Deep Generative Models: Continuous Latent Variables

Philip Schulz and Wilker Aziz
https://github.com/philschulz/VITutorial

Deep Generative Models

First Attempt: Wake-Sleep

This is how we do: Variational Autoencoders

Deep Generative Models

First Attempt: Wake-Sleep

This is how we do: Variational Autoencoders

Generative Models

Joint distribution over observed data x and latent variables Z.

$$p(x, z|\theta) = \underbrace{p(z)}_{\text{prior}}\,\underbrace{p(x|z, \theta)}_{\text{likelihood}}$$

The likelihood and prior are often standard distributions (Gaussian, Bernoulli) with simple dependence on conditioning information.

Deep generative models

Joint distribution with **deep observation model**

$$p(x, z|\theta) = \underbrace{p(z)}_{\text{prior}} \underbrace{p(x|z,\theta)}_{\text{likelihood}}$$

mapping from z to $p(x|z,\theta)$ is a NN with parameters θ

Deep generative models

Joint distribution with **deep observation model**

$$p(x, z|\theta) = \underbrace{p(z)}_{\text{prior}} \underbrace{p(x|z, \theta)}_{\text{likelihood}}$$

mapping from z to $p(x|z, \theta)$ is a NN with parameters θ

Marginal likelihood

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

intractable in general

Goals

We want

- richer probabilistic models

Goals

We want

- richer probabilistic models
- complex observation models parameterised by NNs

Goals

We want

- richer probabilistic models
- complex observation models
  parameterised by NNs

but we can't perform gradient-based MLE

Goals

We want

- richer probabilistic models
- complex observation models parameterised by NNs

but we can't perform gradient-based MLE

We need approximate inference techniques!

Deep Generative Models

First Attempt: Wake-Sleep

This is how we do: Variational Autoencoders

Wake-sleep Algorithm

- Generalise latent variables to Neural Networks
- Train generative neural model
- Use variational inference! (kind of)

Wake-sleep Architecture

2 Neural Networks:

Wake-sleep Architecture

2 Neural Networks:

- A generation network to model the data (the one we want to optimise) – parameters: θ

Wake-sleep Architecture

2 Neural Networks:

- A generation network to model the data (the one we want to optimise) – parameters: θ
- An inference (recognition) network (to model the latent variable) – parameters: λ

Wake-sleep Architecture

2 Neural Networks:

- A generation network to model the data (the one we want to optimise) – parameters: θ
- An inference (recognition) network (to model the latent variable) – parameters: λ
- Original setting: binary hidden units

Wake-sleep Architecture

2 Neural Networks:

- A generation network to model the data (the one we want to optimise) – parameters: θ
- An inference (recognition) network (to model the latent variable) – parameters: λ
- Original setting: binary hidden units
- Training is performed in a "hard EM" fashion

Wake-sleep Training

Wake Phase

- Use inference network to sample hidden unit setting z from $q(z|x, \lambda)$
- Update generation parameters θ to maximize join log-liklelihood of data and latents $p(x, z|\theta)$

Wake-sleep Training

Wake Phase

- Use inference network to sample hidden unit setting z from $q(z|x, \lambda)$
- Update generation parameters θ to maximize join log-liklelihood of data and latents $p(x, z|\theta)$

Sleep Phase

- Produce dream sample $\tilde{x}$ from random hidden unit z
- Update inference parameters λ to maximize probability of latent state $q(z|\tilde{x}, \lambda)$

Wake Phase Objective

Objective

$$\min_{\theta} \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right)$$

Wake Phase Objective

Objective

$$\min_{\theta} \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right)$$

$$= \max_{\theta} \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{G}(\theta)}$$

Wake Phase Objective

Objective

$$\min_{\theta} \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right)$$
$$= \max_{\theta} \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{G}(\theta)}$$

Gradient estimate

$$\boldsymbol{\nabla}_{\theta}\mathcal{G}(\theta) = \mathbb{E}_{q(z|x,\lambda)}\left[\boldsymbol{\nabla}_{\theta} \log p(z,x|\theta)\right] + \boldsymbol{\nabla}_{\theta}\mathbb{H}[q(z|x,\lambda)]$$

Wake Phase Objective

Objective

$$\min_{\theta} \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z|x,\theta)\right)$$
$$= \max_{\theta} \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{G}(\theta)}$$

Gradient estimate

$$\nabla_{\theta}\mathcal{G}(\theta) = \mathbb{E}_{q(z|x,\lambda)}\left[\nabla_{\theta} \log p(z,x|\theta)\right] + \nabla_{\theta}\mathbb{H}[q(z|x,\lambda)]$$
$$\overset{\mathrm{MC}}{\approx} \nabla_{\theta} \log p(z,x|\theta)$$

with z drawn from $q(z|x,\lambda)$

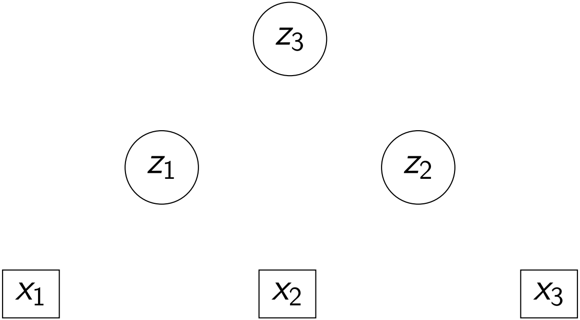
Wake Phase Objective

Assumes latent state z to be fixed random draws from $q(z|x, \lambda)$.

