

# Variational Auto-encoders

Miguel Rios  
University of Amsterdam

April 25, 2019

# Outline

- 1 Variational inference
- 2 Variational auto-encoder
  - Semi supervised VAE
  - Beyond mean field

# The Basic Problem

The marginal likelihood

$$p(x) = \int p(x, z) dz$$

is generally **intractable**, which prevents us from computing quantities that depend on the posterior  $p(z|x)$

- e.g. gradients in MLE
- e.g. predictive distribution in Bayesian modelling

# Strategy

Accept that  $p(z|x)$  is not computable.

# Strategy

Accept that  $p(z|x)$  is not computable.

- approximate it by an auxiliary distribution  $q(z|x)$  that is computable
- choose  $q(z|x)$  as close as possible to  $p(z|x)$  to obtain a faithful approximation

# Evidence lowerbound

$$\log p(x) = \log \int p(x, z) dz$$

# Evidence lowerbound

$$\begin{aligned}\log p(x) &= \log \int p(x, z) dz \\ &= \log \int \textcolor{red}{q(z|x)} \frac{p(x, z)}{\textcolor{red}{q(z|x)}} dz\end{aligned}$$

# Evidence lowerbound

$$\begin{aligned}\log p(x) &= \log \int p(x, z) dz \\ &= \log \int q(z|x) \frac{p(x, z)}{q(z|x)} dz \\ &= \log \left( \mathbb{E}_{q(z|x)} \left[ \frac{p(x, z)}{q(z|x)} \right] \right)\end{aligned}$$



# Evidence lowerbound

$$\begin{aligned}\log p(x) &= \log \int p(x, z) dz \\ &= \log \int \textcolor{red}{q(z|x)} \frac{p(x, z)}{\textcolor{red}{q(z|x)}} dz \\ &= \log \left( \mathbb{E}_{q(z|x)} \left[ \frac{p(x, z)}{q(z|x)} \right] \right) \\ &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}}\end{aligned}$$

# Evidence lowerbound

$$\begin{aligned}\log p(x) &= \log \int p(x, z) dz \\ &= \log \int \textcolor{red}{q(z|x)} \frac{p(x, z)}{\textcolor{red}{q(z|x)}} dz \\ &= \log \left( \mathbb{E}_{q(z|x)} \left[ \frac{p(x, z)}{q(z|x)} \right] \right) \\ &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} [\log p(x, z)] - \mathbb{E}_{q(z|x)} [\log q(z)]\end{aligned}$$

# Evidence lowerbound

$$\begin{aligned}\log p(x) &= \log \int p(x, z) dz \\ &= \log \int q(z|x) \frac{p(x, z)}{q(z|x)} dz \\ &= \log \left( \mathbb{E}_{q(z|x)} \left[ \frac{p(x, z)}{q(z|x)} \right] \right) \\ &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} [\log p(x, z)] - \mathbb{E}_{q(z|x)} [\log q(z)] \\ &= \mathbb{E}_{q(z|x)} [\log p(x, z)] + \mathbb{H}(q(z|x))\end{aligned}$$

# An approximate posterior

$$\log p(x) \geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}}$$

# An approximate posterior

$$\begin{aligned}\log p(x) &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)p(x)}{q(z|x)} \right]\end{aligned}$$

# An approximate posterior

$$\begin{aligned}\log p(x) &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)p(x)}{q(z|x)} \right] \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)}{q(z|x)} \right] + \underbrace{\log p(x)}_{\text{constant}}\end{aligned}$$

# An approximate posterior

$$\begin{aligned}\log p(x) &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)p(x)}{q(z|x)} \right] \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)}{q(z|x)} \right] + \underbrace{\log p(x)}_{\text{constant}} \\ &= - \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{q(z|x)}{p(z|x)} \right]}_{\text{KL}(q(z|x) || p(z|x))} + \log p(x)\end{aligned}$$

# An approximate posterior

$$\begin{aligned}\log p(x) &\geq \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{p(x, z)}{q(z|x)} \right]}_{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)p(x)}{q(z|x)} \right] \\ &= \mathbb{E}_{q(z|x)} \left[ \log \frac{p(z|x)}{q(z|x)} \right] + \underbrace{\log p(x)}_{\text{constant}} \\ &= - \underbrace{\mathbb{E}_{q(z|x)} \left[ \log \frac{q(z|x)}{p(z|x)} \right]}_{\text{KL}(q(z|x) || p(z|x))} + \log p(x)\end{aligned}$$

We have derived a lower bound on the log-evidence whose gap is exactly  $\text{KL}(q(z|x) || p(z|x))$ .



