10-708 Probabilistic Graphical Models

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Variational Inference

Matt Gormley
Lecture 17
Mar. 31, 2021

Reminders

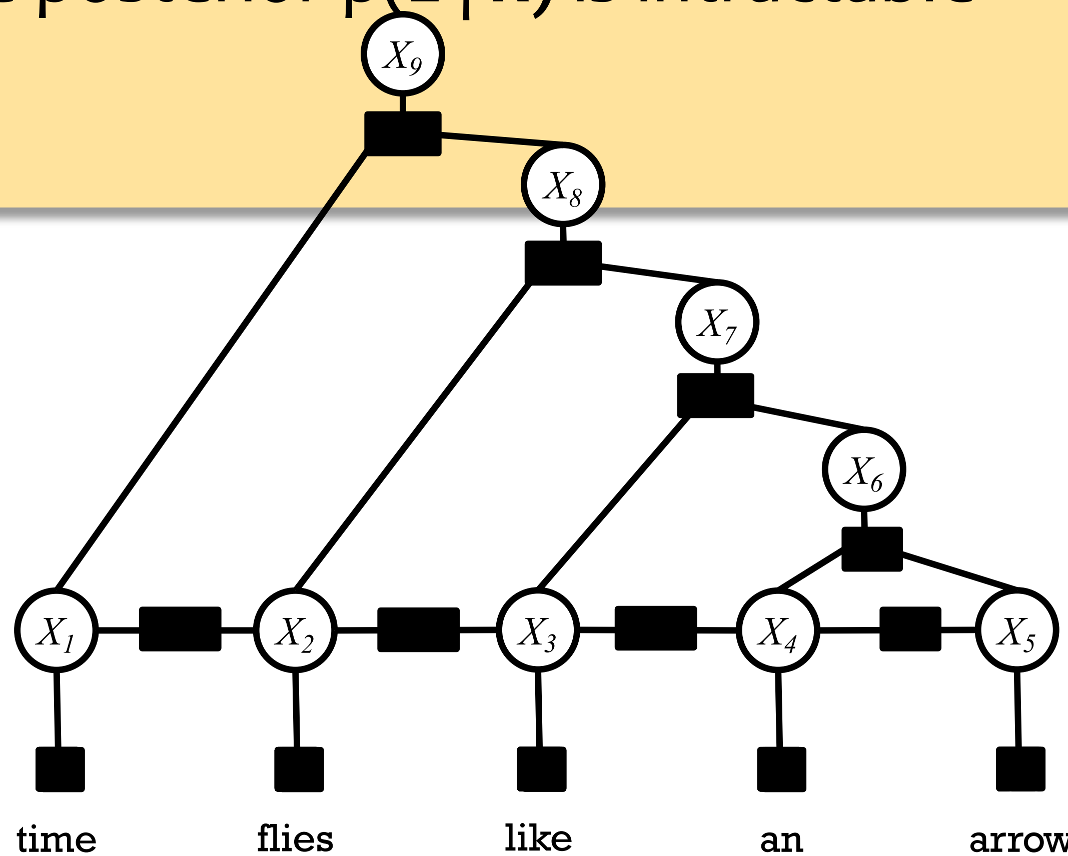
- Project Proposal
  - Due: Wed, Mar. 31 at 11:59pm
- Homework 4: MCMC
  - Out: Wed, Mar. 24
  - Due: Wed, Apr. 7 at 11:59pm

HIGH-LEVEL INTRO TO VARIATIONAL INFERENCE

Variational Inference

Problem:

– For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable

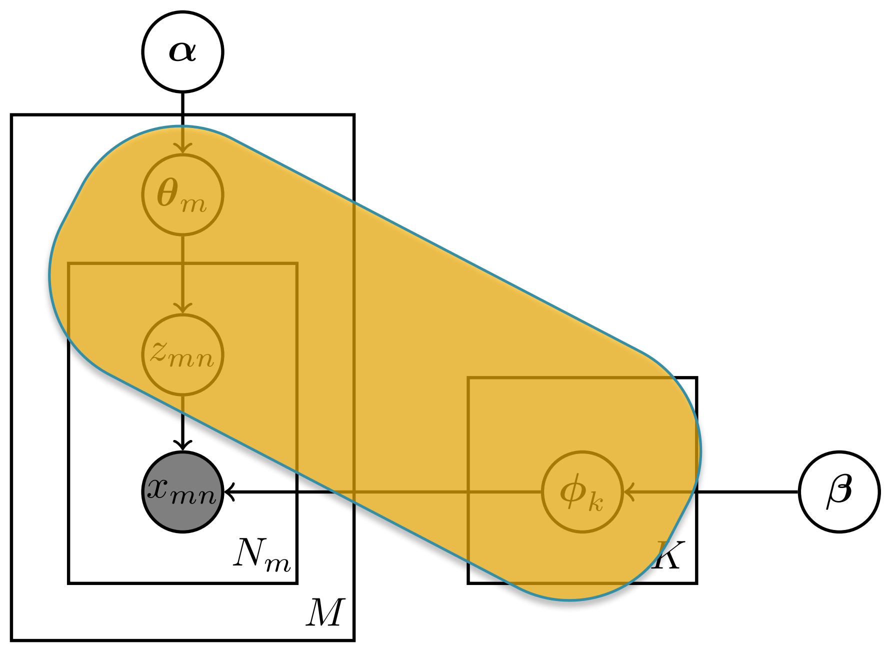


Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Problem:

- For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable
- For training data **x** and parameters **z**, estimating the posterior p(**z** | **x**) is intractable



Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Problem:

- For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable
- For training data **x** and parameters **z**, estimating the posterior p(**z** | **x**) is intractable

Solution:

- Approximate p(**z** | **x**) with a simpler q(**z**)
- Typically q(**z**) has more independence assumptions than p(**z** | **x**) – fine b/c q(**z**) is tuned for a specific **x**
- **Key idea**: pick a single q(**z**) from some family Q that best approximates p(**z** | **x**)

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Terminology:

- $q(\mathbf{z})$: the **variational approximation**
- Q: the **variational family**
- Usually $q_\theta(\mathbf{z})$ is parameterized by some θ called **variational parameters**
- Usually $p_\alpha(\mathbf{z} \mid \mathbf{x})$ is parameterized by some fixed α – we'll call them the parameters

Example Algorithms:

- mean-field variational inference
- loopy belief propagation
- tree-reweighted belief propagation
- expectation propagation

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Is this trivial?

- Note: We are not defining a new distribution simple $q_\theta(\mathbf{z} \mid \mathbf{x})$, there is one simple $q_\theta(\mathbf{z})$ for each $p_\alpha(\mathbf{z} \mid \mathbf{x})$
- Consider the MCMC equivalent of this:
  - you could draw samples $z^{(i)} \sim p(\mathbf{z} \mid \mathbf{x})$
  - then train some simple $q_\theta(\mathbf{z})$ on $z^{(1)}, z^{(2)}, \ldots, z^{(N)}$
  - hope that the sample adequately represents the posterior for the given $\mathbf{x}$
- How is VI different from this?
  - VI doesn't require sampling
  - VI is fast and deterministic
  - Why? b/c we choose an objective function (KL divergence) that defines which q_θ best approximates p_α, and exploit the special structure of q_θ to optimize it

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

V.I. offers a new design decision

- Choose the distribution $p_\alpha(\mathbf{z} \mid \mathbf{x})$ that you really want, i.e. don't just simpify it to make it computationally convenient
- Then design a the structure of another distribution $q_\theta(\mathbf{z})$ such that V.I. is efficient

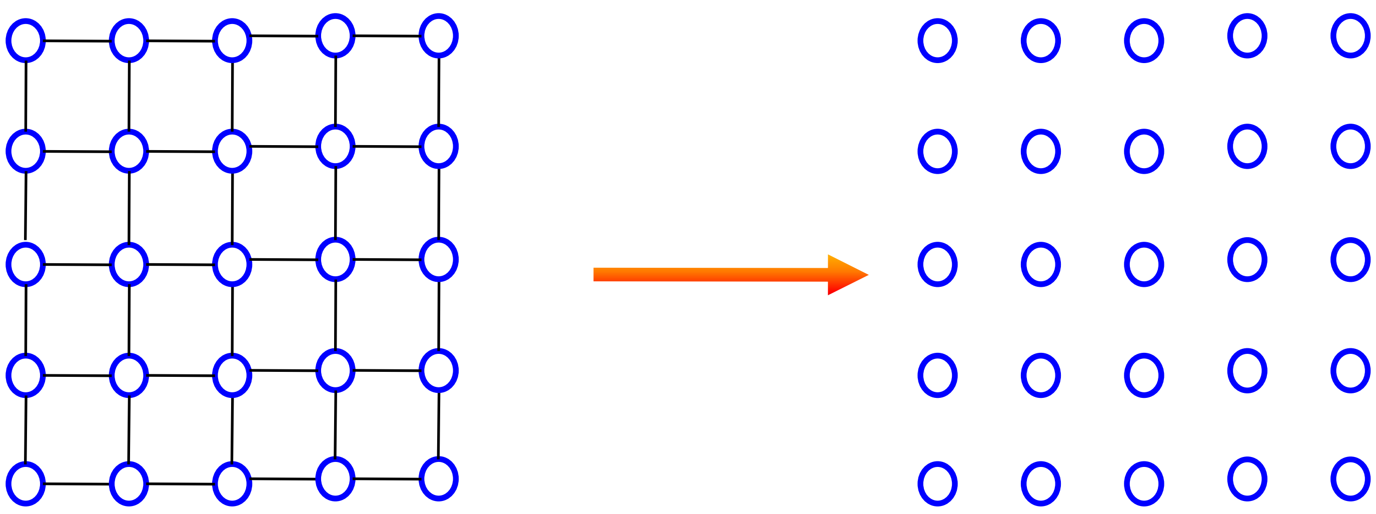
Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

