

# Variational Inference: The Basics

Philip Schulz and Wilker Aziz

# 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.

## 2 Examples of Generative Models

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

# 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: the prior may depend on a non-random quantity  $\alpha$  (write  $p(z|\alpha)$ ). In that case, we call  $\alpha$  a hyperparameter.

# Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \quad (1)$$

# Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{p(x)} \quad (2)$$

# Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{p(x)} \quad (3)$$

# Bayes' rule

Bayes rule asserts that we can *invert* a conditional probability distribution.

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{marginal likelihood/evidence}}} \quad (4)$$



# 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

Under the Bayesian view, model parameters  $\theta$  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. This is because  $\theta$  is often high-dimensional (just think of neural nets!).

- ▶  $p(x) = \int p(x, \theta) d\theta$  (supervised learning)
- ▶  $p(x) = \int \int p(x, z, \theta) dz d\theta$  (unsupervised learning)

# Bayesian Inference

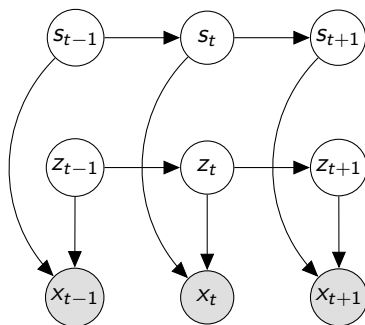
The evidence becomes even harder to compute. This is because  $\theta$  is often high-dimensional (just think of neural nets!).

- ▶  $p(x) = \int p(x, \theta) d\theta$  (supervised learning)
- ▶  $p(x) = \int \int p(x, z, \theta) dz d\theta$  (unsupervised learning)

Again, approximate inference is needed.

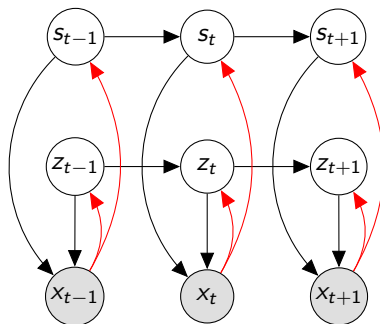
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



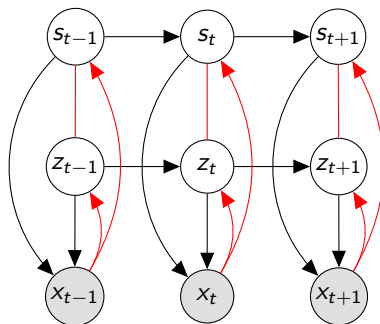
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



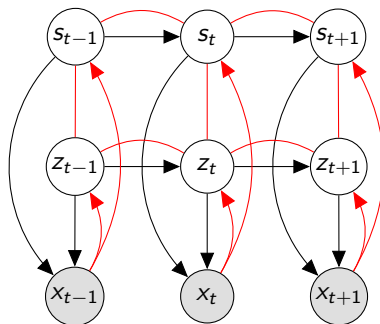
# Factorial HMMs

FHMMs have several Markov chains over latent variables.



# Factorial HMMs

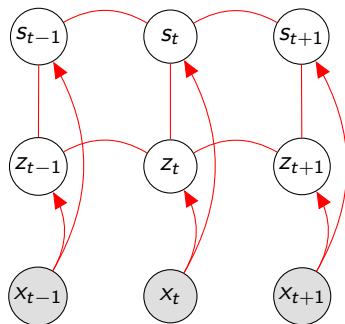
FHMMs have several Markov chains over latent variables.





# 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.

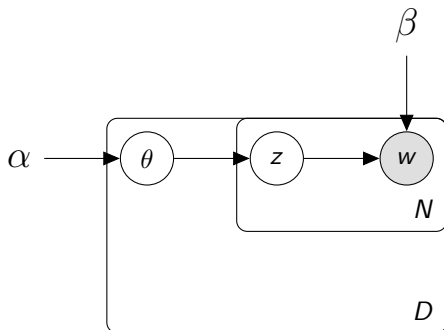
- ▶  $M$  Markov chains over latent variables.
- ▶  $L$  outcomes per latent variable.
- ▶ Sequence of length  $T$ .
- ▶ Complexity of inference:  $\mathcal{O}(L^{2M}T)$ .

