterior to be close to the prior $p_\theta(z)$.





Future Work

Extensions will feature both application-based [3] and theoretical [4] research along the following broad avenues:

- Variational autoencoders for automatic chemical design [3].
- The composition of robust features using denoising (variational) autoencoders [4].

References

[1] Diederik P Kingma, Max Welling *Auto-Encoding Variational Bayes.* arXiv preprint arXiv:1312.6114 (2014)

[2] Paisley, John and Blei, David and Jordan, Michael *Variational Bayesian inference with stochastic search.* arXiv preprint arXiv:1206.6430 (2012)

[3] Gómez-Bombarelli, Rafael and Duvenaud, David and Hernández-Lobato, José Miguel and Aguilera-Iparraguirre, Jorge and Hirzel, Timothy D and Adams, Ryan P and Aspuru-Guzik, Alán *Automatic chemical design using a data-driven continuous representation of molecules.* arXiv preprint arXiv:1610.02415 (2016)

[4] Vincent, Pascal and Larochelle, Hugo and Bengio, Yoshua and Manzagol, Pierre-Antoine *Extracting and composing robust features with denoising autoencoders.* Proceedings of the 25th international conference on Machine learning (ACM 2008)