EXAMPLES OF VARIATIONAL APPROXIMATIONS

Mean Field for MRFs

- Mean field approximation for Markov random field (such as the Ising model):

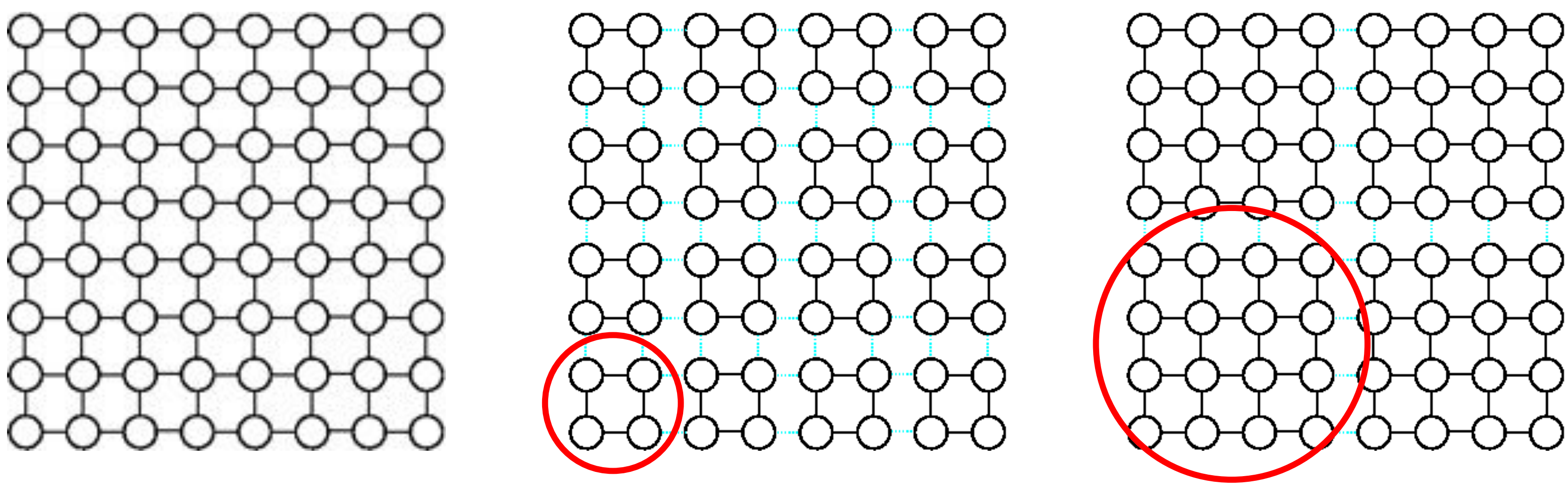
$$q(x) = \prod_{s \in V} q(x_s)$$



Variational Inference for MRFs

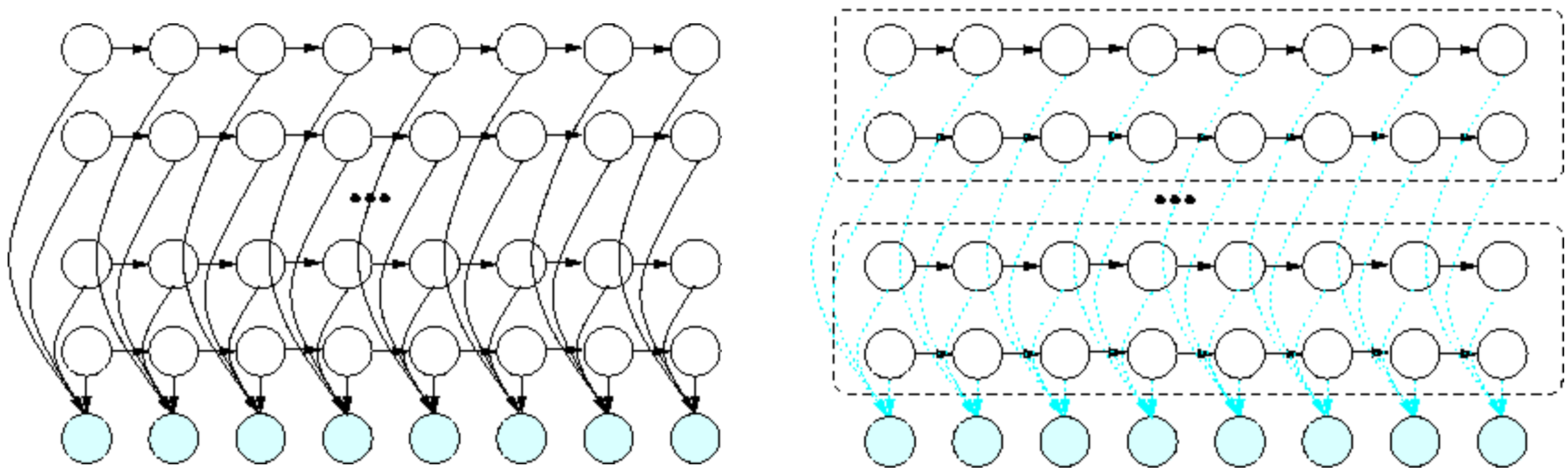
- We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:

- Instead of making all latent variables independent (i.e. naïve mean field, previous figure), clusters of (disjoint) latent variables are independent.



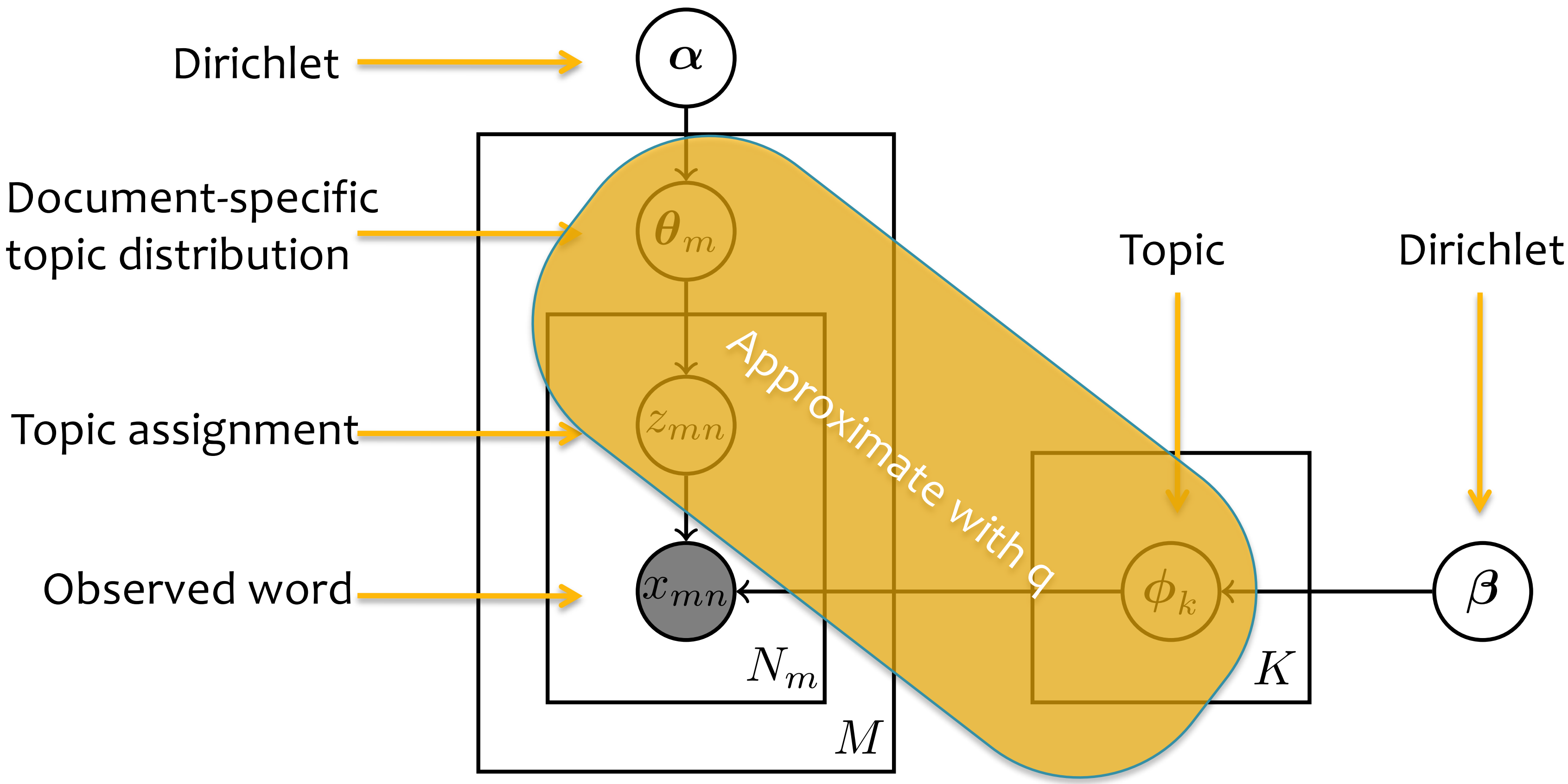
