

# Variational Inference: The Foundations

Bryan Eikema and Wilker Aziz

VI Tutorial @ University of Alicante

<https://vitutorial.github.io/tour/ua2020>

# This class is about approximate inference

- probabilistic models with latent variables often have intractable marginal and posterior
- inference in a probabilistic context means *computation*, it involves computing/infering quantities by manipulation of probability calculus
- we will discuss one class of approximate inference algorithms known as *variational inference* (VI)

# 1 Generative Models

## 2 Examples

## 3 Variational Inference

- Deriving VI with Jensen's Inequality
- Deriving VI from KL Divergence
- Relationship to EM

## 4 Mean Field Inference

# Joint Distribution

Let  $X$  and  $Z$  be random variables. A generative model is any model that defines a joint distribution over these variables.

# Joint Distribution

Let  $X$  and  $Z$  be random variables. A generative model is any model that defines a joint distribution over these variables.

## 3 Examples of Generative Models

- $p(x, z) = p(x)p(z)$

# Joint Distribution

Let  $X$  and  $Z$  be random variables. A generative model is any model that defines a joint distribution over these variables.

## 3 Examples of Generative Models

- $p(x, z) = p(x)p(z)$
- $p(x, z) = p(z)p(x|z)$

# Joint Distribution

Let  $X$  and  $Z$  be random variables. A generative model is any model that defines a joint distribution over these variables.

## 3 Examples of Generative Models

- $p(x, z) = p(x)p(z)$
- $p(x, z) = p(z)p(x|z)$
- $p(x, z) = p(x)p(z|x)$

# Likelihood and prior

From here on,  $x$  is our observed data. On the other hand,  $z$  is an unobserved outcome.

- $p(x|z)$  is the **likelihood**
- $p(z)$  is the **prior** over  $Z$

Notice: both distributions may depend on a non-random quantity  $\alpha$ , we write e.g.  $p(z|\alpha)$  and call  $\alpha$  a hyperparameter.



# Bayes rule

We can *invert* a conditional probability distribution.

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

# Bayes rule

We can *invert* a conditional probability distribution.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{p(x)}$$

# Bayes rule

We can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{p(x)}$$

# Bayes rule

We can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{marginal likelihood/evidence}}}$$

# The Basic Problem

We want to compute the posterior over latent variables  $p(z|x)$ . This involves computing the marginal likelihood

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

which is often **intractable**. This problem motivates the use of **approximate inference** techniques.

# Bayesian Inference

The evidence becomes even harder to compute because  $\theta$  is often high-dimensional (just think of neural nets!).

- $p(x|\theta) = \int p(x, z|\theta)dz$  (frequentist)
- $p(x) = \int \int p(x, z, \theta)dzd\theta$  (Bayesian)

# Bayesian Inference

The evidence becomes even harder to compute because  $\theta$  is often high-dimensional (just think of neural nets!).

- $p(x|\theta) = \int p(x, z|\theta)dz$  (frequentist)
- $p(x) = \int \int p(x, z, \theta)dzd\theta$  (Bayesian)

Today we will mostly focus on the frequentist case!

1 Generative Models

2 Examples

3 Variational Inference

- Deriving VI with Jensen's Inequality
- Deriving VI from KL Divergence
- Relationship to EM

4 Mean Field Inference

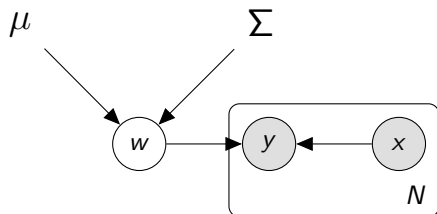


# We cannot compute the posterior when

- 1 The functional form of the posterior is unknown (we don't know which parameters to infer)
- 2 The functional form is known but the computation is intractable

# Bayesian Log-Linear Model

$$p(y|x, \mu, \Sigma) = \int \frac{\exp(w_y^\top x)}{\sum_c \exp(w_c^\top x)} \mathcal{N}(w|\mu, \Sigma) dw$$



The Normal distribution is not conjugate to the Categorical distribution. The form of the posterior is unknown.