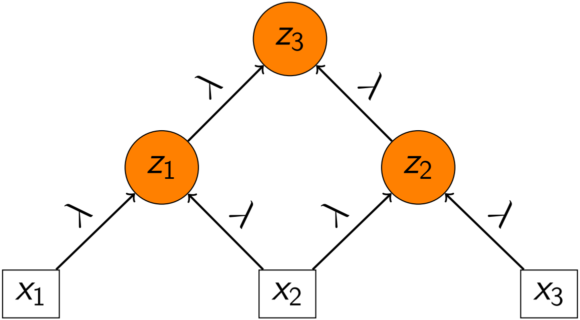
$$\min_{\theta} \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right)$$
$$\overset{\text{MC}}{\approx} \max_{\theta} \log p(z, x|\theta)$$

This is simply supervised learning with imputed latent data!

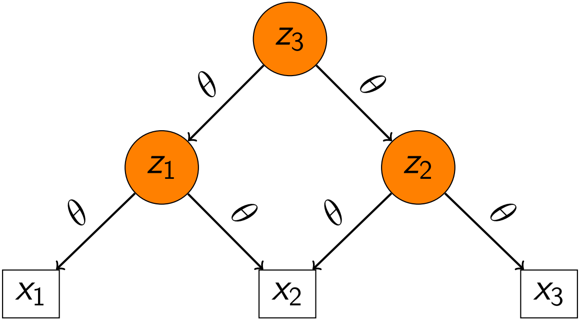
Wake Phase Sampling



Wake Phase Sampling



Wake Phase Update



Sleep Phase Objective

Objective

$$\min_{\lambda} \mathrm{KL}\left(q(z|x,\lambda) \;||\; p(z|x,\theta)\right)$$

Sleep Phase Objective

Objective

$$\min_{\lambda} \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right)$$
$$= \max_{\lambda} \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{R}(\lambda)}$$

Sleep Phase Objective

Objective

$$\min_{\lambda} \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z|x,\theta)\right)$$
$$= \max_{\lambda} \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{R}(\lambda)}$$

Gradient estimate

$$\boldsymbol{\nabla}_{\lambda}\mathcal{R}(\lambda) = \boldsymbol{\nabla}_{\lambda}\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \boldsymbol{\nabla}_{\lambda}\mathbb{H}[q(z|x,\lambda)]$$

Sleep Phase Objective

Objective

$$\begin{aligned}\min_{\lambda} &\, \mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z|x,\theta)\right) \\ = \max_{\lambda} &\; \underbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \mathbb{H}[q(z|x,\lambda)]}_{\mathcal{R}(\lambda)}\end{aligned}$$

Gradient estimate

$$\boldsymbol{\nabla}_{\lambda}\mathcal{R}(\lambda) = \boldsymbol{\nabla}_{\lambda}\mathbb{E}_{q(z|x,\lambda)}\left[\log p(z,x|\theta)\right] + \boldsymbol{\nabla}_{\lambda}\mathbb{H}[q(z|x,\lambda)]$$

Let's change the objective!

Sleep Phase Objective

Assumes fake data $\tilde{x}$ and latent variables z to be fixed random draw from $p(x, z|\theta)$.

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

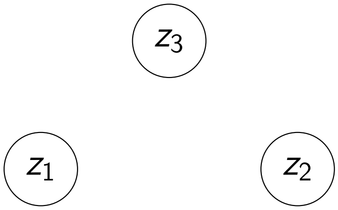
Sleep Phase Objective

Assumes fake data $\tilde{x}$ and latent variables z to be fixed random draw from $p(x, z|\theta)$.

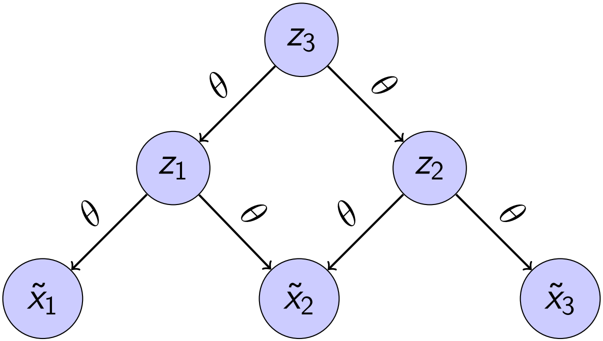
$$\max_{\lambda} \; \mathbb{E}_{p(\tilde{x},z|\theta)} [\log q(z|\tilde{x}, \lambda)] + \mathbb{E}_{p(\tilde{x})} [\mathbb{H}(p(z|\tilde{x}, \theta))]$$

