Variational Inference: The Basics

Philip Schulz and Wilker Aziz

https://github.com/philschulz/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

Model parameters θ are also random. The generative model becomes

- $p(x, \theta)$ for fully observed data (supervised learning)
- $p(x, z, \theta)$ for observed and latent data (unsupervised learning)

Bayesian Inference

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

- $p(x) = \int p(x, \theta) \mathrm{d}\theta$ (supervised learning)
- $p(x) = \int \int p(x, z, \theta) z \mathrm{d}\theta$ (unsupervised learning)

Bayesian Inference

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

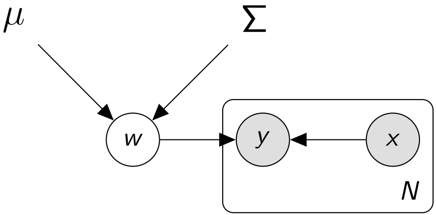
- $p(x) = \int p(x, \theta) \mathrm{d}\theta$ (supervised learning)
- $p(x) = \int \int p(x, z, \theta) z \mathrm{d}\theta$ (unsupervised learning)

Again, approximate inference is needed.

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 Logistic Regression



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

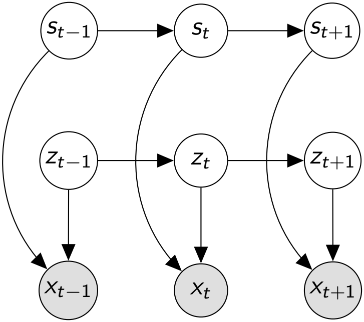
Bayesian Logistic Regression

Intuition

Simply assume that the posterior is Gaussian.

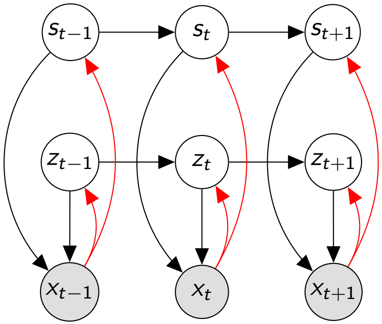
Factorial HMMs

FHMMs have several Markov chains over latent variables.



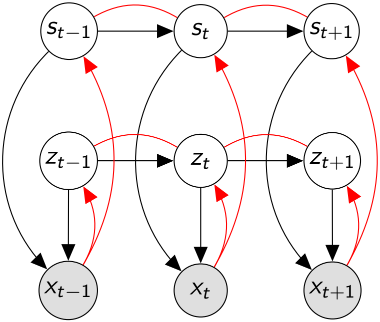
Factorial HMMs

FHMMs have several Markov chains over latent variables.



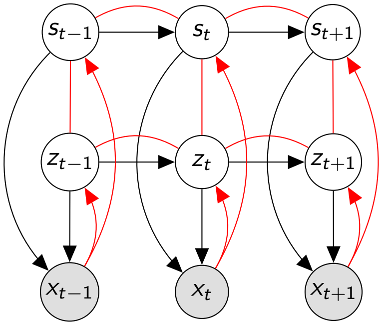
Factorial HMMs

FHMMs have several Markov chains over latent variables.



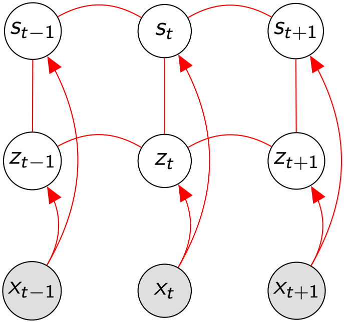
Factorial HMMs

FHMMs have several Markov chains over latent variables.



Factorial HMMs

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.

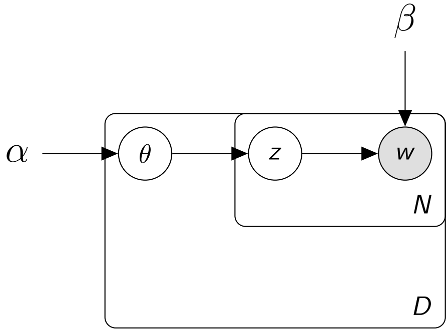
Factorial HMMs

Intuition

Simply assume that the posterior consists of independent Markov chains.