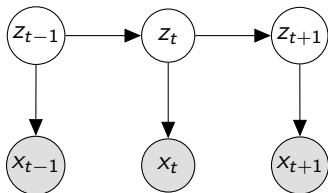
# Bayesian Log-Linear Model

## Intuition

Simply assume that the posterior is Gaussian.

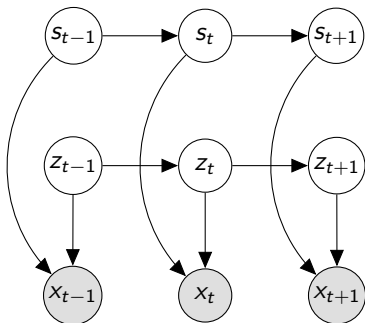
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



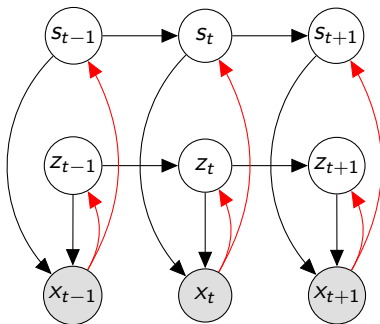
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



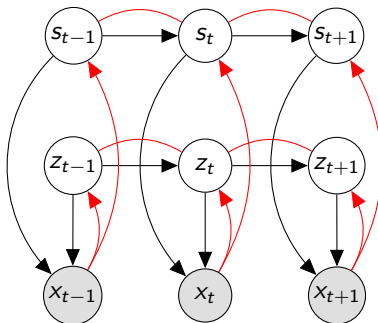
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



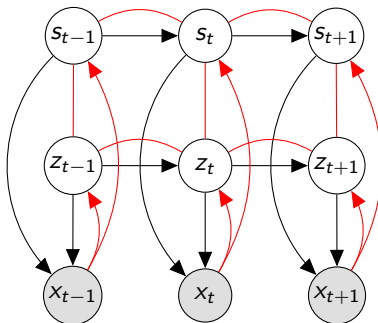
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



# Factorial HMMs

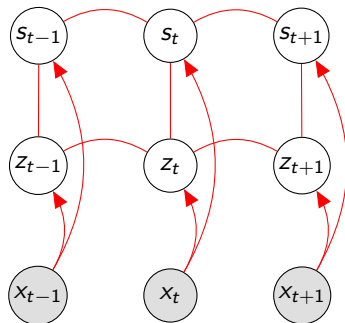
FHMMs have several Markov chains over latent variables.





# Factorial HMMs

Inference network for FHMMs.



# Factorial HMMs

FHMMs have several Markov chains over latent variables.

- $M$  Markov chains over latent variables.
- $L$  outcomes per latent variable.
- Sequence of length  $T$ .
- Complexity of inference:  $\mathcal{O}(L^{2M}T)$ .

# Factorial HMMs

FHMMs have several Markov chains over latent variables.

- $M$  Markov chains over latent variables.
- $L$  outcomes per latent variable.
- Sequence of length  $T$ .
- Complexity of inference:  $\mathcal{O}(L^{2M}T)$ .

## Intractable

Exponential dependency on the number of hidden Markov chains.

# Factorial HMMs

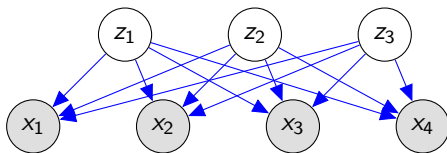
## Intuition

Simply assume that the posterior consists of independent Markov chains.

# Latent Factor Model

**Joint distribution:** latent variables are marginally independent a priori

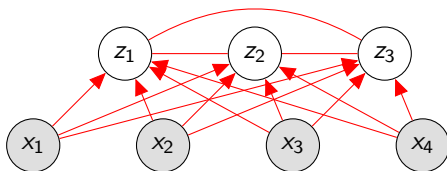
for example,  $K = 3, N = 4$



# Latent Factor Model

**Joint distribution:** latent variables are marginally independent a priori

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



**Posterior:** latent variables are conditionally dependent

# Latent Factor Model

Latent binary variables that together produce an output.

