Variational Inference: The Basics

Philip Schulz and Wilker Aziz

https://github.com/vitutorial/VITutorial

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

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

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 α (write e.g. $p(z|\alpha)$). In that case, we call α 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.

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

Bayes rule

We can *invert* a conditional probability distribution.

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

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

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

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

Bayesian Inference

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

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

Bayesian Inference

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

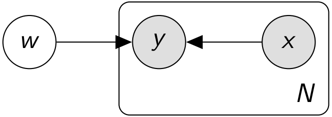
- $p(x) = \int p(x, z|\theta)\mathrm{d}z$ (frequentist)
- $p(x) = \int \int p(x, z, \theta)\mathrm{d}z\mathrm{d}\theta$ (Bayesian)

Today we will only treat the frequentist case!

We cannot compute the posterior when

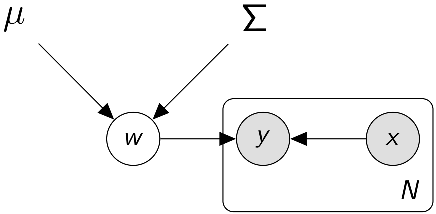
1. The functional form of the posterior is unknown (we don't know which parameters to infer)
2. The functional form is known but the computation is intractable

Bayesian Log-Linear POS Tagger



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

Bayesian Log-Linear POS Tagger



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

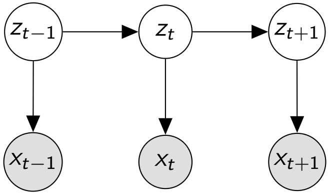
Bayesian Log-Linear POS Tagger

Intuition

Simply assume that the posterior is Gaussian.

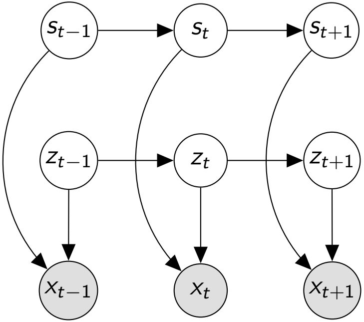
Factorial HMMs

FHMMs have several Markov chains over latent variables.



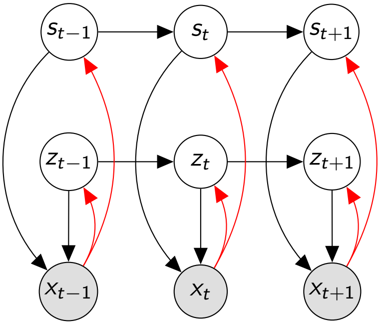
Factorial HMMs

FHMMs have several Markov chains over latent variables.



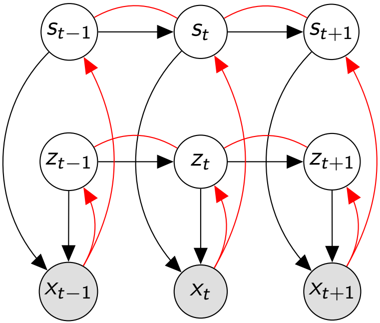
Factorial HMMs

FHMMs have several Markov chains over latent variables.



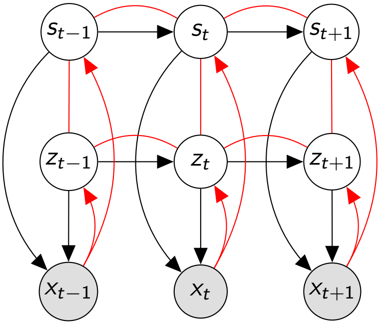
Factorial HMMs

FHMMs have several Markov chains over latent variables.



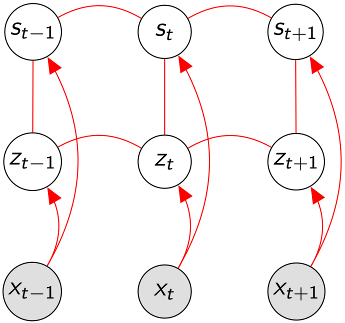
Factorial HMMs

FHMMs have several Markov chains over latent variables.



Factorial HMMs

Inference network for FHHMs.



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

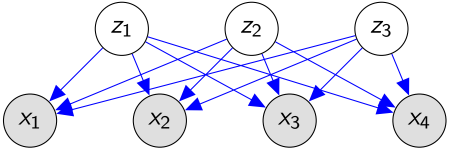
Intuition

Simply assume that the posterior consists of independent Markov chains.

Latent Factor Model

Joint distribution: latent variables are marginally independent a priori

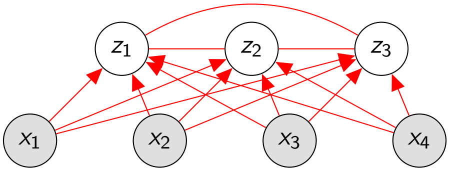
for example, $K = 3, N = 4$



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).
- 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.

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.