V.I. for Factorial HMM

- For a factorial HMM, we could decompose into chains



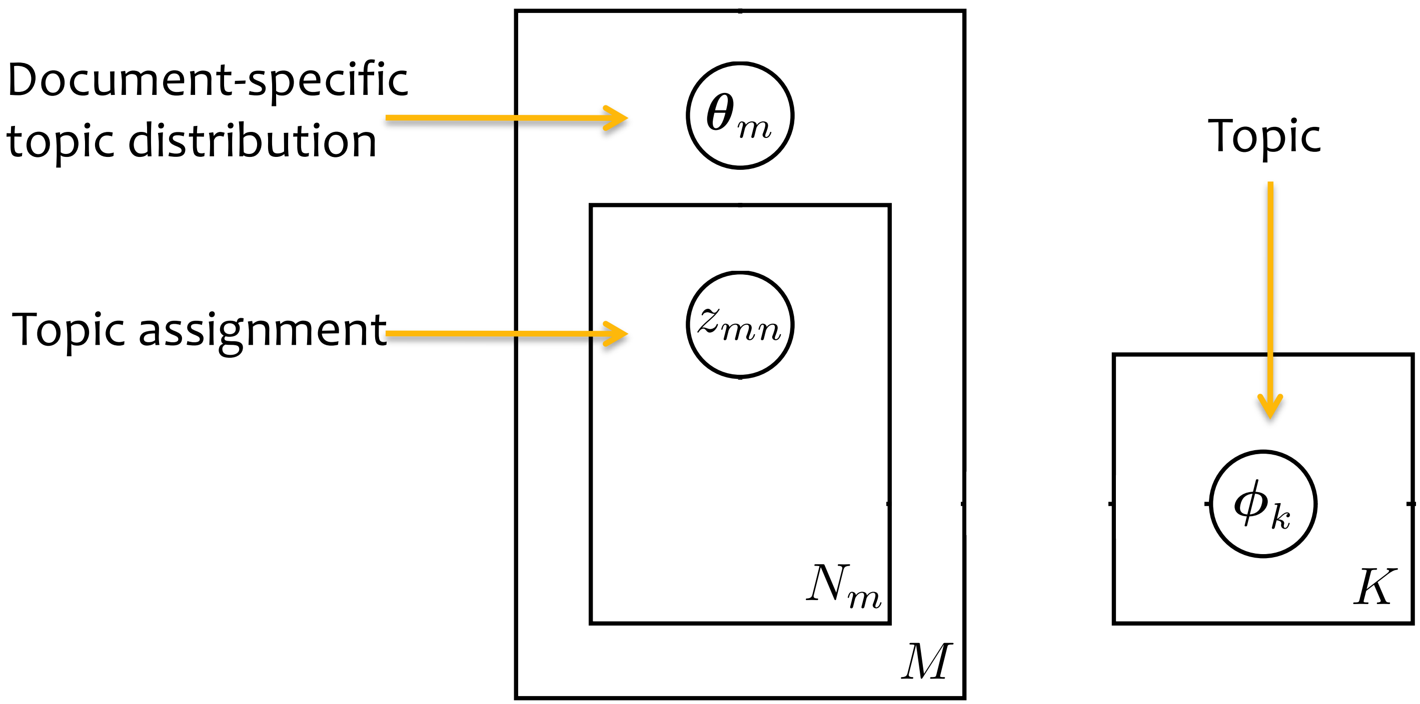
LDA Inference

- Explicit Variational Inference (original distribution)