- $N$  output variables (e.g. pixels, words, sentences).
- $K$  binary factors (usually much less than  $N$ ).
- Complexity of inference:  $\mathcal{O}(2^K)$ .

# Latent Factor Model

## Intuition

Simply assume that the posterior consists of independent Bernoulli variables.



# Latent Factor Model

## Intuition

Simply assume that the posterior consists of independent Bernoulli variables.

## Rule of Thumb

Simply assume that the posterior is in the same family as the prior.

1 Generative Models

2 Examples

3 Variational Inference

- Deriving VI with Jensen's Inequality
- Deriving VI from KL Divergence
- Relationship to EM

4 Mean Field Inference

# The Goal

Assume  $p(z|x)$  is not computable.

# The Goal

Assume  $p(z|x)$  is not computable.

## Idea

Let's approximate it by an auxiliary distribution  $q(z)$  that is computable!

# The Goal

Assume  $p(z|x)$  is not computable.

## Idea

Let's approximate it by an auxiliary distribution  $q(z)$  that is computable!

## Requirement

Choose  $q(z)$  as close as possible to  $p(z|x)$  to obtain a faithful approximation.

# Recap KL divergence

The Kullback-Leibler divergence (or relative entropy) measures the divergence of a distribution  $q$  from a distribution  $p$ .

# Recap KL divergence

The Kullback-Leibler divergence (or relative entropy) measures the divergence of a distribution  $q$  from a distribution  $p$ .

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

# Recap KL divergence

The Kullback-Leibler divergence (or relative entropy) measures the divergence of a distribution  $q$  from a distribution  $p$ .

- $\text{KL}(q(z) \parallel p(z|x)) = \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right]$
- $\text{KL}(q(z) \parallel p(z|x)) = \int q(z) \log \frac{q(z)}{p(z|x)} dz$   
(continuous)



# Recap KL divergence

The Kullback-Leibler divergence (or relative entropy) measures the divergence of a distribution  $q$  from a distribution  $p$ .

- $\text{KL}(q(z) \parallel p(z|x)) = \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right]$
- $\text{KL}(q(z) \parallel p(z|x)) = \int q(z) \log \frac{q(z)}{p(z|x)} dz$   
(continuous)
- $\text{KL}(q(z) \parallel p(z|x)) = \sum_z q(z) \log \frac{q(z)}{p(z|x)}$  (discrete)

# Recap KL divergence

## Properties

- $\text{KL}(q(z) \parallel p(z|x)) \geq 0$  with equality iff  $q(z) = p(z|x)$ .

# Recap KL divergence

## Properties

- $\text{KL}(q(z) \parallel p(z|x)) \geq 0$  with equality iff  $q(z) = p(z|x)$ .
- $-\text{KL}(q(z) \parallel p(z|x)) = \mathbb{E}_{q(z)} \left[ \log \frac{p(z|x)}{q(z)} \right] \leq 0$ .

# Recap KL divergence

## Properties

- $\text{KL}(q(z) \parallel p(z|x)) \geq 0$  with equality iff  $q(z) = p(z|x)$ .
- $-\text{KL}(q(z) \parallel p(z|x)) = \mathbb{E}_{q(z)} \left[ \log \frac{p(z|x)}{q(z)} \right] \leq 0$ .
- We want:  $\text{supp}(q) \subseteq \text{supp}(p)$ ; otherwise  $\text{KL}(q(z) \parallel p(z|x)) = \infty$

# VI derivation I

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

# VI derivation I

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

# VI derivation I

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

# VI derivation I

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



# VI derivation I

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

# VI derivation I

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

# VI derivation I

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

# VI derivation I

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

# VI derivation I

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

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

# VI derivation II

Recall that we want to find  $q(z)$  such that  $\text{KL}(q(z) \parallel p(z|x))$  is small.