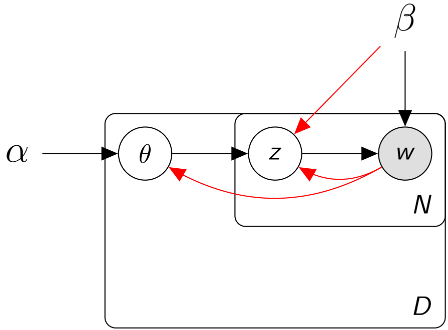
Latent Dirichlet Allocation

An admixture model that changes its mixture weights per document. We assume that the mixture components are fixed.



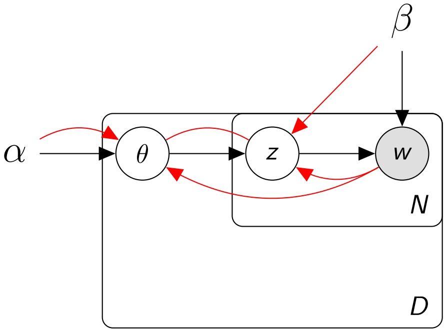
Latent Dirichlet Allocation

An admixture model that changes its mixture weights per document. We assume that the mixture components are fixed.



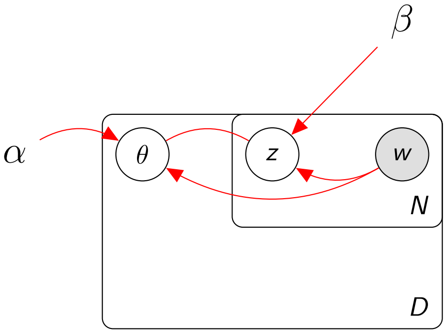
Latent Dirichlet Allocation

An admixture model that changes its mixture weights per document. We assume that the mixture components are fixed.



Latent Dirichlet Allocation

Inference network for LDA.



Latent Dirichlet Allocation

An admixture model that changes its mixture weights per document. Here we assume that the mixture components are fixed.

- D documents.
- N tokens and latent variables per document.
- L outcomes per latent variable.
- Complexity of inference: $\mathcal{O}(L^{DN})$.

Latent Dirichlet Allocation

Intuition

Simply assume that the posterior consists of independent categorical and Dirichlet distributions.

Latent Dirichlet Allocation

Intuition

Simply assume that the posterior consists of independent categorical and Dirichlet distributions.

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.

- $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) = \int q(z) \log\left(\frac{q(z)}{p(z|x)}\right) \mathrm{d}z$ (continuous)

Recap KL divergence

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

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

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

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

Recap KL divergence

Originally known as relative entropy. Read $\mathrm{KL}\left(q(z) \mid\mid p(z)\right)$ as the divergence of p from q or the entropy of p relative to q.

Recap KL divergence

Originally known as relative entropy. Read $\mathrm{KL}\left(q(z) \,||\, p(z)\right)$ as the divergence of p from q or the entropy of p relative to q.

Expand $\mathrm{KL}\left(q \,||\, p\right)$

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

Recap KL divergence

Originally known as relative entropy. Read $\mathrm{KL}\left(q(z) \mid\mid p(z)\right)$ as the divergence of p from q or the entropy of p relative to q.

Expand $\mathrm{KL}\left(q \mid\mid p\right)$

$$\mathbb{E}_{q(z)}\left[\log\left(\frac{q(z)}{p(z)}\right)\right] = -\mathbb{E}_{q(z)}\left[\log p(z)\right] + \mathbb{E}_{q(z)}\left[\log q(z)\right]$$
$$= \mathrm{CE} - \mathbb{H}\left(q(z)\right)$$

Recap KL divergence

Originally known as relative entropy. Read $\mathrm{KL}\left(q(z) \,||\, p(z)\right)$ as the divergence of p from q or the entropy of p relative to q.

Expand $\mathrm{KL}\left(q \,||\, p\right)$

$$\mathbb{E}_{q(z)}\left[\log\left(\frac{q(z)}{p(z)}\right)\right] = -\mathbb{E}_{q(z)}\left[\log p(z)\right] + \mathbb{E}_{q(z)}\left[\log q(z)\right]$$
$$= \mathrm{CE} - \mathbb{H}\left(q(z)\right)$$

Bits needed to encode p once we know q.