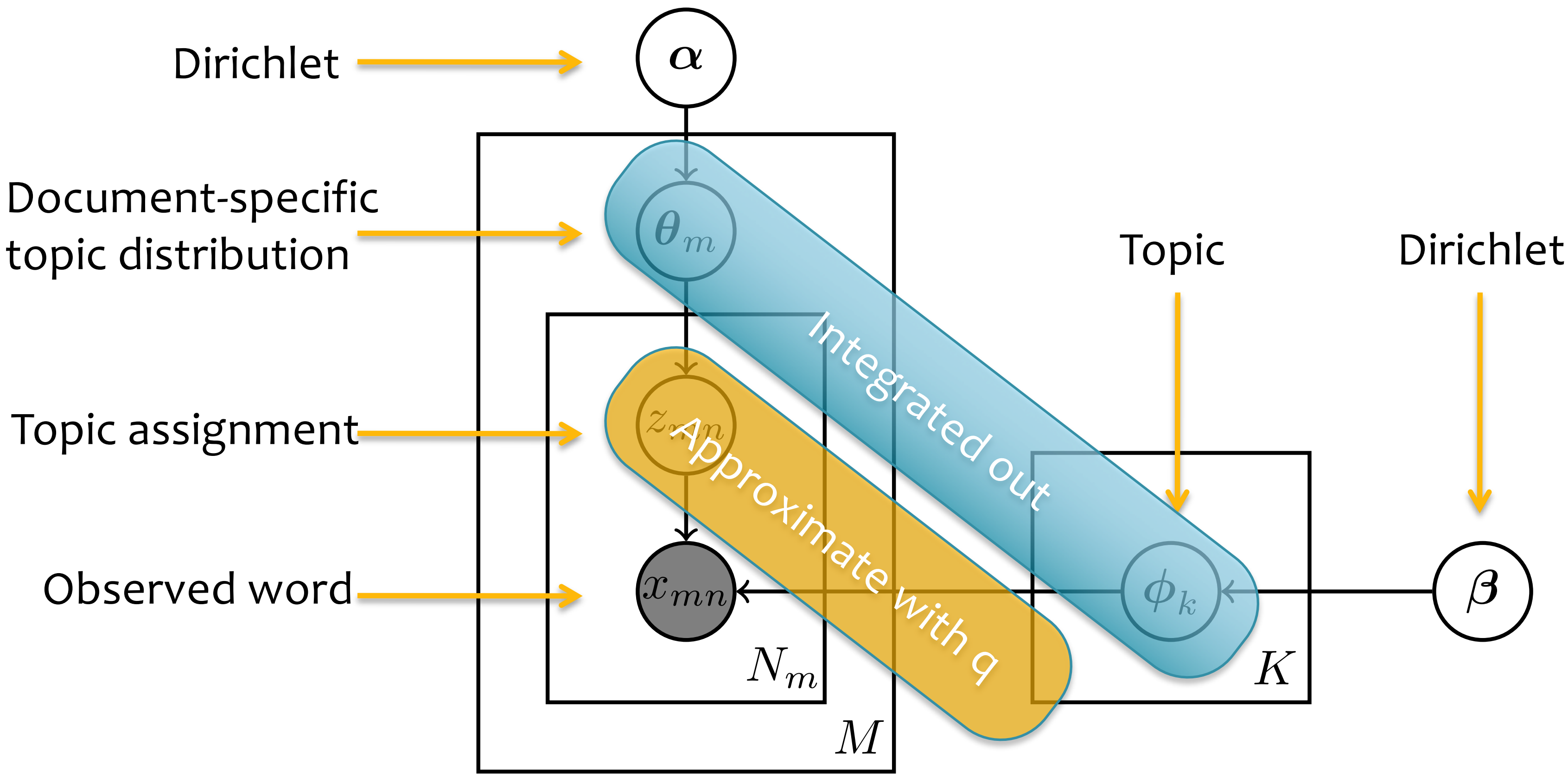
LDA Inference

- Explicit Variational Inference (variational approximation)



LDA Inference

- Collapsed Variational Inference



MEAN FIELD VARIATIONAL INFERENCE

KL Divergence

- *Definition*: for two distributions $q(x)$ and $p(x)$ over $x \in \mathcal{X}$, the **KL Divergence** is:

$$KL(q \,||\, p) = E_{q(x)}[\log q(x)/p(x)]$$

- *Properties*:
  - $KL(q \,||\, p)$ measures the **proximity** of two distributions q and p
  - KL is **not** symmetric: $KL(q \,||\, p) \neq KL(p \,||\, q)$
  - KL is minimized when $q(x) = p(x)$ for all $x \in \mathcal{X}$

Variational Inference

Whiteboard

- Background: KL Divergence
- Mean Field Variational Inference (overview)

Two Cases for Intractability

- <u>*Case 1*</u>:

  given a **joint distribution** p(x, z)

$$\Rightarrow p(z \mid x) = \frac{p(x, z)}{p(x)}$$

we assume p(x) is intractable

- <u>*Case 2*</u>:

  give **factor graph** and potentials

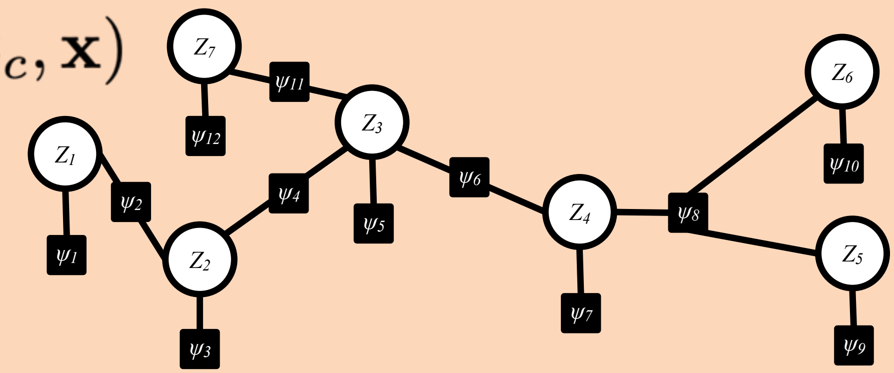
$$\Rightarrow p(z \mid x) = \frac{\tilde{p}(x, z)}{Z(x)}$$

