Variational Auto-encoders

Miguel Rios
University of Amsterdam

April 22, 2019

Outline

The Basic Problem

The marginal likelihood

$$p(x) = \int p(x, z)\mathrm{d}z$$

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

$$\log p(x) = \log \int p(x, z)\mathrm{d}z$$

Evidence lowerbound

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

Evidence lowerbound

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

Evidence lowerbound

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

Evidence lowerbound

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

Evidence lowerbound

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

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

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

$$= \mathbb{E}_{q(z|x)}\left[\log \frac{p(z|x)p(x)}{q(z|x)}\right]$$

An approximate posterior

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

$$= \mathbb{E}_{q(z|x)}\left[\log \frac{p(z|x)p(x)}{q(z|x)}\right]$$

$$= \mathbb{E}_{q(z|x)}\left[\log \frac{p(z|x)}{q(z|x)}\right] + \underbrace{\log p(x)}_{\text{constant}}$$

An approximate posterior

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

An approximate posterior

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

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

Variational Inference

Objective

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

- 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

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

Amortised variational inference

Amortise the cost of inference using NNs

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

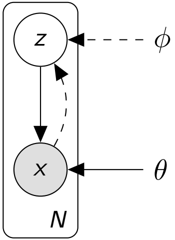
with a shared set of parameters

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

Outline

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)

?

Objective

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

Objective

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

Objective

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

Objective

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

Parameter estimation

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

Objective

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

Parameter estimation

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

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

Objective

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

Parameter estimation

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

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

Generative Network Gradient

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

Generative Network Gradient

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

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x|z)\right]}_{\text{expected gradient :)}}$$

Generative Network Gradient

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

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\frac{\partial}{\partial \theta}\log p_\theta(x|z)\right]}_{\text{expected gradient :)}}$$

$$\overset{\text{MC}}{\approx} \frac{1}{K}\sum_{k=1}^{K}\frac{\partial}{\partial \theta}\log p_\theta(x|z^{(k)})$$

$$z^{(k)} \sim q_\phi(z|x)$$

Generative Network Gradient

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

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x|z)\right]}_{\text{expected gradient :)}}$$

$$\overset{\text{MC}}{\approx} \frac{1}{K}\sum_{k=1}^{K}\frac{\partial}{\partial\theta}\log p_\theta(x|z^{(k)})$$

$$z^{(k)} \sim q_\phi(z|x)$$

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

Inference Network Gradient

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

Inference Network Gradient

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

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

Inference Network Gradient

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

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

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

Inference Network Gradient

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

Inference Network Gradient

$$\frac{\partial}{\partial \phi} \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z) \right]$$
$$= \frac{\partial}{\partial \phi} \int q_\phi(z|x) \log p_\theta(x|z) \mathrm{d}z$$

Inference Network Gradient

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

Inference Network Gradient

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

- MC estimator is non-differentiable: cannot sample first

Inference Network Gradient

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

- 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

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

Score function estimator

We can again use the log identity for derivatives

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

Score function estimator

We can again use the log identity for derivatives

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

Score function estimator: high variance

We can now build an MC estimator

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

Score function estimator: high variance

We can now build an MC estimator

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

Score function estimator: high variance

We can now build an MC estimator

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

but

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

Score function estimator: high variance

We can now build an MC estimator

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

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

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

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)
  excellent idea, but not just yet

When variance is high we can

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

Reparametrisation

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

(???)

Reparametrisation

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

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

(???)

Reparametrisation

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

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

(???)

Reparametrisation

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

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

Invertibility implies

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

(???)