# Variational Inference

## Objective

$$\max_{q(z|x)} \mathbb{E} [\log p(x, z)] + \mathbb{H}(q(z|x))$$

- The ELBO is a lower bound on  $\log p(x)$

# Mean field assumption

Suppose we have  $N$  latent variables

- assume the posterior factorises as  $N$  independent terms
- each with an independent set of parameters

$$q(z_1, \dots, z_N) = \underbrace{\prod_{i=1}^N q_{\lambda_i}(z_i)}_{\text{mean field}}$$

# Amortised variational inference

Amortise the cost of inference using NNs

$$q(z_1, \dots, z_N | x_1, \dots, x_N) = \prod_{i=1}^N q_{\lambda}(z_i | x_i)$$

with a shared set of parameters

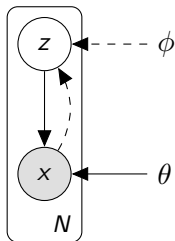
- e.g.  $Z|x \sim \mathcal{N}(\underbrace{\mu_{\lambda}(x), \sigma_{\lambda}(x)^2}_{\text{inference network}})$

# Outline

- 1 Variational inference
- 2 Variational auto-encoder
  - Semi supervised VAE
  - Beyond mean field

# Variational auto-encoder

## Generative model with NN likelihood



- complex (non-linear) observation model  $p_{\theta}(x|z)$
- complex (non-linear) mapping from data to latent variables  $q_{\phi}(z|x)$

Jointly optimise generative model  $p_{\theta}(x|z)$  and inference model  $q_{\phi}(z|x)$  under the same objective (ELBO)

$$\log p_{\theta}(x) \geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}}$$

$$\begin{aligned}\log p_{\theta}(x) &\geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z) + \log p(z)] + \mathbb{H}(q_{\phi}(z|x))\end{aligned}$$

$$\begin{aligned}\log p_{\theta}(x) &\geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z) + \log p(z)] + \mathbb{H}(q_{\phi}(z|x)) \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) \parallel p(z))\end{aligned}$$



$$\begin{aligned}\log p_{\theta}(x) &\geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z) + \log p(z)] + \mathbb{H}(q_{\phi}(z|x)) \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))\end{aligned}$$

Parameter estimation

$$\arg \max_{\theta, \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))$$

$$\begin{aligned}\log p_{\theta}(x) &\geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z) + \log p(z)] + \mathbb{H}(q_{\phi}(z|x)) \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))\end{aligned}$$

Parameter estimation

$$\arg \max_{\theta, \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))$$

- assume  $\text{KL}(q_{\phi}(z|x) || p(z))$  analytical  
true for exponential families

$$\begin{aligned}\log p_{\theta}(x) &\geq \overbrace{\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z)] + \mathbb{H}(q_{\phi}(z|x))}^{\text{ELBO}} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z) + \log p(z)] + \mathbb{H}(q_{\phi}(z|x)) \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))\end{aligned}$$

Parameter estimation

$$\arg \max_{\theta, \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p(z))$$

- assume  $\text{KL}(q_{\phi}(z|x) || p(z))$  analytical  
true for exponential families
- approximate  $\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)]$  by sampling  
true because we design  $q_{\phi}(z|x)$  to be simple

# Generative Network Gradient

$$\frac{\partial}{\partial \theta} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) || p(z))}^{\text{constant wrt } \theta} \right)$$