$$\stackrel{\text{MC}}{\approx} \max_{\lambda} \; \log q(z|\tilde{x}, \lambda)$$

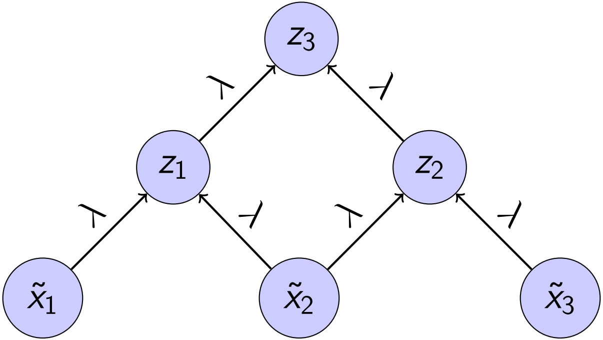
Sleep Phase Sampling



Sleep Phase Sampling



Sleep Phase Update



Wake-sleep Algorithm

Advantages

- Simple layer-wise updates
- Amortised inference: all latent variables are inferred from the same weights λ

Wake-sleep Algorithm

Advantages

- Simple layer-wise updates
- Amortised inference: all latent variables are inferred from the same weights λ

Drawbacks

- Inference and generative networks are trained on different objectives
- Inference weights λ are updated on fake data $\tilde{x}$
- Generative weights are bad initially, giving wrong signal to the updates of λ

Deep Generative Models

First Attempt: Wake-Sleep

This is how we do: Variational Autoencoders

Generative Model with NN Likelihood

Goal

Define model $p(x, z|\theta) = p(x|z, \theta)p(z)$ where the likelihood $p(x|z, \theta)$ is given by a neural network. (We fix $p(z)$ for simplicity.)

Generative Model with NN Likelihood

Goal

Define model $p(x, z|\theta) = p(x|z, \theta)p(z)$ where the likelihood $p(x|z, \theta)$ is given by a neural network. (We fix $p(z)$ for simplicity.)

Problem

$p(x) = \int p(x|z, \theta)p(z)\mathrm{d}z$ is hard to compute.

Generative Model with NN Likelihood

Goal

Define model $p(x, z|\theta) = p(x|z, \theta)p(z)$ where the likelihood $p(x|z, \theta)$ is given by a neural network. (We fix $p(z)$ for simplicity.)

Problem

$p(x) = \int \underbrace{p(x|z, \theta)}_{\text{highly non-linear!}} p(z)\mathrm{d}z$ is hard to compute.

Solution: Variational Inference

$$\log p(x|\theta) \geq \overbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x, Z|\theta)\right] + \mathbb{H}\left(q(z|x,\lambda)\right)}^{\text{ELBO}}$$

Solution: Variational Inference

$$\begin{aligned}\log p(x|\theta) &\geq \overbrace{\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x, Z|\theta)\right] + \mathbb{H}\left(q(z|x,\lambda)\right)}^{\text{ELBO}} \\ &= \mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|Z,\theta) + \log p(Z)\right] + \mathbb{H}\left(q(z|x,\lambda)\right)\end{aligned}$$

Solution: Variational Inference

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

Solution: Variational Inference

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

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

Solution: Variational Inference

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

$$
\arg\max_{\theta,\lambda} \mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|Z,\theta)\right] - \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)
$$

- assume $\mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)$ analytical
  true for exponential families

Solution: Variational Inference

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

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

- assume $\mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z)\right)$ analytical
  true for exponential families
- approximate $\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right]$ by sampling
  feasible because $q(z|x,\lambda)$ is simple

Generator Network Gradient

$$\frac{\partial}{\partial \theta} \mathbb{E}_{q(z|x,\lambda)} \left[\log p(x|z,\theta) \right] - \overbrace{\mathrm{KL}\left(q(z|x,\lambda) \;||\; p(z) \right)}^{constant}$$

Generator Network Gradient

$$\frac{\partial}{\partial \theta}\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \overbrace{\mathrm{KL}\left(q(z|x,\lambda) \,||\, p(z)\right)}^{constant}$$
$$= \mathbb{E}_{q(z|x,\lambda)}\left[\frac{\partial}{\partial \theta}\log p(x|z,\theta)\right]$$

Generator Network Gradient

$$\frac{\partial}{\partial\theta}\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \overbrace{\mathrm{KL}\left(q(z|x,\lambda)\ ||\ p(z)\right)}^{\textit{constant}}$$

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

$$\overset{\text{MC}}{\approx} \frac{1}{S}\sum_{i=1}^{S}\frac{\partial}{\partial\theta}\log p(x|z_i,\theta)$$

where $z_i \sim q(z|x,\lambda)$

Generator Network Gradient

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

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

$$\overset{\mathrm{MC}}{\approx} \frac{1}{S} \sum_{i=1}^{S} \frac{\partial}{\partial \theta} \log p(x|z_i,\theta)$$

where $z_i \sim q(z|x,\lambda)$

Note: $q(z|x,\lambda)$ does not depend on θ.

