Importance weighted autoencoders

Presented by Bence Halpern

June 17, 2019

Introduction

Introduction

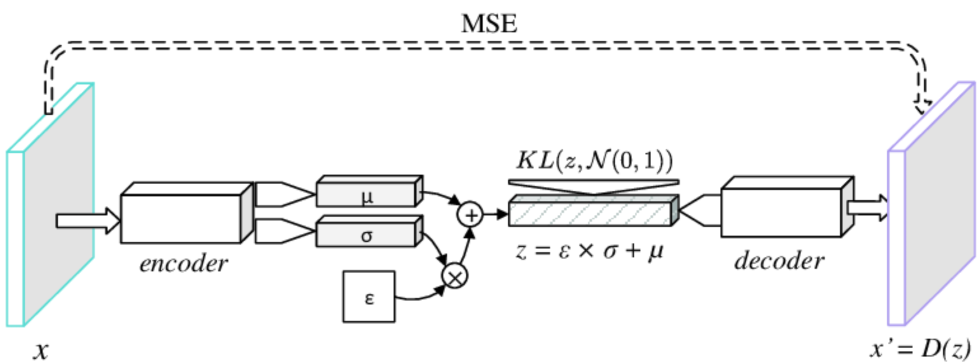
The aim of this talk is the following:

- revisit VAE,
- obtain a probabilistic understanding of VAEs ("Bayesian"),
- build intuition on what importance sampling is,
- descibe how IWAE builds on both to alleviate limitations of VAE.

Revisiting the VAE

Architectural view

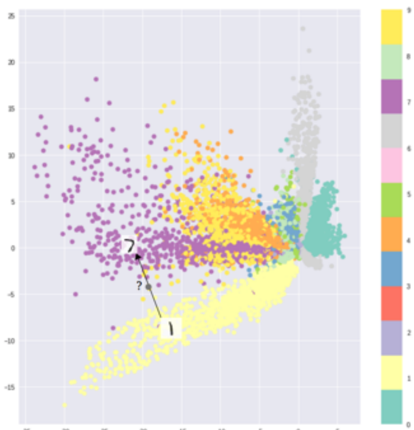
Neural network compresses the data to a lower dimensional probability density function in order to learn about structure.



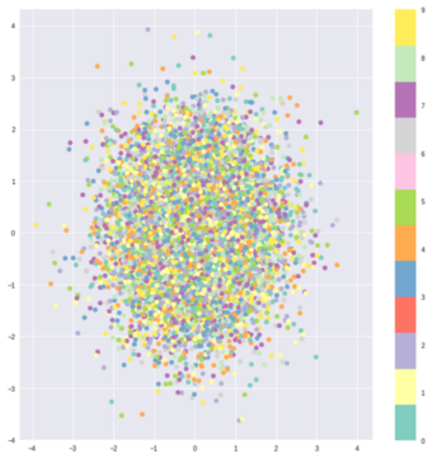
Latent space

Latent space is penalised by KL to have a form specified by or prior (usually $\mathcal{N}(0,1)$)

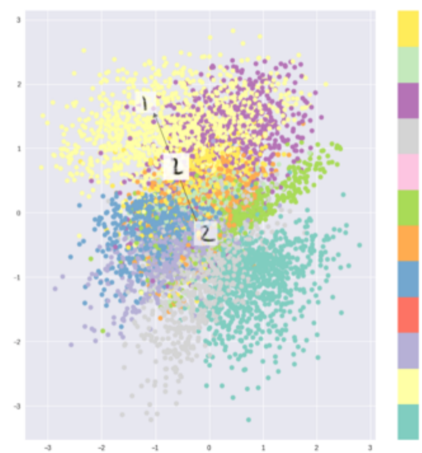
Only reconstruction loss



Only KL divergence



Combination



Our view so far - the regularisation view

We so far looked at the VAE as a latent space regulariser,

$$\mathcal{L}_{\beta\text{-VAE}} = \text{MSE/NLL} + \beta KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \tag{1}$$

where we have a reconstruction loss and a β parameter which can be tuned with your own favourite technique:

- Cross-validation
- information criterions (AIC,BIC)

Probabilistic VAE view

It is difficult to learn posteriors

This is a recurring theme in Bayesian statistics - it is difficult to learn the posterior distribution and the marginal distribution.