## Intractable

Exponential dependency on the number of hidden Markov chains.

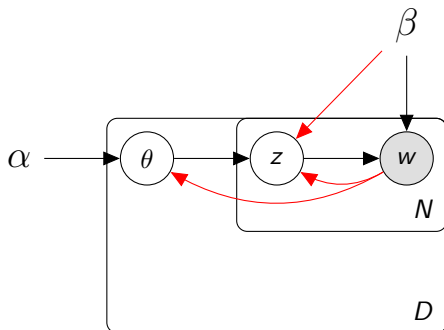
# Latent Dirichlet Allocation

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



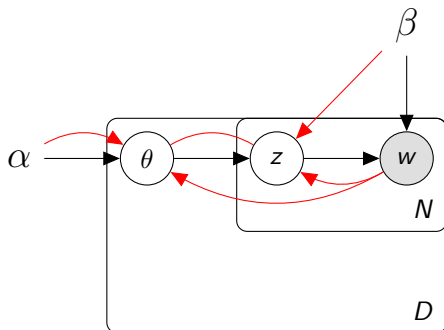
# Latent Dirichlet Allocation

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



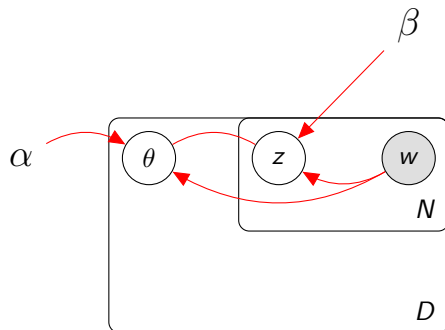
# Latent Dirichlet Allocation

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:  $\mathcal{O}(L^{DN})$ .



# The Goal

Assume  $p(z|x)$  is intractable.

# The Goal

Assume  $p(z|x)$  is intractable.

## Idea

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

# The Goal

Assume  $p(z|x)$  is intractable.

## Idea

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

## Requirement

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

# The Goal

Assume  $p(z|x)$  is intractable.

## Idea

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

## Requirement

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

## Implementation

Minimize  $\text{KL}(q(z) \parallel p(z|x))$ .

# Recap KL divergence

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

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

# 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  $q(z) = p(z|x)$ .
- ▶  $\text{KL}(q(z) \parallel p(z|x)) = \infty$  if  $\exists z$  s.t.  $p(z|x) = 0$  and  $q(z) > 0$ .

# Recap KL divergence

## Properties

- ▶  $\text{KL}(q(z) \parallel p(z|x)) \geq 0$  with equality iff  $q(z) = p(z|x)$ .
- ▶  $\text{KL}(q(z) \parallel p(z|x)) = \infty$  if  $\exists z$  s.t.  $p(z|x) = 0$  and  $q(z) > 0$ .
- ▶ In general  $\text{KL}(q(z) \parallel p(z|x)) \neq \text{KL}(p(z|x) \parallel q(z))$ .



# Recap KL divergence

## Properties

- ▶  $\text{KL}(q(z) \parallel p(z|x)) \geq 0$  with equality iff  $q(z) = p(z|x)$ .
- ▶  $\text{KL}(q(z) \parallel p(z|x)) = \infty$  if  $\exists z$  s.t.  $p(z|x) = 0$  and  $q(z) > 0$ .
- ▶ In general  $\text{KL}(q(z) \parallel p(z|x)) \neq \text{KL}(p(z|x) \parallel q(z))$ .
- ▶  $-\text{KL}(q(z) \parallel p(z|x)) \leq 0$ .