Inference Network Gradient

$$\frac{\partial}{\partial \lambda}\left[\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)\right]$$

Inference Network Gradient

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda}\left[\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)\right] \\
=&\frac{\partial}{\partial \lambda}\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \underbrace{\frac{\partial}{\partial \lambda}\mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)}_{\text{analytical computation}}
\end{aligned}
$$

Inference Network Gradient

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

The first term again requires approximation by sampling

Inference Network Gradient

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

Inference Network Gradient

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

Inference Network Gradient

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

Inference Network Gradient

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

Not an expected gradient!

Inference Network Gradient

Reparametrisation trick

Find a transformation $h : z \mapsto \epsilon$ such that ϵ does not depend on λ.

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

Inference Network Gradient

Reparametrisation trick

Find a transformation $h : z \mapsto \epsilon$ such that ϵ does not depend on λ.

- $h(z, \lambda)$ needs to be invertible
- $h(z, \lambda)$ needs to be differentiable
- $h(z, \lambda) = \epsilon$
- $h^{-1}(\epsilon, \lambda) = z$

Gaussian Transformation

Gaussian Transformation

Affine property

$$Az + b \sim \mathcal{N}\left(\mu + b, A\Sigma A^T\right) \text{ for } z \sim \mathcal{N}\left(\mu, \Sigma\right)$$

Gaussian Transformation

Affine property

$$Az + b \sim \mathcal{N}\left(\mu + b, A\Sigma A^T\right) \text{ for } z \sim \mathcal{N}(\mu, \Sigma)$$

Special case

$$Az + b \sim \mathcal{N}\left(b, AA^T\right) \text{ for } z \sim \mathcal{N}(0, \mathrm{I})$$

Gaussian Transformation

Affine property

$$Az + b \sim \mathcal{N}\left(\mu + b, A\Sigma A^T\right) \text{ for } z \sim \mathcal{N}\left(\mu, \Sigma\right)$$

Special case

$$Az + b \sim \mathcal{N}\left(b, AA^T\right) \text{ for } z \sim \mathcal{N}\left(0, \mathrm{I}\right)$$

Gaussian transformation

$$h(z, \lambda) = \frac{z - \mu(\phi, \lambda)}{\sigma(\phi, \lambda)} = \epsilon \sim \mathcal{N}\left(0, \mathrm{I}\right)$$

$$\underbrace{h^{-1}(\epsilon, \lambda)}_{=z} = \mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon \quad \epsilon \sim \mathcal{N}\left(0, \mathrm{I}\right)$$

Inference Network Gradient

$$= \frac{\partial}{\partial \lambda} \int q(z|x, \lambda) \log p(x|z, \theta) \mathrm{d}z$$

Inference Network Gradient

$$
\begin{aligned}
&= \frac{\partial}{\partial \lambda} \int q(z|x, \lambda) \log p(x|z, \theta) \mathrm{d}z \\
&= \frac{\partial}{\partial \lambda} \int q(\epsilon) \log \left(p(x| \overbrace{h^{-1}(\epsilon, \lambda)}^{=z}, \theta) \right) \mathrm{d}\epsilon
\end{aligned}
$$

Inference Network Gradient

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

Inference Network Gradient

$$\mathbb{E}_{q(\epsilon)}\left[\frac{\partial}{\partial\lambda}\log p(x|\overbrace{h^{-1}(\epsilon,\lambda)}^{=z},\theta)\right]$$

Inference Network Gradient

$$\mathbb{E}_{q(\epsilon)}\left[\frac{\partial}{\partial \lambda} \log p(x|\overbrace{h^{-1}(\epsilon, \lambda)}^{=z}, \theta)\right]$$

$$\stackrel{\text{MC}}{\approx} \frac{1}{S}\sum_{i=1}^{S} \frac{\partial}{\partial \lambda} \log p(x|\overbrace{h^{-1}(\epsilon_i, \lambda)}^{=z}, \theta)$$

where $\epsilon_i \sim q(\epsilon)$

Inference Network Gradient

$$\mathbb{E}_{q(\epsilon)}\left[\frac{\partial}{\partial\lambda}\log p(x|\overbrace{h^{-1}(\epsilon,\lambda)}^{=z},\theta)\right]$$

$$\overset{\text{MC}}{\approx}\frac{1}{S}\sum_{i=1}^{S}\frac{\partial}{\partial\lambda}\log p(x|\overbrace{h^{-1}(\epsilon_i,\lambda)}^{=z},\theta)$$

where $\epsilon_i \sim q(\epsilon)$

$$\overset{\text{MC}}{\approx}\frac{1}{S}\sum_{i=1}^{S}\underbrace{\frac{\partial}{\partial z}\log p(x|\overbrace{h^{-1}(\epsilon_i,\lambda)}^{=z},\theta)\times\frac{\partial}{\partial\lambda}h^{-1}(\epsilon_i,\lambda)}_{\text{chain rule}}$$

Derivatives of Gaussian transformation

Recall:

$$h^{-1}(\epsilon, \lambda) = \mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon \ .$$

We get two gradient paths!