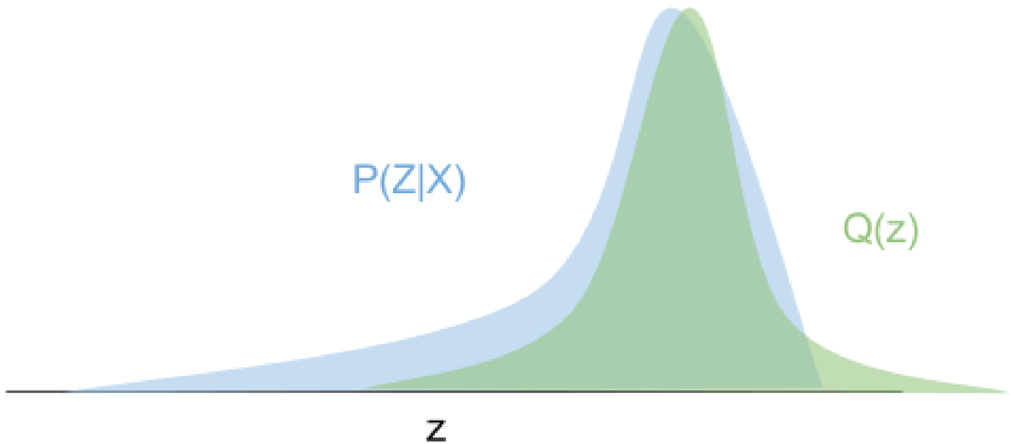
$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z})d\mathbf{z}} \tag{2}$$

The general reason for this is that we have to integrate highly multidiensional functions, which is often analytically impossible and computationally intractable. Solutions:

- Markov Chain (Markov Chain Monte Carlo) - slow but exact
- Variational Inference - faster but inexact

Graphical view on variational inference

Choose a simpler form of the posterior. This makes marginal (more?) tractable so we can optimise.



$$q(\mathbf{z}) = \text{argmax}_{q(\mathbf{z})} KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x})) \quad (3)$$

Obtaining the ELBO

$$KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = E_q[\ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}] \tag{4}$$

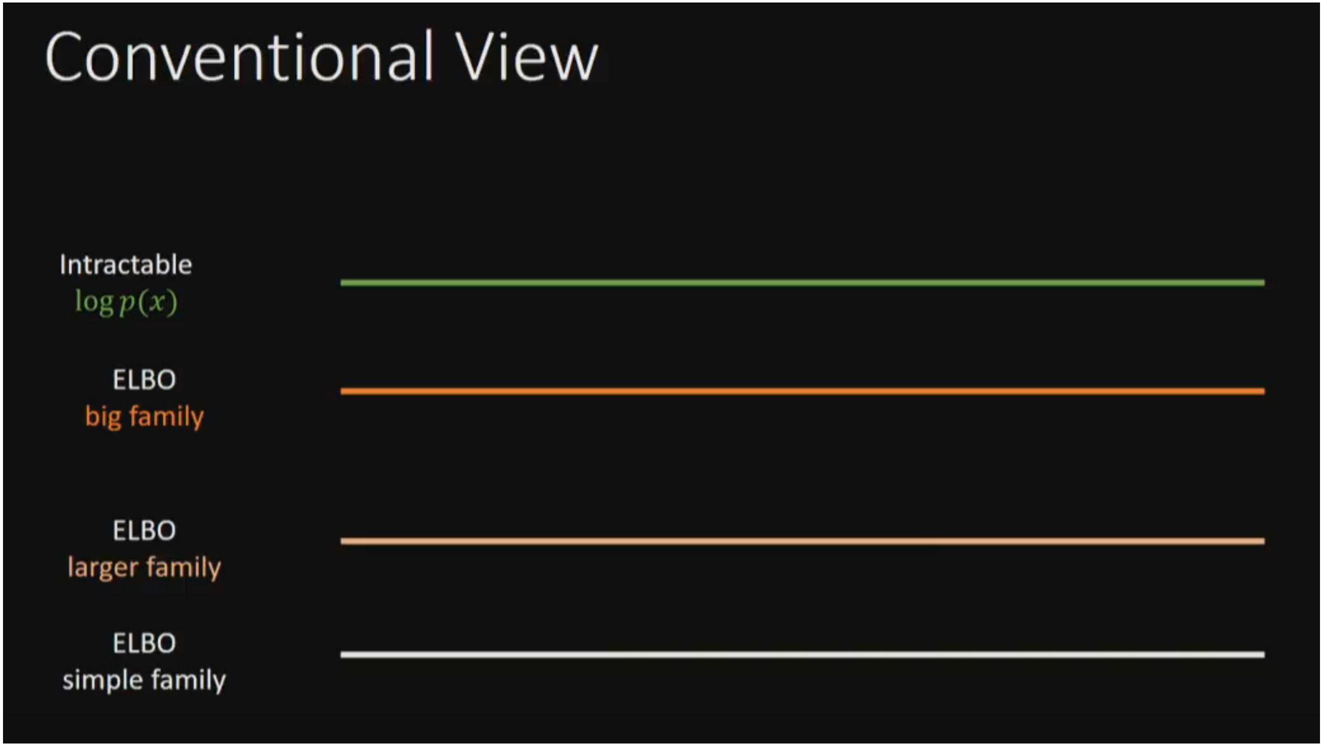
$$KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = E_q[\ln q(\mathbf{z})] - E_q[\ln p(\mathbf{z},\mathbf{x})] + \ln p(\mathbf{x}) \tag{5}$$