# Generative Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \theta} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) \parallel p(z))}^{\text{constant wrt } \theta} \right) \\ &= \underbrace{\mathbb{E}_{q_{\phi}(z|x)} \left[ \frac{\partial}{\partial \theta} \log p_{\theta}(x|z) \right]}_{\text{expected gradient :)}} \end{aligned}$$

# Generative Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \theta} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) \parallel p(z))}^{\text{constant wrt } \theta} \right) \\ &= \underbrace{\mathbb{E}_{q_{\phi}(z|x)} \left[ \frac{\partial}{\partial \theta} \log p_{\theta}(x|z) \right]}_{\text{expected gradient :)}} \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \frac{\partial}{\partial \theta} \log p_{\theta}(x|z^{(k)}) \\ &z^{(k)} \sim q_{\phi}(z|x) \end{aligned}$$

# Generative Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \theta} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) \parallel p(z))}^{\text{constant wrt } \theta} \right) \\ &= \underbrace{\mathbb{E}_{q_{\phi}(z|x)} \left[ \frac{\partial}{\partial \theta} \log p_{\theta}(x|z) \right]}_{\text{expected gradient :)}} \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \frac{\partial}{\partial \theta} \log p_{\theta}(x|z^{(k)}) \\ &z^{(k)} \sim q_{\phi}(z|x) \end{aligned}$$

Note:  $q_{\phi}(z|x)$  does not depend on  $\theta$ .

# Inference Network Gradient

$$\frac{\partial}{\partial \phi} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) || p(z))}^{\text{analytical}} \right)$$



# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) || p(z))}^{\text{analytical}} \right) \\ &= \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \underbrace{\frac{\partial}{\partial \phi} \text{KL}(q_{\phi}(z|x) || p(z))}_{\text{analytical computation}} \end{aligned}$$

# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \left( \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \overbrace{\text{KL}(q_{\phi}(z|x) || p(z))}^{\text{analytical}} \right) \\ &= \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \underbrace{\frac{\partial}{\partial \phi} \text{KL}(q_{\phi}(z|x) || p(z))}_{\text{analytical computation}} \end{aligned}$$

The first term again requires approximation by sampling,  
but there is a problem

# Inference Network Gradient

$$\frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)]$$

# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \end{aligned}$$

# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \end{aligned}$$

# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \end{aligned}$$

- MC estimator is non-differentiable: cannot sample first

# Inference Network Gradient

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \end{aligned}$$

- MC estimator is non-differentiable: cannot sample first
- Differentiating the expression does not yield an expectation: cannot approximate via MC

# Score function estimator

We can again use the log identity for derivatives

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \end{aligned}$$



# Score function estimator

We can again use the log identity for derivatives

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \\ &= \int q_{\phi}(z|x) \frac{\partial}{\partial \phi} (\log q_{\phi}(z|x)) \log p_{\theta}(x|z) dz \end{aligned}$$

# Score function estimator

We can again use the log identity for derivatives

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) dz \\ &= \underbrace{\int \frac{\partial}{\partial \phi} (q_{\phi}(z|x)) \log p_{\theta}(x|z) dz}_{\text{not an expectation}} \\ &= \int q_{\phi}(z|x) \frac{\partial}{\partial \phi} (\log q_{\phi}(z|x)) \log p_{\theta}(x|z) dz \\ &= \underbrace{\mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right]}_{\text{expected gradient :)}} \end{aligned}$$

## Score function estimator: high variance

We can now build an MC estimator

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right] \end{aligned}$$

# Score function estimator: high variance

We can now build an MC estimator

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right] \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \log p_{\theta}(x|z^{(k)}) \frac{\partial}{\partial \phi} \log q_{\phi}(z^{(k)}|x) \\ & z^{(k)} \sim q_{\phi}(Z|x) \end{aligned}$$