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

Recap KL divergence

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

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

Recap KL divergence

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

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

Recap KL divergence

Properties

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

Recap KL divergence

Properties

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

Recap KL divergence

Properties

- $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = \mathbb{E}_{q(z)}\left[\log\left(\frac{q(z)}{p(z|x)}\right)\right] \geq 0$ with equality iff $q(z) = p(z|x)$.
- $-\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = \mathbb{E}_{q(z)}\left[\log\left(\frac{p(z|x)}{q(z)}\right)\right] \leq 0.$
- $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = \infty$
  if $\exists z$ s.t. $p(z|x) = 0$ and $q(z) > 0$.

VI derivation I

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

VI derivation I

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

VI derivation I

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

VI derivation I

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

VI derivation I

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

VI derivation I

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

VI derivation I

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

VI derivation I

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

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

VI derivation II

Recall that we want to find $q(z)$ such that $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right)$ is small.

VI derivation II

Recall that we want to find $q(z)$ such that $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right)$ is small.

Formal Objective

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

VI derivation II

Recall that we want to find $q(z)$ such that $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right)$ is small.

Formal Objective

$$\min_{q(z)} \mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = \max_{q(z)} -\mathrm{KL}\left(q(z) \,||\, p(z|x)\right)$$

VI derivation II

$$\max_{q(z)} -\mathrm{KL}\left(q(z) \,||\, p(z|x)\right)$$

VI derivation II

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

VI derivation II

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

VI derivation II

$$
\begin{aligned}
&\max_{q(z)} -\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) \\
&= \max_{q(z)} \int q(z) \log\left(\frac{p(z|x)}{q(z)}\right) \mathrm{d}z \\
&= \max_{q(z)} \int q(z) \log\left(\frac{p(z,x)}{p(x)q(z)}\right) \mathrm{d}z \\
&= \max_{q(z)} \int q(z) \log\left(p(z,x)\right) \mathrm{d}z - \int q(z) \log q(z) \mathrm{d}z - \overbrace{\log p(x)}^{\textit{constant}}
\end{aligned}
$$

VI derivation II

$$
\begin{aligned}
&\max_{q(z)} -\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) \\
&= \max_{q(z)} \int q(z) \log\left(\frac{p(z|x)}{q(z)}\right) \mathrm{d}z \\
&= \max_{q(z)} \int q(z) \log\left(\frac{p(z,x)}{p(x)q(z)}\right) \mathrm{d}z \\
&= \max_{q(z)} \int q(z) \log\left(p(z,x)\right) \mathrm{d}z - \int q(z) \log q(z) \mathrm{d}z - \overbrace{\log p(x)}^{\textit{constant}} \\
&= \max_{q(z)} \mathbb{E}_{q(z)}\left[\log p(x,z)\right] + \mathbb{H}\left(q(z)\right)
\end{aligned}
$$

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

ELBO

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

Performing VI (Frequentist Case)

Variational Objective

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

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

Performing VI (Frequentist Case)

Variational Objective

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

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

Also optimize the model!

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

Performing VI (Frequentist Case)

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

Performing VI (Frequentist Case)

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

1. Maximize (regularised) expected log-density.

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

Performing VI (Frequentist Case)

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

1. Maximize (regularised) expected log-density.

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

2. Optimise generative model.

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

Recap: EM Algorithm

E-step Compute: $\mathbb{E}_{p(z|x)}\left[\log\left(p(x,z)\right)\right]$.

Same as: $\max_{p(z|x)} \mathbb{E}_{p(z|x)}\left[\log p(x,z)\right]$

Recap: EM Algorithm

E-step Compute: $\mathbb{E}_{p(z|x)}\left[\log\left(p(x,z)\right)\right]$.

Same as: $\max\limits_{p(z|x)} \mathbb{E}_{p(z|x)}\left[\log p(x,z)\right]$

M-step $\max\limits_{p(x,z)} \mathbb{E}_{p(z|x)}\left[\log p(x,z)\right] + \underbrace{\mathbb{H}\left(p(z|x)\right)}_{\text{constant}}$

Recap: EM Algorithm

E-step Compute: $\mathbb{E}_{p(z|x)}\left[\log\left(p(x,z)\right)\right]$.

Same as: $\max_{p(z|x)} \mathbb{E}_{p(z|x)}\left[\log p(x,z)\right]$

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

EM is variational inference!

$$q(z) = p(z|x)$$

$$\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = 0$$

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

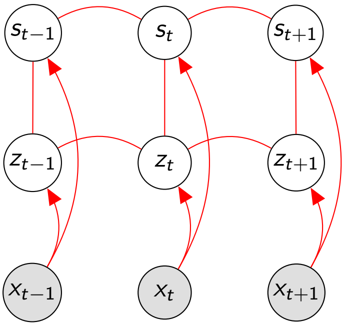
- Recall: The approximation $q(z)$ needs to be tractable.
- Common solution: make **all** latent variables independent under $q(z)$.
- Formal assumption: $q(z) = \prod_{i=1}^{N} q(z_i)$

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

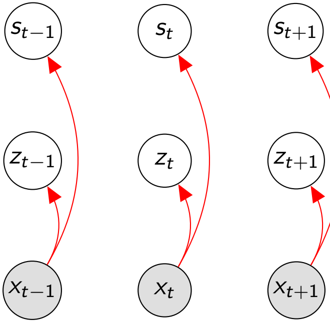
This approximation strategy is commonly known as **mean field** approximation.