Gaussian Transformation

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

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

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

Inference Network – Reparametrised Gradient

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

Inference Network – Reparametrised Gradient

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

$$= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_{\theta}(x| \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \, \mathrm{d}\epsilon$$

Inference Network – Reparametrised Gradient

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

$$= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_\theta(x| \overbrace{h^{-1}(\epsilon, \phi)}^{=z}) \, \mathrm{d}\epsilon$$

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

Inference Network – Reparametrised Gradient

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

$$= \frac{\partial}{\partial \phi} \int q(\epsilon) \log p_\theta(x|\overbrace{h^{-1}(\epsilon, \phi)}^{=z})\, \mathrm{d}\epsilon$$

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

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

Reparametrised gradient estimate

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

Reparametrised gradient estimate

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

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

Reparametrised gradient estimate

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

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

$$\overset{\text{MC}}{\approx} \frac{1}{K}\sum_{k=1}^{K}\underbrace{\frac{\partial}{\partial z}\log p_\theta(x|\overbrace{h^{-1}(\epsilon^{(k)},\phi)}^{=z}) \times \frac{\partial}{\partial \phi}h^{-1}(\epsilon^{(k)},\phi)}_{\text{backprop's job}}$$

$$\epsilon^{(k)} \sim q(\epsilon)$$

Note that both models contribute with gradients

Gaussian KL

ELBO

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

Gaussian KL

ELBO

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

Analytical computation of $-\mathrm{KL}\left(q_\phi(z|x) \mid\mid p(z)\right)$:

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

Gaussian KL

ELBO

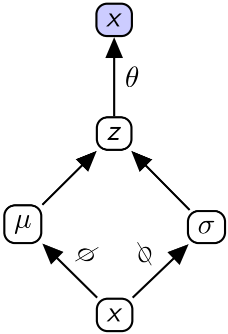
$$\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}\left(q_\phi(z|x) \mid\mid p(z)\right)$$

Analytical computation of $-\mathrm{KL}\left(q_\phi(z|x) \mid\mid p(z)\right)$:

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

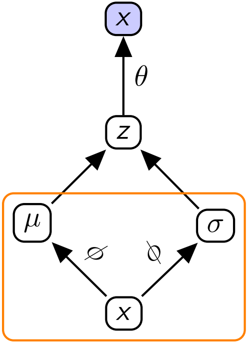
Thus backprop will compute $-\frac{\partial}{\partial\phi}\mathrm{KL}\left(q_\phi(z|x) \mid\mid p(z)\right)$ for us

Computation Graph



Computation Graph

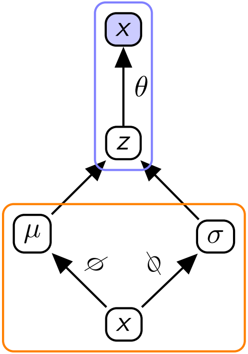
inference model



Computation Graph

generative model

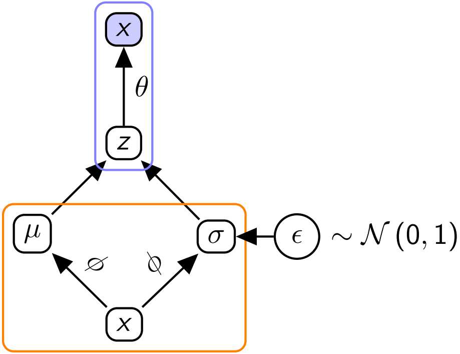
inference model



Computation Graph

generative model

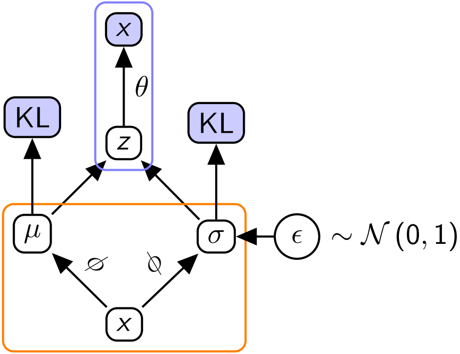
inference model



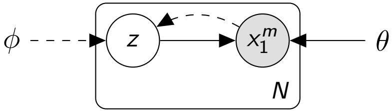
Computation Graph

generative model

inference model



Example



Generative model

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

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 **?** and **?**

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