10-418/10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Variational Inference

Matt Gormley
Lecture 16
Oct. 31, 2022

Q&A

Q: The parameters of a K-dimensional Dirichlet($\boldsymbol{\alpha}$) are a vector $\boldsymbol{\alpha}$ of length K, so why are Dirichlet parameters sometimes given as a scalar? For example…

"We use a Dirichlet prior with parameter α = 0.1."

A: Great question!

A K-dimensional Dirichlet prior is said to be *symmetric* if all the values in the vector $\boldsymbol{\alpha}$ are the same, i.e. for all k, $\alpha_k = c$ where c is a scalar constant.

We sometimes call this restricted version the *symmetric Dirichlet distribution*.

Reminders

- Exam Rubrics and Exam Viewings
- Homework 4: MCMC
  - Out: Mon, Oct 24
  - Due: Fri, Nov 3 at 11:59pm
- Homework 5: ~~MCMC~~ VI
  - ~~Out: Mon, Oct 24~~
  - ~~Due: Fri, Nov 3 at 11:59pm~~

Reminders

Happy Halloween!

SEMANTIC SEGMENTATION

Case Study: Image Segmentation

- Image segmentation (FG/BG) by modeling of interactions btw RVs
  - Images are noisy.
  - Objects occupy continuous regions in an image.

[Nowozin,Lampert 2012]

Input image

Pixel-wise separate optimal labeling

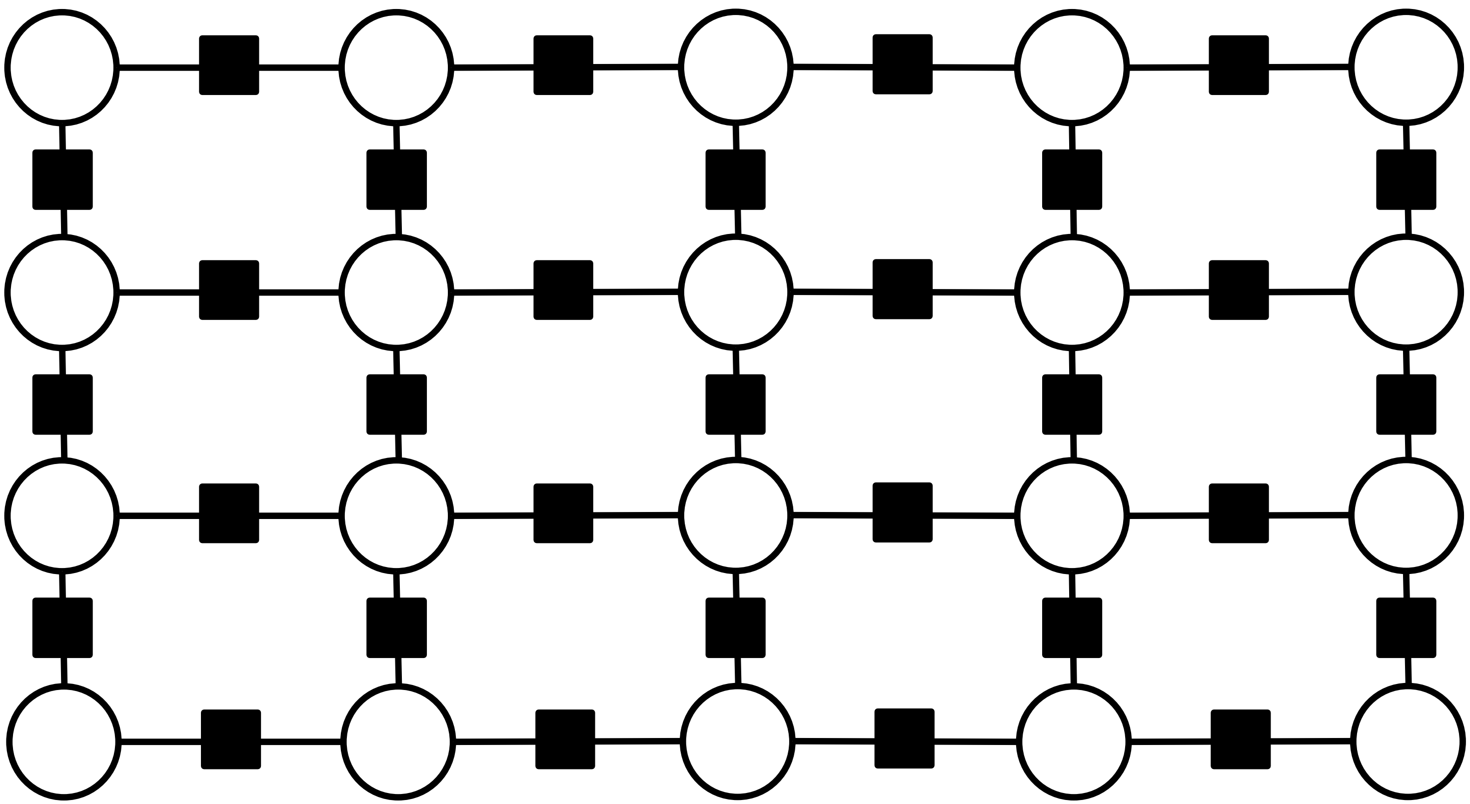
Locally-consistent joint optimal labeling