# VI derivation II

Recall that we want to find  $q(z)$  such that  $\text{KL}(q(z) \parallel p(z|x))$  is small.

## Formal Objective

$$\arg \min_{q(z)} \text{KL}(q(z) \parallel p(z|x))$$

# VI derivation II

Recall that we want to find  $q(z)$  such that  $\text{KL}(q(z) \parallel p(z|x))$  is small.

## Formal Objective

$$\begin{aligned} & \arg \min_{q(z)} \text{KL}(q(z) \parallel p(z|x)) \\ &= \arg \max_{q(z)} - \text{KL}(q(z) \parallel p(z|x)) \end{aligned}$$



# VI derivation II

$$\arg \max_{q(z)} - \text{KL} (q(z) || p(z|x))$$

# VI derivation II

$$\begin{aligned} & \arg \max_{q(z)} - \text{KL} (q(z) || p(z|x)) \\ &= \arg \max_{q(z)} \int q(z) \log \frac{p(z|x)}{q(z)} dz \end{aligned}$$

# VI derivation II

$$\begin{aligned} & \arg \max_{q(z)} - \text{KL} (q(z) \parallel p(z|x)) \\ &= \arg \max_{q(z)} \int q(z) \log \frac{p(z|x)}{q(z)} dz \\ &= \arg \max_{q(z)} \int q(z) \log \frac{p(z, x)}{p(x)q(z)} dz \end{aligned}$$

# VI derivation II

$$\begin{aligned}
 & \arg \max_{q(z)} - \text{KL} (q(z) \parallel p(z|x)) \\
 &= \arg \max_{q(z)} \int q(z) \log \frac{p(z|x)}{q(z)} dz \\
 &= \arg \max_{q(z)} \int q(z) \log \frac{p(z, x)}{p(x)q(z)} dz \\
 &= \arg \max_{q(z)} \int q(z) \log p(z, x) dz - \int q(z) \log q(z) dz - \overbrace{\log p(x)}^{\text{constant}}
 \end{aligned}$$

# VI derivation II

$$\begin{aligned}
 & \arg \max_{q(z)} - \text{KL} (q(z) \parallel p(z|x)) \\
 &= \arg \max_{q(z)} \int q(z) \log \frac{p(z|x)}{q(z)} dz \\
 &= \arg \max_{q(z)} \int q(z) \log \frac{p(z, x)}{p(x)q(z)} dz \\
 &= \arg \max_{q(z)} \int q(z) \log p(z, x) dz - \int q(z) \log q(z) dz - \overbrace{\log p(x)}^{\text{constant}} \\
 &= \arg \max_{q(z)} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H} (q(z))
 \end{aligned}$$

As before, we have derived a lower bound on the log-evidence. This **evidence lower bound** or **ELBO** is our optimisation objective.

## ELBO

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

# Performing VI (Frequentist Case)

## Variational Objective

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

This finds us the best posterior approximation for a **given model**.

# Performing VI (Frequentist Case)

## Variational Objective

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

This finds us the best posterior approximation for a **given model**.

## Also optimize the model!

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



# Performing VI (Frequentist Case)

VI in its basic form can be performed via coordinate ascent. This can be done as a 2-step procedure.

# Performing VI (Frequentist Case)

VI in its basic form can be performed via coordinate ascent. This can be done as a 2-step procedure.

- 1 Maximize (regularised) expected log-density.

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

# Performing VI (Frequentist Case)

VI in its basic form can be performed via coordinate ascent. This can be done as a 2-step procedure.

- 1 Maximize (regularised) expected log-density.