# VI derivation I

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

# VI derivation I

$$\begin{aligned} \int q(z) \log \left( \frac{p(z|x)}{q(z)} \right) dz + \log(p(x)) \\ = -\text{KL}(q(z) \parallel p(z|x)) + \log(p(x)) \end{aligned}$$

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

$$\min_{q(z)} \text{KL}(q(z) \parallel p(z|x)) = \max_{q(z)} - \text{KL}(q(z) \parallel p(z|x))$$

# VI derivation II

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

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

## ELBO

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

# Performing VI

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



# Performing VI

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

1. Compute the expected log-density

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

# Performing VI

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

1. Compute the expected log-density  $\mathbb{E}_{q(z)} [\log (p(x, z))]$ .
2. Maximize with respect to  $q(z)$  and while trying to keep  $q(z)$  as broad as possible (through entropy regularisation):

$$\max_{q(z)} \mathbb{E}_{q(z)} [\log (p(x, z))] + \mathbb{H} (q(z)) \quad (5)$$

# What if $q(z) = p(z|x)$ ?

If  $q(z) = p(z|x)$  then  $\text{KL}(q(z) \parallel p(z|x)) = 0$  and thus we are directly optimising the log-evidence.

1. Compute the expected log-density  $\mathbb{E}_{p(z|x)} [\log(p(x, z))]$ .
2. Maximize with respect to  $p(z|x)$  and while trying to keep  $p(z|x)$  as broad as possible (through entropy regularisation):

$$\max_{p(z|x)} \mathbb{E}_{p(z|x)} [\log(p(x, z))] + \mathbb{H}(p(z|x)) \quad (6)$$

# What if $q(z) = p(z|x)$ ?

If  $q(z) = p(z|x)$  then  $\text{KL}(q(z) || p(z|x)) = 0$  and thus we are directly optimising the log-evidence.

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

**M-step** Maximize with respect to  $p(z|x)$  and while trying to keep  $p(z|x)$  as broad as possible (through entropy regularisation):

$$\max_{p(z|x)} \mathbb{E}_{p(z|x)} [\log(p(x, z))] + \mathbb{H}(p(z|x)) \quad (7)$$

# Relationship to EM

- ▶ Variational Inference where  $q(z) = p(z|x)$  is EM!
- ▶ The implementation of the M-step does not change.
- ▶ The implementation of the E-step changes when  $q(z) \neq p(z|x)$  because the expectation term changes.

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

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

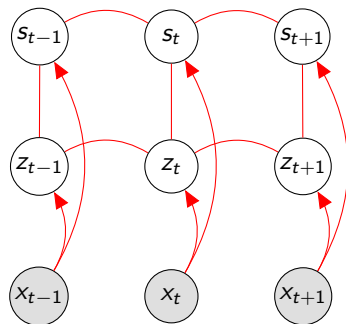
# Designing a tractable approximation

- ▶ 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.

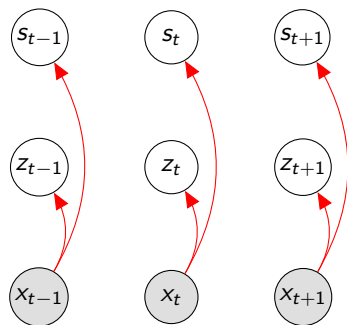


# Original FHMM Inference



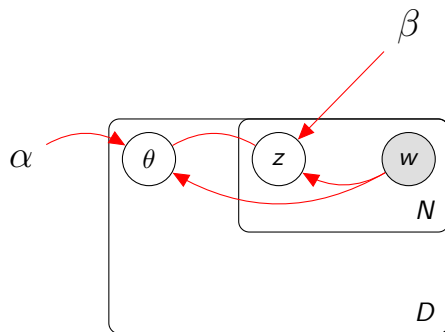
Exact posterior  $p(s, z|x)$

# Mean field FHMM Inference



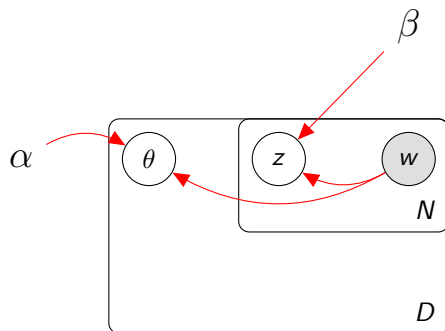
Approximate posterior  $q(s, z) = q(s)q(z)$

# Original LDA Inference



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

# Mean field LDA Inference



Approximate posterior

$$q(z, \theta | w, \alpha, \beta) = q(z | w, \beta) q(\theta | w, \alpha)$$

# 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)**:

$$\mathbb{E}_{q(z)} [\log (p(x, z))] + \mathbb{H} (q(z)) \quad (8)$$

# 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>.