Recap KL divergence

Properties

- $\mathrm{KL}\left(q(z) \mid\mid p(z|x)\right) \geq 0$ with equality iff $q(z) = p(z|x)$.

Recap KL divergence

Properties

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

- $\mathrm{KL}\left(q(z) \,||\, p(z|x)\right) \geq 0$ 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 $q(z) > 0$.

Recap KL divergence

Properties

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

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

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

VI derivation I

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

VI derivation I

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

VI derivation I

$$\log p(x) \geq \int q(z) \log \left(\frac{p(z|x)}{q(z)} \right) \mathrm{d}z + \log p(x)$$

VI derivation I

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

VI derivation I

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

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

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

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

VI derivation II

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

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)

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.

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

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.

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

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

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!

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

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

Performing VI (Bayesian Case)

We have latent variables z (e.g. POS tags) and θ (e.g. model parameters).

1. Maximise over local variables z.

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

Performing VI (Bayesian Case)

We have latent variables z (e.g. POS tags) and θ (e.g. model parameters).

1. Maximise over local variables z.

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

2. Maximise over global variables θ.

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

Differences between frequentist VI and VB (Variational Bayes)

- Frequentist VI optimises two sets of parameters, VB only optimises variational parameters
- Entropy term matters in the M-step for VB but not for VI

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

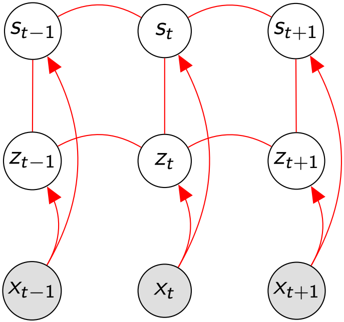
- Recall: The approximation $q(z)$ needs to be tractable.
- Common solution: make **all** latent variables independent under $q(z)$.
- Formal assumption: $q(z) = \prod_{i=1}^{N} q(z_i)$

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

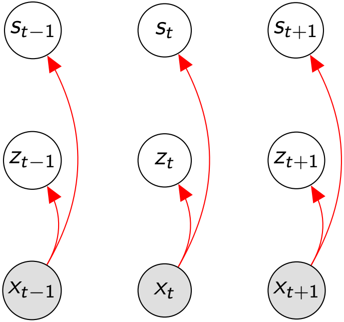
This approximation strategy is commonly known as **mean field** approximation.

Original FHHM Inference



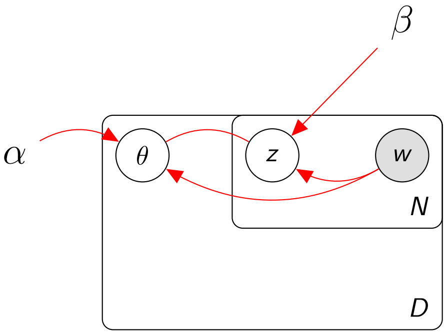
Exact posterior $p(s, z|x)$

Mean field FHHM Inference



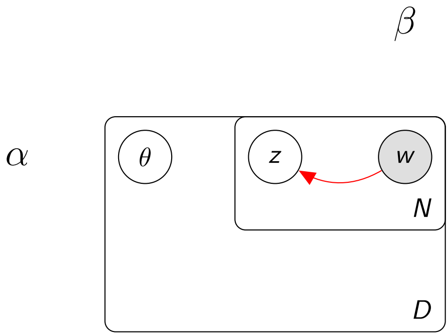
Approximate posterior $q(s, z) = \prod_{t=1}^{T} q(s_t)q(z_t)$

Original LDA Inference



Exact posterior $p(z, \theta | w, \alpha, \beta)$

Mean field LDA Inference



Approximate posterior

$$q(z, \theta | w, \alpha, \beta) = \prod_{d=1}^{D} q(\theta_d) \prod_{i=1}^{N} q(z_i | w)$$

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)**:

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

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. ISSN 1532-4435. 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.

Literature II

Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996. URL http://papers.nips.cc/paper/1144-factorial-hidden-markov-models.pdf.

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.