# Score function estimator: high variance

We can now build an MC estimator

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right] \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \log p_{\theta}(x|z^{(k)}) \frac{\partial}{\partial \phi} \log q_{\phi}(z^{(k)}|x) \\ & z^{(k)} \sim q_{\phi}(Z|x) \end{aligned}$$

but

- magnitude of  $\log p_{\theta}(x|z)$  varies widely

# Score function estimator: high variance

We can now build an MC estimator

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right] \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \log p_{\theta}(x|z^{(k)}) \frac{\partial}{\partial \phi} \log q_{\phi}(z^{(k)}|x) \\ & z^{(k)} \sim q_{\phi}(Z|x) \end{aligned}$$

but

- magnitude of  $\log p_{\theta}(x|z)$  varies widely
- model likelihood does not contribute to direction of gradient

# Score function estimator: high variance

We can now build an MC estimator

$$\begin{aligned} & \frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] \\ &= \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_{\theta}(x|z) \frac{\partial}{\partial \phi} \log q_{\phi}(z|x) \right] \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \log p_{\theta}(x|z^{(k)}) \frac{\partial}{\partial \phi} \log q_{\phi}(z^{(k)}|x) \\ & z^{(k)} \sim q_{\phi}(Z|x) \end{aligned}$$

but

- magnitude of  $\log p_{\theta}(x|z)$  varies widely
- model likelihood does not contribute to direction of gradient
- too much variance to be useful

# When variance is high we can

- sample more



## When variance is high we can

- sample more  
won't scale

## When variance is high we can

- sample more  
won't scale
- use variance reduction techniques (e.g. baselines and control variates)

# When variance is high we can

- sample more  
won't scale
- use variance reduction techniques (e.g. baselines and control variates)
- stare at this  $\frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)]$

# When variance is high we can

- sample more  
won't scale
- use variance reduction techniques (e.g. baselines and control variates)
- stare at this  $\frac{\partial}{\partial \phi} \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)]$   
until we find a way to rewrite the expectation in terms of a density that **does not depend on  $\phi$**

# Reparametrisation

Find a transformation  $h : z \mapsto \epsilon$  that expresses  $z$  through a random variable  $\epsilon$  such that  $q(\epsilon)$  does not depend on  $\phi$

---

(Kingma and Welling, 2013; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014)

# Reparametrisation

Find a transformation  $h : z \mapsto \epsilon$  that expresses  $z$  through a random variable  $\epsilon$  such that  $q(\epsilon)$  does not depend on  $\phi$

- $h(z, \phi)$  needs to be invertible

---

(Kingma and Welling, 2013; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014)

# Reparametrisation

Find a transformation  $h : z \mapsto \epsilon$  that expresses  $z$  through a random variable  $\epsilon$  such that  $q(\epsilon)$  does not depend on  $\phi$

- $h(z, \phi)$  needs to be invertible
- $h(z, \phi)$  needs to be differentiable

---

(Kingma and Welling, 2013; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014)

# Reparametrisation

Find a transformation  $h : z \mapsto \epsilon$  that expresses  $z$  through a random variable  $\epsilon$  such that  $q(\epsilon)$  does not depend on  $\phi$

- $h(z, \phi)$  needs to be invertible
- $h(z, \phi)$  needs to be differentiable

Invertibility implies

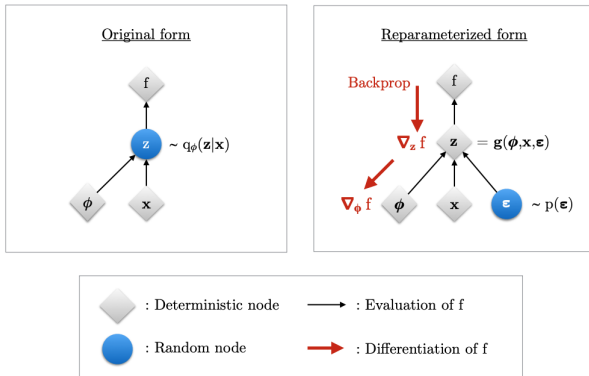
- $h(z, \phi) = \epsilon$
- $h^{-1}(\epsilon, \phi) = z$

---

(Kingma and Welling, 2013; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014)



# Reparametrisation



(Kingma and Welling, 2013)