Derivatives of Gaussian transformation

Recall:

$$h^{-1}(\epsilon, \lambda) = \mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon \ .$$

We get two gradient paths!

- one is deterministic
  $\frac{\partial h^{-1}(\epsilon,\lambda)}{\partial \mu(\phi,\lambda)} = \frac{\partial}{\partial \mu(\phi,\lambda)} [\mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon] = 1$

Derivatives of Gaussian transformation

Recall:

$$h^{-1}(\epsilon, \lambda) = \mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon \ .$$

We get two gradient paths!

- one is deterministic
  $\frac{\partial h^{-1}(\epsilon,\lambda)}{\partial \mu(\phi,\lambda)} = \frac{\partial}{\partial \mu(\phi,\lambda)}[\mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon] = 1$
- the other is stochastic
  $\frac{\partial h^{-1}(\epsilon,\lambda)}{\partial \sigma(\phi,\lambda)} = \frac{\partial}{\partial \sigma(\phi,\lambda)}[\mu(\phi, \lambda) + \sigma(\phi, \lambda) \odot \epsilon] = \epsilon$

Gaussian KL

ELBO

$$\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)$$

Gaussian KL

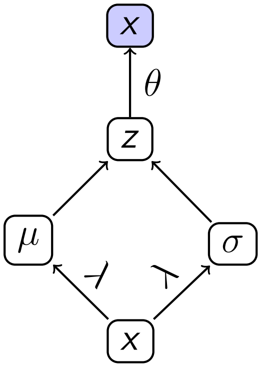
ELBO

$$\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)$$

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

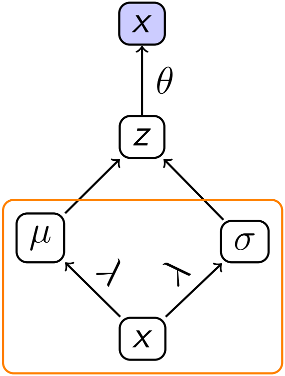
$$\frac{1}{2}\sum_{i=1}^{N}\left(1 + \log\left(\sigma_i^2\right) - \mu_i^2 - \sigma_i^2\right)$$