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

- 2 Optimise generative model.

$$\arg \max_{p(x, z)} \mathbb{E}_{q(z)} [\log p(x, z)] + \underbrace{\mathbb{H}(q(z))}_{\text{constant}}$$

# Unconstrained (exact) optimisation

What's the solution to the following?

$$\arg \max_{q(z) \in \mathcal{Q}} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H}(q(z))$$

(assume  $\mathcal{Q}$  is large enough a family)

# Unconstrained (exact) optimisation

What's the solution to the following?

$$\arg \max_{q(z) \in \mathcal{Q}} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H}(q(z))$$

(assume  $\mathcal{Q}$  is large enough a family)

The true posterior  $p(z|x)$ ! Exactly because

$$\arg \max_{q(z) \in \mathcal{Q}} \text{ELBO} = \arg \min_{q(z) \in \mathcal{Q}} \text{KL}(q(z) \parallel p(z|x))$$

and KL is never negative and 0 iff  $q(z) = p(z|x)$ .

# Recap: EM Algorithm

$$\begin{aligned} \text{E-step} \quad & \arg \max_{q(z)} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H}(p(z|x)) \\ & = p(z|x) \end{aligned}$$

# Recap: EM Algorithm

$$\begin{aligned} \text{E-step} \quad & \arg \max_{q(z)} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H}(p(z|x)) \\ & = p(z|x) \end{aligned}$$

$$\text{M-step} \quad \arg \max_{p(x, z)} \mathbb{E}_{p(z|x)} [\log p(x, z)] + \underbrace{\mathbb{H}(p(z|x))}_{\text{constant}}$$

# Recap: EM Algorithm

$$\begin{aligned} \text{E-step} \quad & \arg \max_{q(z)} \mathbb{E}_{q(z)} [\log p(x, z)] + \mathbb{H}(p(z|x)) \\ & = p(z|x) \end{aligned}$$

$$\text{M-step} \quad \arg \max_{p(x, z)} \mathbb{E}_{p(z|x)} [\log p(x, z)] + \underbrace{\mathbb{H}(p(z|x))}_{\text{constant}}$$

EM is variational inference!

$$\begin{aligned} q(z) &= p(z|x) \\ \text{KL}(q(z) || p(z|x)) &= 0 \end{aligned}$$



1 Generative Models

2 Examples

3 Variational Inference

- Deriving VI with Jensen's Inequality
- Deriving VI from KL Divergence
- Relationship to EM

4 Mean Field Inference

# Designing a tractable approximation

- Recall: The approximation  $q(z)$  needs to be tractable.
- Common solution: make **all** latent variables independent under  $q(z)$ .

# Designing a tractable approximation

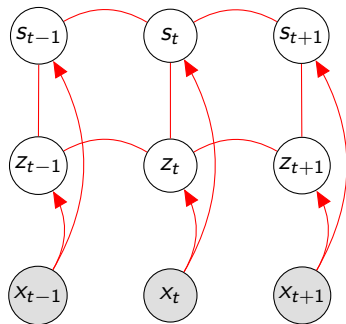
- Recall: The approximation  $q(z)$  needs to be tractable.
- Common solution: make **all** latent variables independent under  $q(z)$ .
- Formal assumption:  $q(z) = \prod_{i=1}^N q(z_i)$

# Designing a tractable approximation

- Recall: The approximation  $q(z)$  needs to be tractable.
- Common solution: make **all** latent variables independent under  $q(z)$ .
- Formal assumption:  $q(z) = \prod_{i=1}^N q(z_i)$

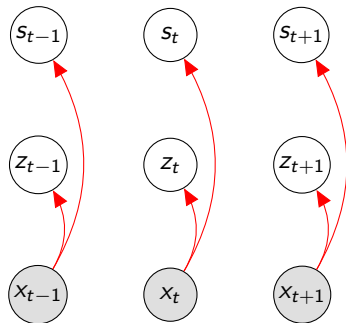
This approximation strategy is commonly known as **mean field** approximation.