we assume Z(x) is intractable

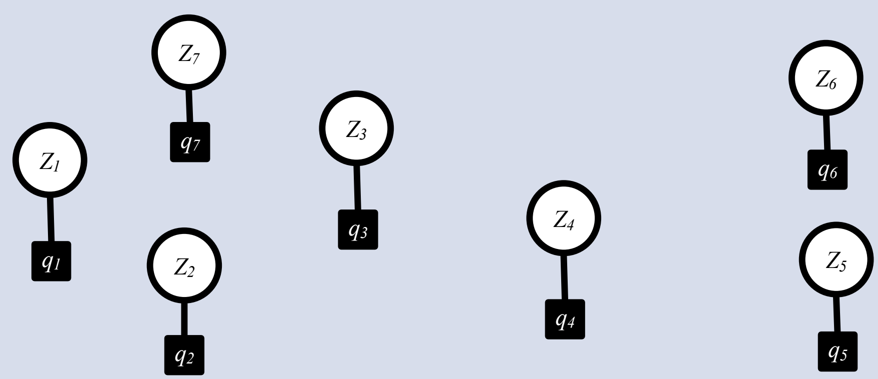
Mean Field Approximation

The **mean field approximation** assumes our variational approximation $q_\theta(\mathbf{z})$ treats each variable as independent

$$p_\alpha(\mathbf{z} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{z}_c, \mathbf{x})$$

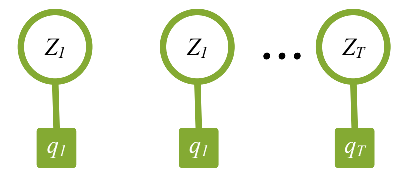


$$q_\theta(\mathbf{z}) = \prod_{t=1}^{T} q_t(z_t)$$



Mean Field V.I. Overview

1. *Goal*: estimate $p_\alpha(\mathbf{z} \mid \mathbf{x})$
   we assume this is intractable to compute exactly
2. *Idea*: approximate with another distribution $q_\theta(\mathbf{z}) \approx p_\alpha(\mathbf{z} \mid \mathbf{x})$ for each $\mathbf{x}$
3. *Mean Field*: assume $q_\theta(\mathbf{z}) = \prod_t q_t(z_t; \theta)$
   i.e., we decompose over variables
   other choices for the decomposition of $q_\theta(\mathbf{z})$ give rise to "structured mean field"



4. *Optimization Problem*: pick the q that minimizes KL(q || p)

$$\hat{q}(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\operatorname{argmin}} \, \mathsf{KL}(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x}))$$

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathsf{KL}(q_\theta(\mathbf{z}) || p_\alpha(\mathbf{z} \mid \mathbf{x}))$$

equivalent

5. *Optimization Algorithm*: coordinate descent
   i.e. pick the best $q_t(z_t)$ based on the other $\{ q_s(z_s) \}_{s \neq t}$ being fixed

Optimizing KL Divergence

- <u>*Question*</u>: How do we minimize KL?

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \underbrace{\mathsf{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x}))}$$

- <u>*Answer #1*</u>: Oh no! We can't even compute this KL.

Why we can't compute KL…

$$\begin{aligned}
\mathsf{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) &= E_{q(\mathbf{z})}\left[\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right)\right] \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - E_{q(\mathbf{z})}\left[\log p(\mathbf{z} \mid \mathbf{x})\right] \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - \boxed{E_{q(\mathbf{z})}\left[\log p(\mathbf{x}, \mathbf{z})\right] + E_{q(\mathbf{z})}\left[\log p(\mathbf{x})\right]} \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - E_{q(\mathbf{z})}\left[\log p(\mathbf{x}, \mathbf{z})\right] + \underbrace{\log p(\mathbf{x})}
\end{aligned}$$

we have the same problem with an intractable data likelihood p(x) or an intractable partition function Z(x)

we assumed this is intractable to compute!

Optimizing KL Divergence

- *Question*: How do we minimize KL?

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \underbrace{\mathsf{KL}(q_\theta(\mathbf{z}) \,\|\, p_\alpha(\mathbf{z} \mid \mathbf{x}))}$$

- *Answer #1*: Oh no! We can't even compute this KL.