# Gaussian Transformation

If  $Z \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x)^2)$  then

$$h(z, \phi) = \frac{z - \mu_\phi(x)}{\sigma_\phi(x)} = \epsilon \sim \mathcal{N}(0, 1)$$

$$h^{-1}(\epsilon, \phi) = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \quad \epsilon \sim \mathcal{N}(0, 1)$$

$$= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) \mathrm{d}z$$

$$\begin{aligned} &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) \, dz \\ &= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_{\theta}(x | \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \, d\epsilon \end{aligned}$$

$$\begin{aligned}
 &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) \, dz \\
 &= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_{\theta}(x | \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \, d\epsilon \\
 &= \int q(\epsilon) \frac{\partial}{\partial \phi} \left[ \log p_{\theta}(x | \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \right] \, d\epsilon
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\partial}{\partial \phi} \int q_{\phi}(z|x) \log p_{\theta}(x|z) \, dz \\
 &= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_{\theta}(x | \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \, d\epsilon \\
 &= \int q(\epsilon) \frac{\partial}{\partial \phi} \left[ \log p_{\theta}(x | \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \right] \, d\epsilon \\
 &= \underbrace{\mathbb{E}_{q(\epsilon)} \left[ \frac{\partial}{\partial \phi} \log p_{\theta}(x | h^{-1}(\epsilon, \phi)) \right]}_{\text{expected gradient :D}} \, d\epsilon
 \end{aligned}$$

## Reparametrised gradient estimate

$$= \underbrace{\mathbb{E}_{q(\epsilon)} \left[ \frac{\partial}{\partial \phi} \log p_{\theta}(x|h^{-1}(\epsilon, \phi)) \right]}_{\text{expected gradient :D}} d\epsilon$$

# Reparametrised gradient estimate

$$\begin{aligned} &= \underbrace{\mathbb{E}_{q(\epsilon)} \left[ \frac{\partial}{\partial \phi} \log p_{\theta}(x|h^{-1}(\epsilon, \phi)) \right]}_{\text{expected gradient :D}} d\epsilon \\ &= \mathbb{E}_{q(\epsilon)} \left[ \underbrace{\frac{\partial}{\partial z} \log p_{\theta}(x|\overbrace{h^{-1}(\epsilon, \phi)}^{=z})}_{\text{chain rule}} \times \frac{\partial}{\partial \phi} h^{-1}(\epsilon, \phi) \right] \end{aligned}$$



# Reparametrised gradient estimate

$$\begin{aligned} &= \underbrace{\mathbb{E}_{q(\epsilon)} \left[ \frac{\partial}{\partial \phi} \log p_{\theta}(x|h^{-1}(\epsilon, \phi)) \right]}_{\text{expected gradient :D}} d\epsilon \\ &= \mathbb{E}_{q(\epsilon)} \left[ \underbrace{\frac{\partial}{\partial z} \log p_{\theta}(x|\overset{=z}{h^{-1}(\epsilon, \phi)})}_{\text{chain rule}} \times \frac{\partial}{\partial \phi} h^{-1}(\epsilon, \phi) \right] \\ &\stackrel{\text{MC}}{\approx} \frac{1}{K} \sum_{k=1}^K \underbrace{\frac{\partial}{\partial z} \log p_{\theta}(x|\overset{=z}{h^{-1}(\epsilon^{(k)}, \phi)}) \times \frac{\partial}{\partial \phi} h^{-1}(\epsilon^{(k)}, \phi)}_{\text{backprop's job}} \\ &\epsilon^{(k)} \sim q(\epsilon) \end{aligned}$$

Note that both models contribute with gradients

# Gaussian KL

## ELBO

$$\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL} (q_{\phi}(z|x) || p(z))$$

# Gaussian KL

## ELBO

$$\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL} (q_{\phi}(z|x) || p(z))$$

Analytical computation of  $-\text{KL} (q_{\phi}(z|x) || p(z))$ :

$$\frac{1}{2} \sum_{i=1}^d (1 + \log (\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

# Gaussian KL

## ELBO

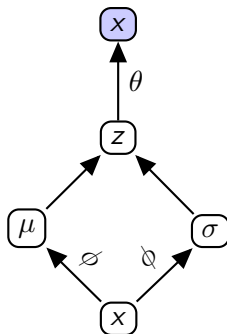
$$\mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL} (q_{\phi}(z|x) || p(z))$$