# Original FHMM Inference



Exact posterior  $p(s, z|x)$

# Mean field FHMM Inference

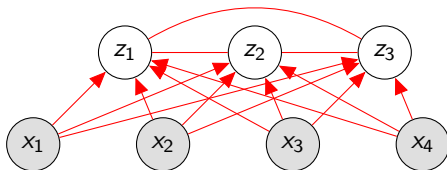


Approximate posterior  $q(s, z) = \prod_{t=1}^T q(s_t)q(z_t)$

# Original Latent Factor Model Inference

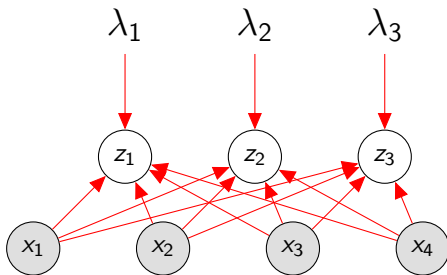
**Joint distribution:** latent variables are marginally independent a priori

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



**Posterior:** latent variables are marginally dependent given observations

# Mean Field Latent Factor Model Inference



$$Z_j \sim \text{Bernoulli}(\lambda_j)$$



# Amortised variational inference

Amortise the cost of inference using NNs

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

# Amortised variational inference

Amortise the cost of inference using NNs

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

still mean field

$$Z_j | x \sim \text{Bernoulli}(b_j)$$

# Amortised variational inference

Amortise the cost of inference using NNs

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

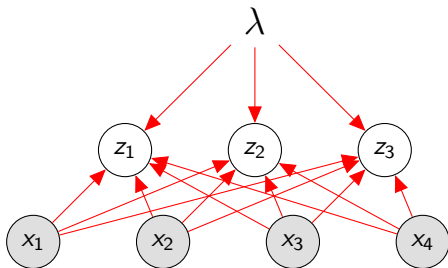
still mean field

$$Z_j | x \sim \text{Bernoulli}(b_j)$$

but with a shared set of parameters

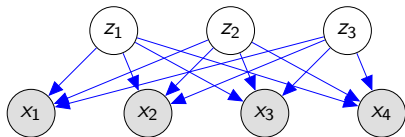
- where  $b_1^K = \text{NN}(x; \lambda)$

# Amortised Mean Field Inference for Latent Factor Model

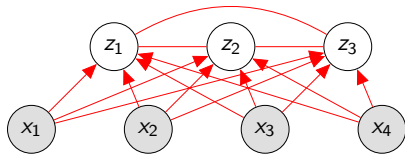


# Overview

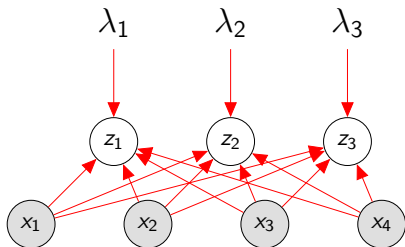
Joint distribution



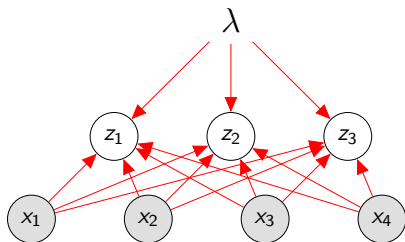
Posterior



Mean field



Amortised VI



# Summary

- Posterior inference is often **intractable** because the marginal likelihood (or **evidence**)  $p(x)$  cannot be computed efficiently.
- Variational inference approximates the posterior  $p(z|x)$  with a simpler distribution  $q(z)$ .

# Summary

- The variational objective is the **evidence lower bound (ELBO)**:

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

- The **ELBO** is a lower bound on the log-evidence.
- The solution to the ELBO minimises  $\text{KL}(q(z) \parallel p(z|x))$
- When  $q(z) = p(z|x)$  we recover EM.

# Summary

- We design  $q(z)$  to be simple
- A common approximation is the **mean field** approximation which assumes that all latent variables are independent:

$$q(z) = \prod_{i=1}^N q(z_i)$$



# Literature I

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5): 993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993. URL <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>.

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

Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *NIPS*, pages 472–478, 1996. URL <http://papers.nips.cc/paper/1144-factorial-hidden-markov-models.pdf>.

# Literature II

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998. URL <http://www.cs.toronto.edu/~fritz/absps/emk.pdf>.