Computation Graph



Computation Graph

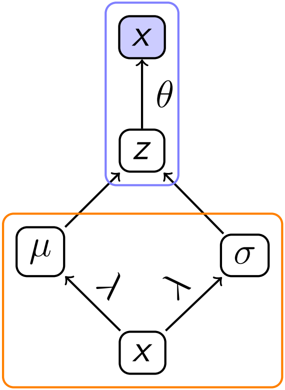
inference model



Computation Graph

generation model

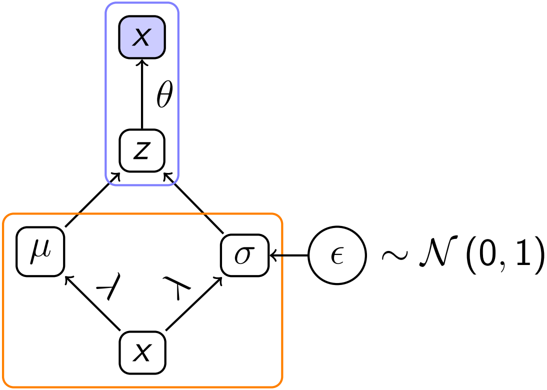
inference model



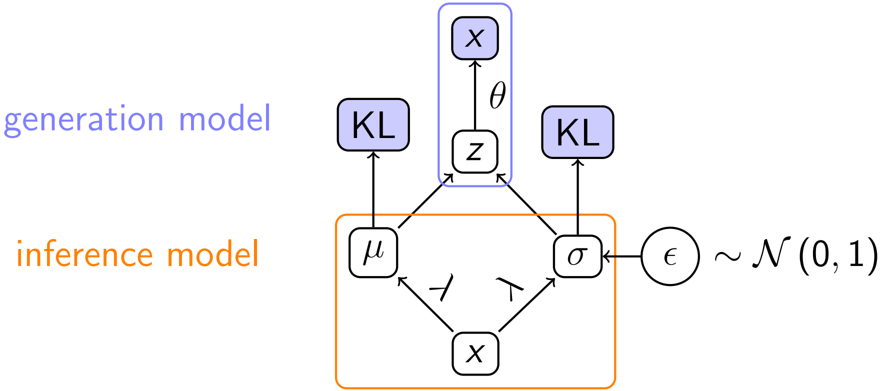
Computation Graph

generation model

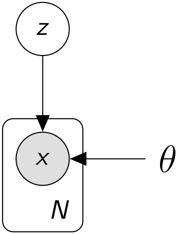
inference model



Computation Graph



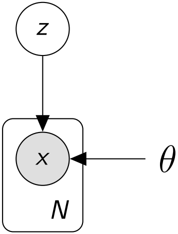
Example: Unigram Document Model



Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \text{Cat}(f(z, \theta))$

Example: Unigram Document Model

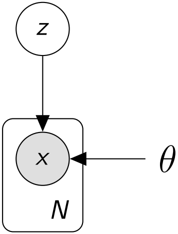


Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \text{Cat}(f(z, \theta))$

Designing $f(z, \theta)$

Example: Unigram Document Model



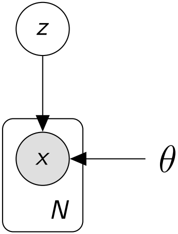
Generative story

- Draw a document embedding $Z \sim \mathcal{N}(0, I)$
- Draw N words $X_i | z \sim \mathrm{Cat}(f(z, \theta))$

Designing $f(z, \theta)$

$$h = \mathrm{relu}(W_1 z + b_1)$$

Example: Unigram Document Model



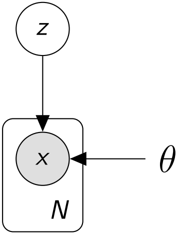
Generative story

- Draw a document embedding $Z \sim \mathcal{N}(0, I)$
- Draw N words $X_i | z \sim \text{Cat}(f(z, \theta))$

Designing $f(z, \theta)$

$$h = \text{relu}(W_1 z + b_1)$$
$$f(z, \theta) = \text{softmax}(W_2 h + b_2)$$

Example: Unigram Document Model



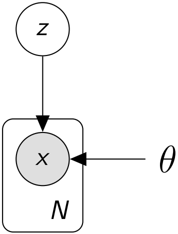
Generative story

- Draw a document embedding $Z \sim \mathcal{N}(0, I)$
- Draw N words $X_i | z \sim \mathrm{Cat}(f(z, \theta))$

Designing $f(z, \theta)$

$$h = \mathrm{relu}(W_1 z + b_1)$$
$$f(z, \theta) = \mathrm{softmax}(W_2 h + b_2)$$
$$\theta = \{W_1, b_1, W_2, b_2\}$$

Example: Unigram Document Model

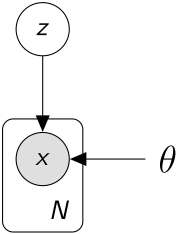


Likelihood

Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \text{Cat}(f(z, \theta))$

Example: Unigram Document Model



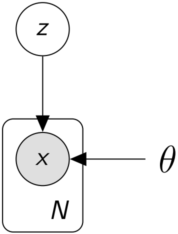
Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \mathrm{Cat}(f(z, \theta))$

Likelihood

$$p(x_1^N | z, \theta) = \prod_{i=1}^{N} p(x_i | z, \theta)$$

Example: Unigram Document Model



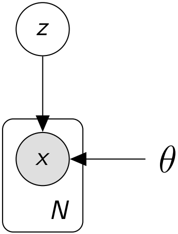
Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \text{Cat}(f(z, \theta))$

Likelihood

$$p(x_1^N | z, \theta) = \prod_{i=1}^{N} p(x_i | z, \theta) = \prod_{i=1}^{N} \text{Cat}(x_i | \underbrace{f(z, \theta)}_{=\psi})$$

Example: Unigram Document Model



Generative story

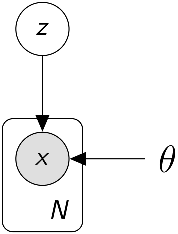
- Draw a document embedding $Z \sim \mathcal{N}(0, I)$
- Draw N words $X_i | z \sim \text{Cat}(f(z, \theta))$

Likelihood

$$p(x_1^N | z, \theta) = \prod_{i=1}^{N} p(x_i | z, \theta) = \prod_{i=1}^{N} \text{Cat}(x_i | \underbrace{f(z, \theta)}_{=\psi})$$

$$= \prod_{i=1}^{N} \psi_{x_i}$$

Example: Unigram Document Model

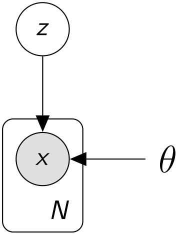


Marginal

Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i|z \sim \mathrm{Cat}(f(z, \theta))$

Example: Unigram Document Model



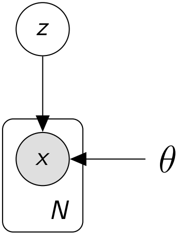
Generative story

- Draw a document embedding
  $Z \sim \mathcal{N}(0, I)$
- Draw N words
  $X_i | z \sim \text{Cat}(f(z, \theta))$

Marginal

$$p(x_1^N | \theta) = \int p(z) \prod_{i=1}^{N} p(x_i | z, \theta) \, \mathrm{d}z$$