$$\ln p(\mathbf{x}) - KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = E_q[\ln p(\mathbf{z},\mathbf{x})] - E_q[\ln q(\mathbf{z})] \tag{6}$$

$$\ln p(\mathbf{x}) - KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = E_q[\ln p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})] \tag{7}$$

Conclusion: ELBO is strict if the variational distribution matches the posterior exactly.

Conventional view of ELBO



From probabilistic view to loss function

By rearranging we reobtain the

$$\mathcal{L}_{VAE} = -\mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] + KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{8}$$

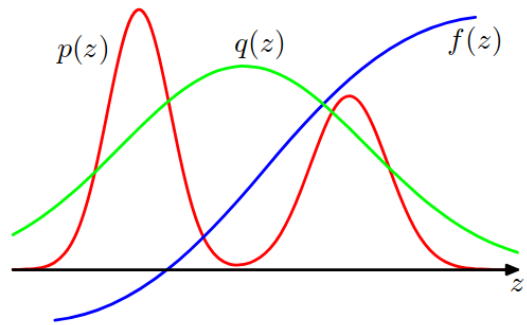
which can be approximated by Monte Carlo estimation (the fanciest expression ever for taking an average of samples)

$$\mathcal{L}_{VAE} = -\frac{1}{L}\sum_{l=1}^{L} \log p(\mathbf{x}|\mathbf{z}) + KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{9}$$

Importance sampling view

What is importance sampling?

Figure 11.8 Importance sampling addresses the problem of evaluating the expectation of a function $f(z)$ with respect to a distribution $p(z)$ from which it is difficult to draw samples directly. Instead, samples $\{z^{(l)}\}$ are drawn from a simpler distribution $q(z)$, and the corresponding terms in the summation are weighted by the ratios $p(z^{(l)})/q(z^{(l)})$.



$$\int 1 \cdot p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z} \tag{10}$$

Exactly our problem: the posterior $p(\mathbf{z}|\mathbf{x})$ is intractable, so we cannot sample from it. However, we could sample from $\mathbf{q}(\mathbf{z}|\mathbf{x})$.

So let's do it!

The expectation we want to have from an importance sampling perspective

$$\log p(\mathbf{x}) = \log \mathbb{E}_q[\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})}] \tag{11}$$

$$\log p(\mathbf{x}) \approx \mathbb{E}_{q\sim z_1} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \mathcal{L}_k \tag{12}$$

from which distributions we can sample from and use the same reparametrisation based backprop as in standard VAE.
Code: to enlighten what this really means in context of automatic differentation.