Analytical computation of  $-\text{KL} (q_{\phi}(z|x) || p(z))$ :

$$\frac{1}{2} \sum_{i=1}^d (1 + \log (\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

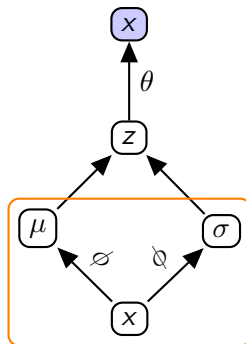
Thus backprop will compute  $-\frac{\partial}{\partial \phi} \text{KL} (q_{\phi}(z|x) || p(z))$  for us

# Computation Graph



# Computation Graph

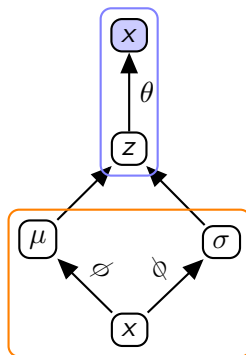
inference model



# Computation Graph

generative model

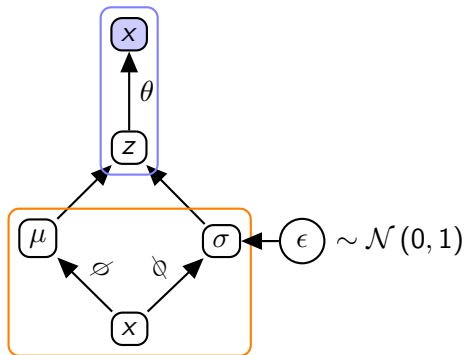
inference model



# Computation Graph

generative model

inference model

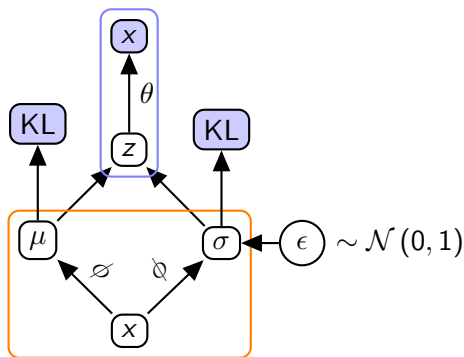




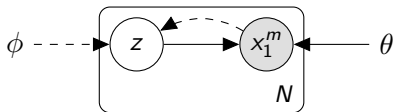
# Computation Graph

generative model

inference model



# Example



Generative model

- $Z \sim \mathcal{N}(0, I)$
- $X_i | z, x_{<i} \sim \text{Cat}(f_\theta(z, x_{<i}))$

Inference model

- $Z \sim \mathcal{N}(\mu_\phi(x_1^m), \sigma_\phi(x_1^m)^2)$

# VAEs – Summary

## Advantages

- Backprop training
- Easy to implement
- Posterior inference possible
- One objective for both NNs

# VAEs – Summary

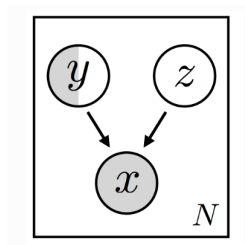
## Advantages

- Backprop training
- Easy to implement
- Posterior inference possible
- One objective for both NNs

## Drawbacks

- Discrete latent variables are difficult
- Optimisation may be difficult with several latent variables
- Location-scale families only  
but see Ruiz et al. (2016) and Kucukelbir et al. (2017)

# Semi-supervised VAE



---

(Kingma et al., 2014)

# Semi-supervised VAE

- Generative model:

$$\begin{aligned}p(y) &= \text{cat}(y|\boldsymbol{\pi}); \\p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}); \\p_{\theta}(\mathbf{x}|y, \mathbf{z}) &= f(\mathbf{x}; y, \mathbf{z}, \boldsymbol{\theta})\end{aligned}\tag{1}$$