Example: Unigram Document Model



Generative story

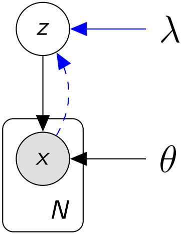
- Draw a document embedding $Z \sim \mathcal{N}(0, I)$
- Draw N words $X_i|z \sim \mathrm{Cat}(f(z,\theta))$

Marginal

$$p(x_1^N|\theta) = \int p(z) \prod_{i=1}^{N} p(x_i|z,\theta)\,\mathrm{d}z$$

$$= \int \mathcal{N}(z|0,I) \prod_{i=1}^{N} \mathrm{Cat}(x_i|f(z,\theta))\,\mathrm{d}z$$

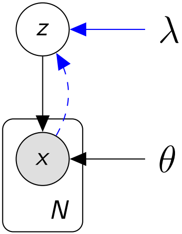
Example: Unigram Document Model



Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Example: Unigram Document Model

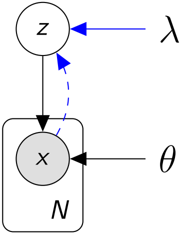


Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Designing the *inference network*

Example: Unigram Document Model



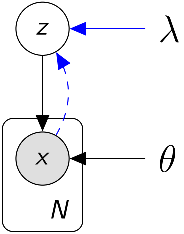
Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Designing the *inference network*

$$s = \sum_{i=1}^{N} E_{x_i}$$

Example: Unigram Document Model



Inference model

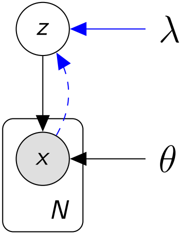
- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Designing the *inference network*

$$s = \sum_{i=1}^{N} E_{x_i}$$

$$h = \mathrm{relu}(M_1 s + c_1)$$

Example: Unigram Document Model



Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

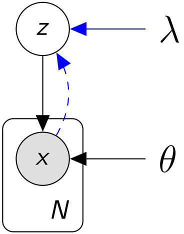
Designing the *inference network*

$$s = \sum_{i=1}^{N} E_{x_i}$$

$$\mu(x_1^N, \lambda) = M_2 h + c_2$$

$$h = \text{relu}(M_1 s + c_1)$$

Example: Unigram Document Model



Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Designing the *inference network*

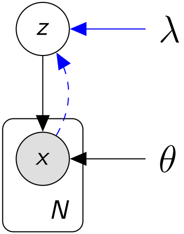
$$s = \sum_{i=1}^{N} E_{x_i}$$

$$h = \text{relu}(M_1 s + c_1)$$

$$\mu(x_1^N, \lambda) = M_2 h + c_2$$

$$\sigma(x_1^N, \lambda) = \text{softplus}(M_3 h + c_3)$$

Example: Unigram Document Model



Inference model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

Designing the *inference network*

$$s = \sum_{i=1}^{N} E_{x_i}$$

$$h = \text{relu}(M_1 s + c_1)$$

$$\mu(x_1^N, \lambda) = M_2 h + c_2$$

$$\sigma(x_1^N, \lambda) = \text{softplus}(M_3 h + c_3)$$

$$\lambda = \{E, M_1^3, c_1^3\}$$

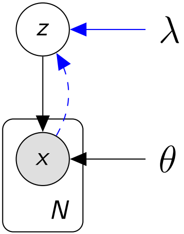
Example: Unigram Document Model

Generative Model

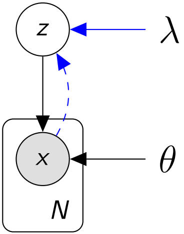
- Prior: $Z \sim \mathcal{N}(0, I)$
- Likelihood: $X_i | z \sim \mathrm{Cat}(f(z, \theta))$

Inference Model

- $Z | x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$



Example: Unigram Document Model



Generative Model

- Prior: $Z \sim \mathcal{N}(0, I)$
- Likelihood: $X_i|z \sim \text{Cat}(f(z, \theta))$

Inference Model

- $Z|x_1^N \sim \mathcal{N}(\mu(x_1^N, \lambda), \sigma(x_1^n, \lambda)^2)$

ELBO

$$\log p(x_1^N|\theta) \geq \mathbb{E}_{\frac{z-u}{s} \sim \mathcal{N}(0,I)} \left[\sum_{i=1}^N \log \psi_{x_i} \right] - \text{KL}\left(\mathcal{N}(z|u, s^2) \,||\, \mathcal{N}(z|0, I) \right)$$

where $u = \mu(x_1^N, \lambda)$, $s = \sigma(x_1^N, \lambda)$, and $\psi = f(z, \theta)$

Aside

If your likelihood model is able to express dependencies between the output variables (e.g. an RNN), the model may simply ignore the latent code. In that case one often scales the KL term. The scale factor is increased gradually.

$$\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \beta\, \mathrm{KL}\left(q(z|x,\lambda) \mid\mid p(z)\right)$$

where $\beta \to 1$.