Importance sampling approaches evidence

Now notice,

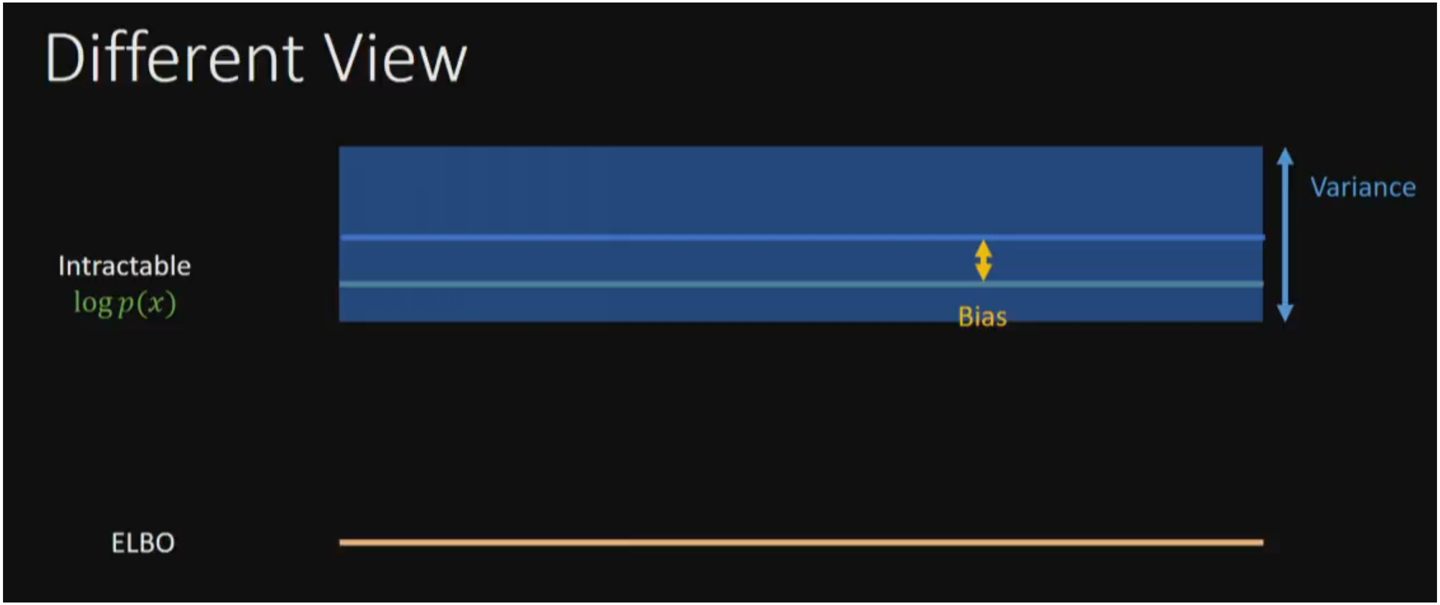
$$\mathcal{L}_1 = \log p(\mathbf{x}) \approx \mathbb{E}_{q\sim z_1} \log \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} = \text{ELBO} \tag{13}$$

and the ELBO is bounded by the log-marginal. It is also possible to show that for all $k \geq m$, $\mathcal{L}_k \geq \mathcal{L}_m$ which means by the strong law of large numbers,

$$\lim_{k\to\infty} \mathcal{L}_k = \log p(\mathbf{x}) \tag{14}$$

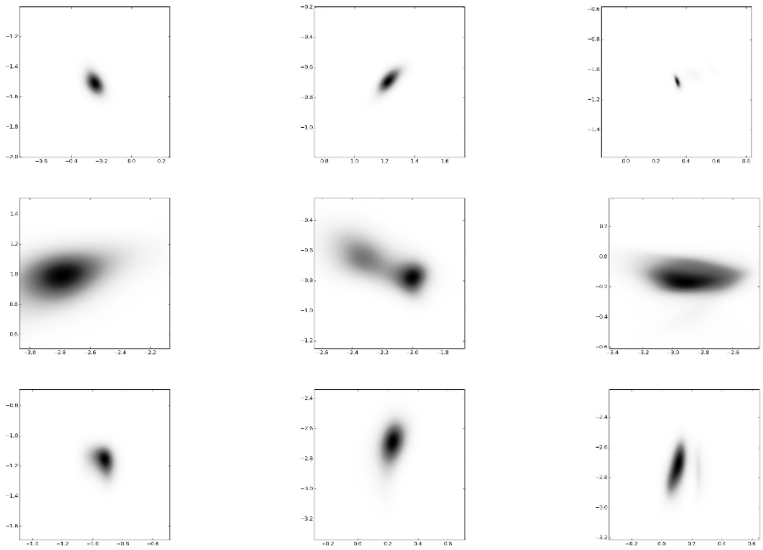
TL;DR: With more samples we obtain **tighter ELBO**, thus we obtain an increasingly better estimate for the evidence.

Estimation view of ELBO



Density results

Left VAE, Middle IWAE $k = 5$, Right IWAE $k = 50$, $q(\mathbf{z}|\mathbf{x})$ for $n = 3$ training examples



References

- C. Bishop: Pattern Recognition and Machine Learning (Chapters on Approximate Inference and Sampling)
- Debiasing Evidence Approximations: IWAE and Jacknife VIhttps://www.youtube.com/watch?v=nRgjvACKNAQ&t=552s
- Various tutorials in VAE
- Y. Burda: Importance Weighted Autoencoders