$$Y^* = \underset{y \in \{0,1\}^n}{\arg\max} \left[\underbrace{\sum_{i \in S} V_i(y_i, X)}_{\text{Unary Term}} + \underbrace{\sum_{i \in S} \sum_{j \in N_i} V_{i,j}(y_i, y_j)}_{\text{Pairwise Term}} \right].$$

Y: labels
X: data (features)
S: pixels
N_i: neighbors of pixel i

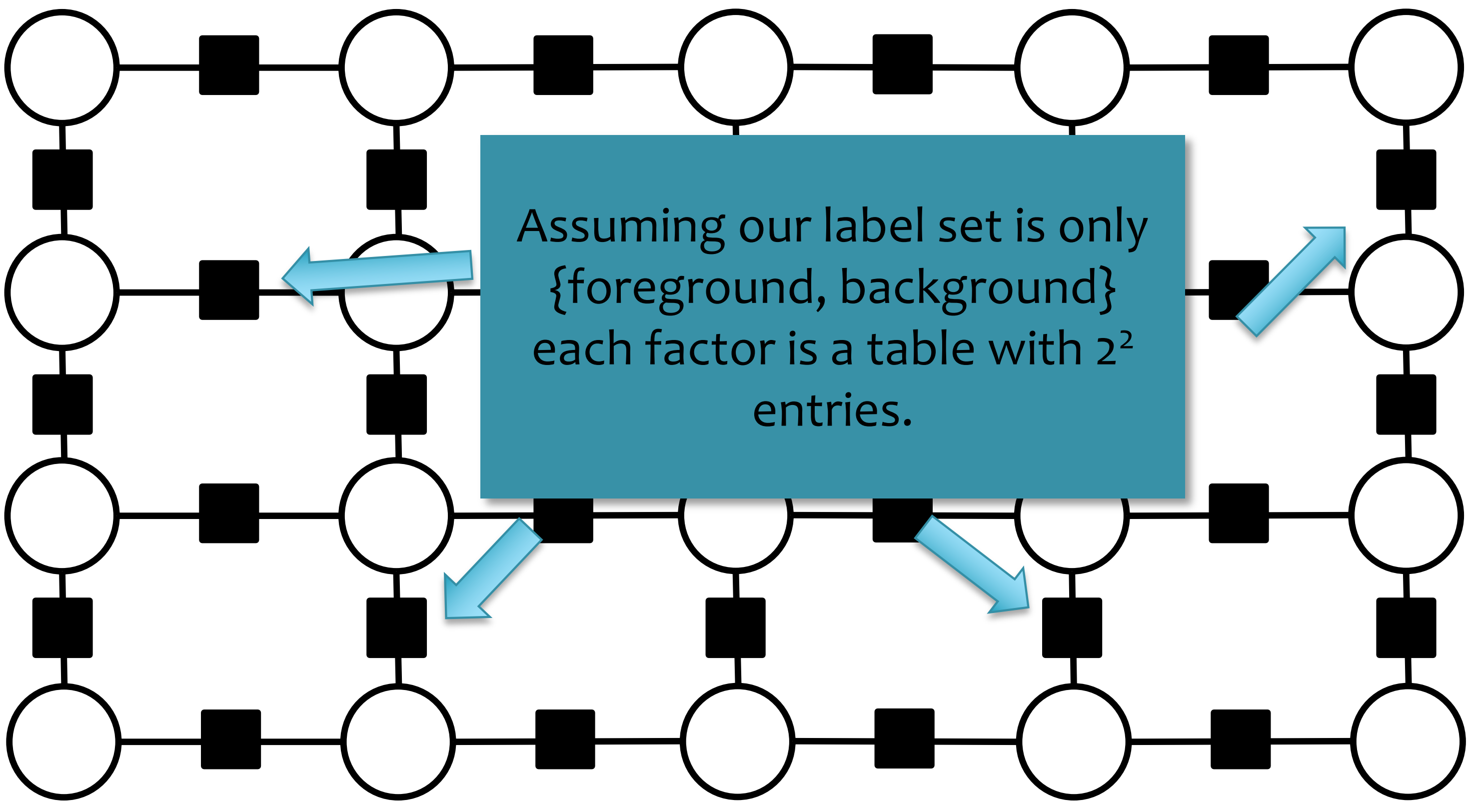
Grid CRF

- Suppose we want to image segmentation using a grid model