- Inference model:

$$\begin{aligned}q_{\phi}(\mathbf{z}|y, \mathbf{x}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi}(y, \mathbf{x}), \text{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}))); \\q_{\phi}(y|\mathbf{x}) &= \text{Cat}(y|\boldsymbol{\pi}_{\phi}(\mathbf{x}))\end{aligned}\tag{2}$$

# Objective

- Labelled data:

$$\log p_{\theta}(\mathbf{x}, y) \geq \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}, y)} [\log p_{\theta}(\mathbf{x}|y, \mathbf{z}) + \log p_{\theta}(y) + \log p(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}, y)] = \quad (3)$$

- Unlabelled data:

$$\begin{aligned} \log p_{\theta}(\mathbf{x}) &\geq \mathbb{E}_{q_{\phi}(y, \mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|y, \mathbf{z}) + \log p_{\theta}(y) + \log p(\mathbf{z}) - \log q_{\phi}(y, \mathbf{z}|\mathbf{x})] \\ &= \sum_y q_{\phi}(y|\mathbf{x}) (-\mathcal{L}(\mathbf{x}, y)) + \mathcal{H}(q_{\phi}(y|\mathbf{x})) = -\mathcal{U}(\mathbf{x}) \end{aligned} \quad (4)$$

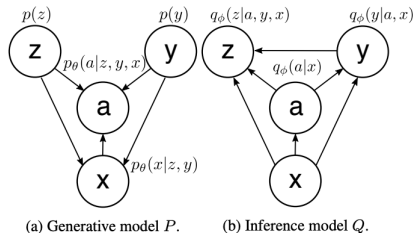
# Beyond the mean field





# Auxiliary variable

The mean field assumption might result in models that do not capture all dependencies in the observations:



(Maaløe et al., 2016)

# Auxiliary variable

- Generative model:

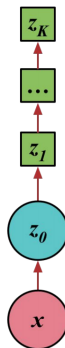
$$\begin{aligned}p(z) &= \mathcal{N}(z|0, I) \\p(y) &= \text{Cat}(y|\pi) \\p_{\theta}(a|z, y, x) &= f(a; z, y, x, \theta) \\p_{\theta}(x|z, y) &= f(x; z, y, \theta)\end{aligned}\tag{5}$$

- Inference model:

$$\begin{aligned}q_{\phi}(a|x) &= \mathcal{N}(a|\mu_{\phi}(x), \text{diag}(\sigma_{\phi}^2(x))) \\q_{\phi}(y|a, x) &= \text{Cat}(y|\pi_{\phi}(a, x)) \\q_{\phi}(z|a, y, x) &= \mathcal{N}(z|\mu_{\phi}(a, y, x), \text{diag}(\sigma_{\phi}^2(a, y, x)))\end{aligned}\tag{6}$$

# Normalizing flow

Normalising Flows



# NF for NLP

- (Pelsmaecker and Aziz, 2019) tackle issues present in VAE models for language.
- Annealing
- Expressive posterior

# Summary

## Deep learning in NLP

- task-driven feature extraction
- models with more realistic assumptions

## Probabilistic modelling

- better (or at least more explicit) statistical assumptions
- compact models
- semi-supervised learning

# Literature I

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. 01 2016. URL <https://arxiv.org/abs/1601.00670>.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/K16-1002>.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. 2013. URL <http://arxiv.org/abs/1312.6114>.

## Literature II

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc., 2014. URL <http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf>.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. URL <http://jmlr.org/papers/v18/16-107.html>.

## Literature III

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1445–1453, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/maaloe16.html>.

Tom Pelsmaeker and Wilker Aziz. Effective estimation of deep generative language models. *arXiv preprint arXiv:1904.08194*, 2019.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014. URL <http://jmlr.org/proceedings/papers/v32/rezende14.pdf>.



## Literature IV

Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. The generalized reparameterization gradient. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, pages 460–468. 2016. URL <http://papers.nips.cc/paper/6328-the-generalized-reparameterization-gradient.pdf>.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Tony Jebara and Eric P. Xing, editors, *ICML*, pages 1971–1979, 2014. URL <http://jmlr.org/proceedings/papers/v32/titsias14.pdf>.