Variational Autoencoder

Advantages

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

Variational Autoencoder

Advantages

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

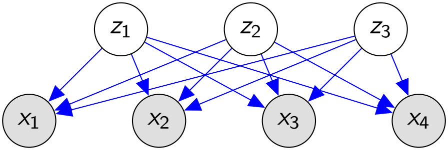
Drawbacks

- Discrete latent variables are difficult
- Optimisation may be difficult with several latent variables

What about amortised inference?

Joint distribution: latent variables are marginally independent a priori

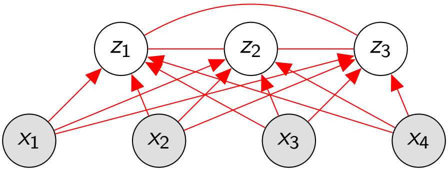
for example, $K = 3, N = 4$



What about amortised inference?

Joint distribution: latent variables are marginally independent a priori

for example, $K = 3, N = 4$



Posterior: latent variables are marginally dependent given observations

Mean field assumption

We have K latent variables

- assume the posterior factorises as K independent terms

$$q(z_1, \ldots, z_K) = \underbrace{\prod_{j=1}^{K} q_{\lambda_j}(z_j)}_{\text{mean field}}$$

Mean field assumption

We have K latent variables

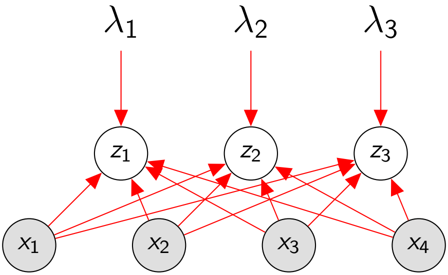
- assume the posterior factorises as K independent terms

$$q(z_1, \ldots, z_K) = \underbrace{\prod_{j=1}^{K} q_{\lambda_j}(z_j)}_{\text{mean field}}$$

with independent sets of parameters $\lambda_j = \{\mu_j, \sigma_j\}$

$$Z_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

Mean field: example



Amortised variational inference

Amortise the cost of inference using NNs

$$q(z_1, \ldots, z_K|x) = \prod_{j=1}^{K} q_{\lambda}(z_j|x)$$

Amortised variational inference

Amortise the cost of inference using NNs

$$q(z_1, \ldots, z_K|x) = \prod_{j=1}^{K} q_{\lambda}(z_j|x)$$

still mean field

$$Z_j|x \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

Amortised variational inference

Amortise the cost of inference using NNs

$$q(z_1, \ldots, z_K|x) = \prod_{j=1}^{K} q_{\lambda}(z_j|x)$$

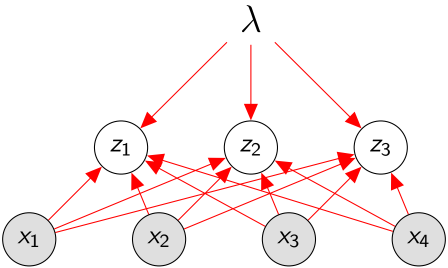
still mean field

$$Z_j|x \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

but with a shared set of parameters

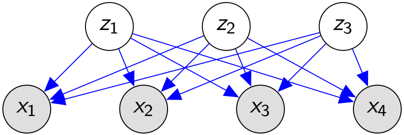
- where $\mu_1^K, \sigma_1^K = g_{\lambda}(x)$

Amortised VI: example

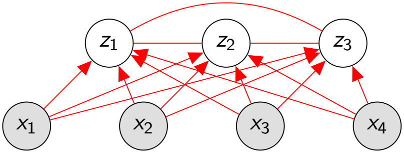


Overview

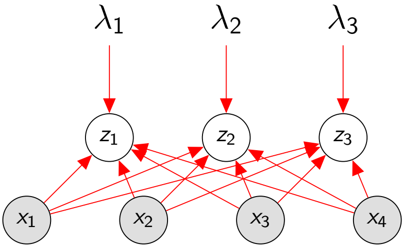
Joint distribution



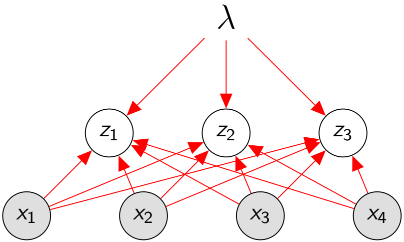
Posterior



Mean field



Amortised VI



Summary

- Wake-Sleep: train inference and generation networks with separate objectives
- VAE: train both networks with same objective
- Reparametrisation
  - Transform parameter-free variable ϵ into latent value z
  - Update parameters with stochastic gradient estimates

Literature I

G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995. URL http://www.gatsby.ucl.ac.uk/~dayan/papers/hdfn95.pdf.

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

Literature II

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.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014. URL

Literature III

http://jmlr.org/proceedings/papers/v32/rezende14.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.