Original FHHM Inference



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

Mean field FHHM Inference

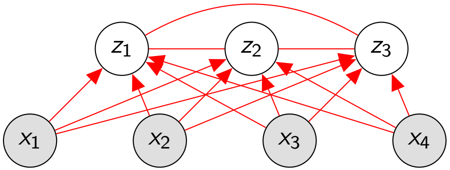


Approximate posterior $q(s, z) = \prod_{t=1}^{T} q(s_t)q(z_t)$

Original Latent Factor Model Inference

Joint distribution: latent variables are marginally independent a priori

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

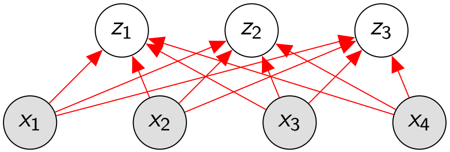


Exact Posterior: latent variables are dependent given observations

Mean Field Latent Factor Model Inference

Joint distribution: latent variables are marginally independent a priori

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



Approximate Posterior: latent variables are conditionally independent

Mean field assumption

We have K latent variables and D data points

- assume the posterior factorises as KD independent terms

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

Mean field assumption

We have K latent variables and D data points

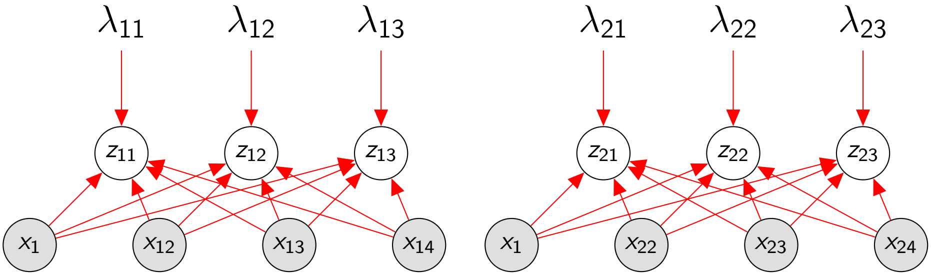
- assume the posterior factorises as KD independent terms

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

with independent sets of parameters $\lambda_j = \{b_j\}$

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

Mean field: example



Amortised variational inference

Amortise the cost of inference using NNs

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

Amortised variational inference

Amortise the cost of inference using NNs

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

still mean field

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

Amortised variational inference

Amortise the cost of inference using NNs

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

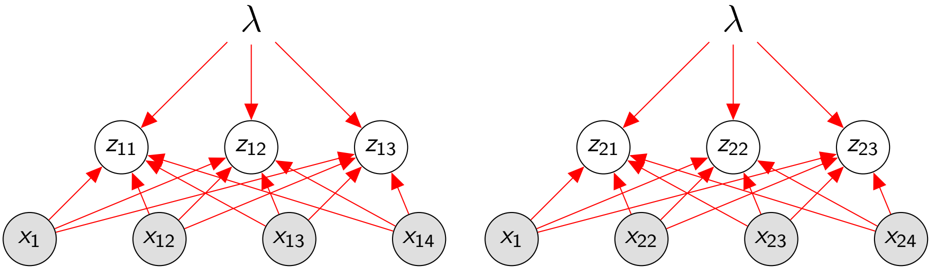
still mean field

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

but with a shared set of parameters

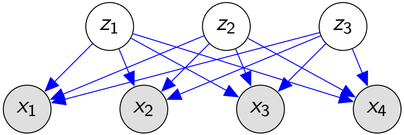
- where $b_1^K = g_{\lambda}(x)$

Amortized mean field: example

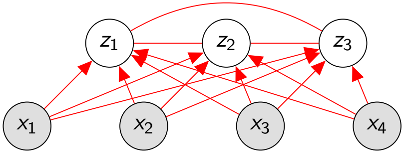


Overview

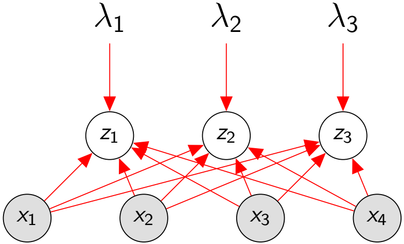
Joint distribution



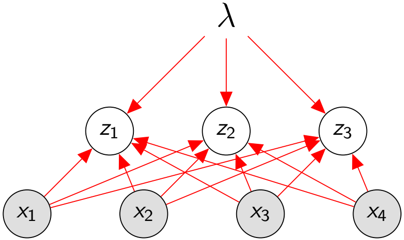
Exact Posterior



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 $q(z)$.
- The variational objective is the **evidence lower bound (ELBO)**:

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

Summary

- The **ELBO** is a lower bound on the log-evidence.
- When $q(z) = p(z|x)$ we recover EM.
- 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.