Why we can't compute KL…

$$\begin{aligned}
\mathsf{KL}(q(\mathbf{z}) \,\|\, p(\mathbf{z} \mid \mathbf{x})) &= E_{q(\mathbf{z})}\left[\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right)\right] \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - E_{q(\mathbf{z})}\left[\log p(\mathbf{z} \mid \mathbf{x})\right] \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - \boxed{E_{q(\mathbf{z})}\left[\log \tilde{p}(\mathbf{z} \mid \mathbf{x})\right] + E_{q(\mathbf{z})}\left[\log Z(\mathbf{x})\right]} \\
&= E_{q(\mathbf{z})}\left[\log q(\mathbf{z})\right] - E_{q(\mathbf{z})}\left[\log \tilde{p}(\mathbf{z} \mid \mathbf{x})\right] + \underbrace{\log Z(\mathbf{x})}
\end{aligned}$$

we have the same problem with an intractable data likelihood p(x) or an intractable partition function Z(x)

we assumed this is intractable to compute!

Optimizing KL Divergence

- *Question*: How do we minimize KL?

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \underbrace{\mathsf{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x}))}$$

- *Answer #2*: We don't need to compute this KL
We can instead maximize the ELBO (i.e. **E**vidence **L**ower **BO**und)

$$\mathsf{ELBO}(q_\theta) = E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})]$$

The ELBO for a DGM

Here is why…

$$\begin{aligned}
\theta &= \underset{\theta}{\operatorname{argmin}} \mathsf{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x})) \\
&= \underset{\theta}{\operatorname{argmin}} E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] + \underbrace{\log p_\alpha(\mathbf{x})} \\
&= \underset{\theta}{\operatorname{argmin}} E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] \\
&= \underset{\theta}{\operatorname{argmax}} \mathsf{ELBO}(q_\theta)
\end{aligned}$$

dropping the
intractable term
gives the ELBO

Optimizing KL Divergence

- *Question*: How do we minimize KL?

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \underbrace{\mathrm{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x}))}$$

- *Answer #2*: We don't need to compute this KL
  We can instead maximize the ELBO (i.e. **E**vidence **L**ower **BO**und)

$$\mathrm{ELBO}(q_\theta) = E_{q_\theta(\mathbf{z})}\left[\log \tilde{p}_\alpha(\mathbf{z} \mid \mathbf{x})\right] - E_{q_\theta(\mathbf{z})}\left[\log q_\theta(\mathbf{z})\right]$$

The ELBO for a UGM

Here is why…

$$\begin{aligned}
\theta &= \operatorname*{argmin}_{\theta} \mathrm{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x})) \\
&= \operatorname*{argmin}_{\theta} E_{q_\theta(\mathbf{z})}\left[\log q_\theta(\mathbf{z})\right] - E_{q_\theta(\mathbf{z})}\left[\log \tilde{p}_\alpha(\mathbf{z} \mid \mathbf{x})\right] + \underbrace{\log Z_\alpha(\mathbf{x})} \\
&= \operatorname*{argmin}_{\theta} E_{q_\theta(\mathbf{z})}\left[\log q_\theta(\mathbf{z})\right] - E_{q_\theta(\mathbf{z})}\left[\log \tilde{p}_\alpha(\mathbf{z} \mid \mathbf{x})\right] \\
&= \operatorname*{argmax}_{\theta} \mathrm{ELBO}(q_\theta)
\end{aligned}$$

dropping the intractable term gives the ELBO

ELBO as Objective Function

What does maximizing ELBO(q_θ) accomplish?

$$\text{ELBO}(q_\theta) = \boxed{E_{q_\theta(\mathbf{z})}\left[\log p_\alpha(\mathbf{x}, \mathbf{z})\right]} \boxed{- E_{q_\theta(\mathbf{z})}\left[\log q_\theta(\mathbf{z})\right]}$$

1. The first expectation is high if q_θ puts probability mass on the same values of **z** that p_α puts probability mass

2. The second expectation is the entropy of q_θ and the negative entropy will be high if q_θ spreads its probability mass evenly

ELBO as lower bound

- *For a DGM*:
  - ELBO(q) is a lower bound for log p(x)
- *For a UGM*:
  - ELBO(q) is a lower bound for log Z(x)

Takeaway: in variational inference, we find the q that gives the **tightest bound** on the normalization constant for p(z | x)

Variational Inference

Whiteboard

- Evidence Lower Bound (ELBO)
- ELBO's relation to log p(x)

COORDINATE ASCENT VARIATIONAL INFERENCE (CAVI)

Variational Inference

Whiteboard

- Coordinate Ascent Variational Inference (CAVI) Algorithm
  - Connecting CAVI to BP and Gibbs sampling
  - Computing marginals from a trained mean field approximation
- CAVI algorithm derivation
  - Chain rule decomposition of log p(x, z)
  - Decomposing the entropy
  - Decomposing the ELBO
  - Derivatives and closed form solution