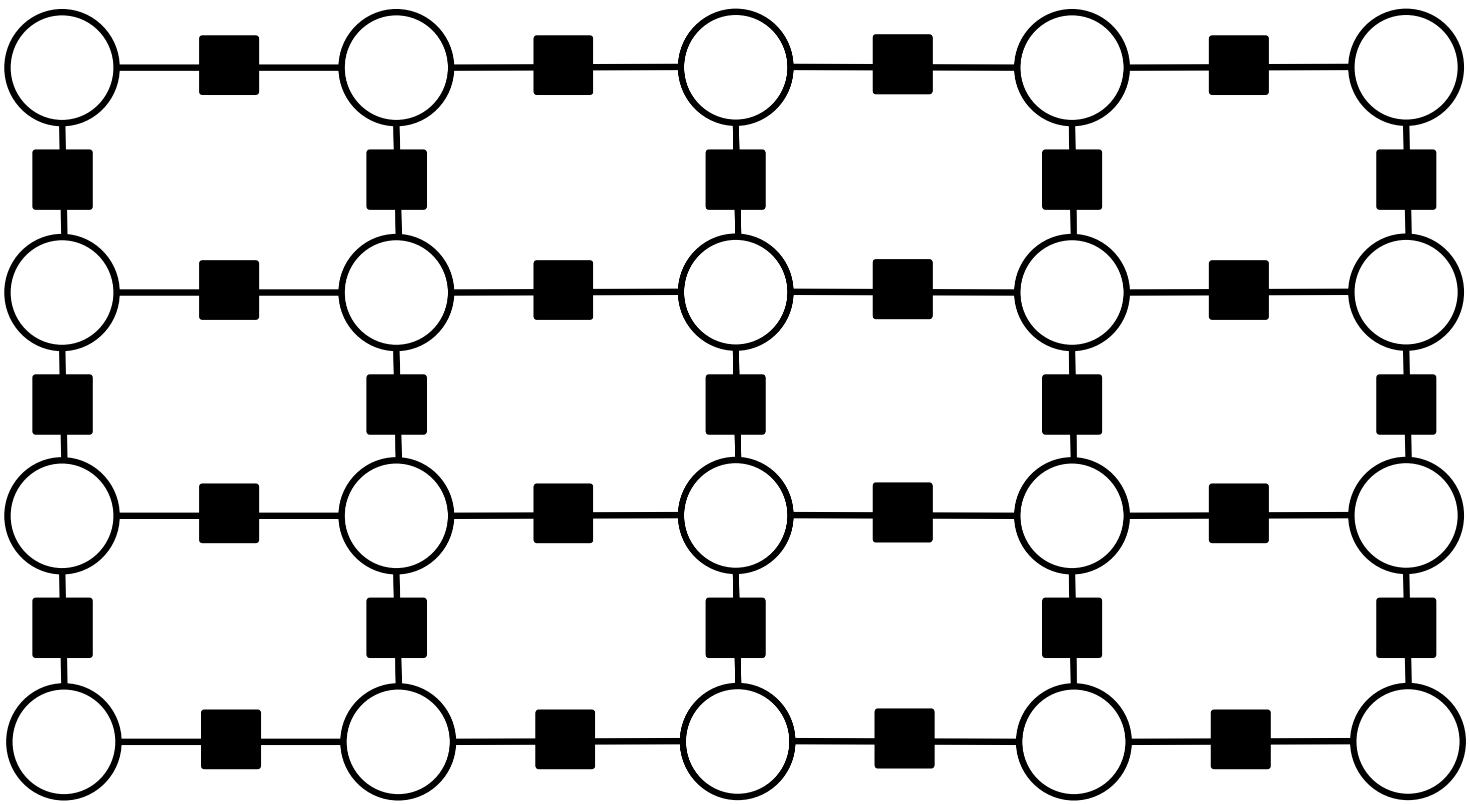
Grid CRF

- Suppose we want to image segmentation using a grid model



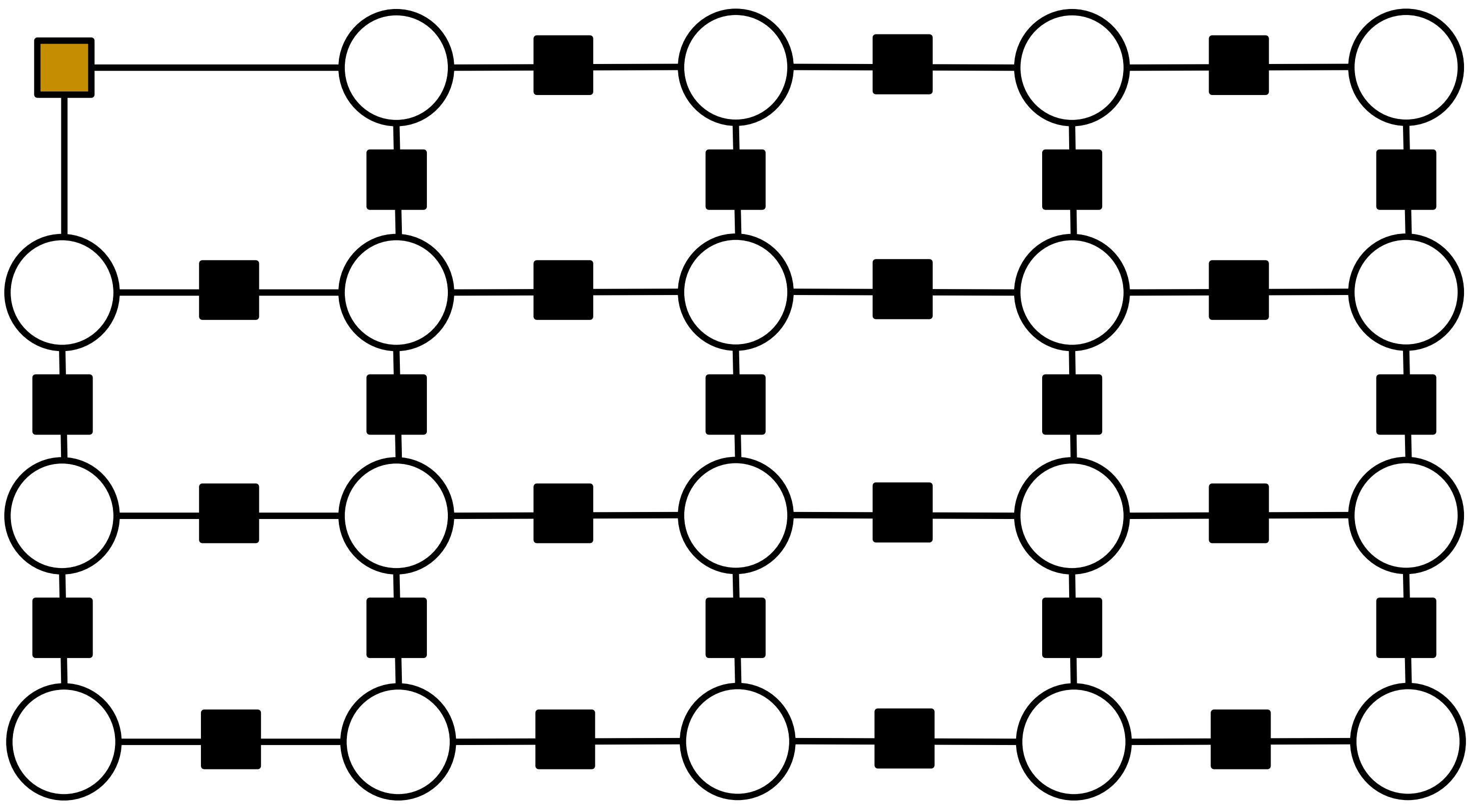
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



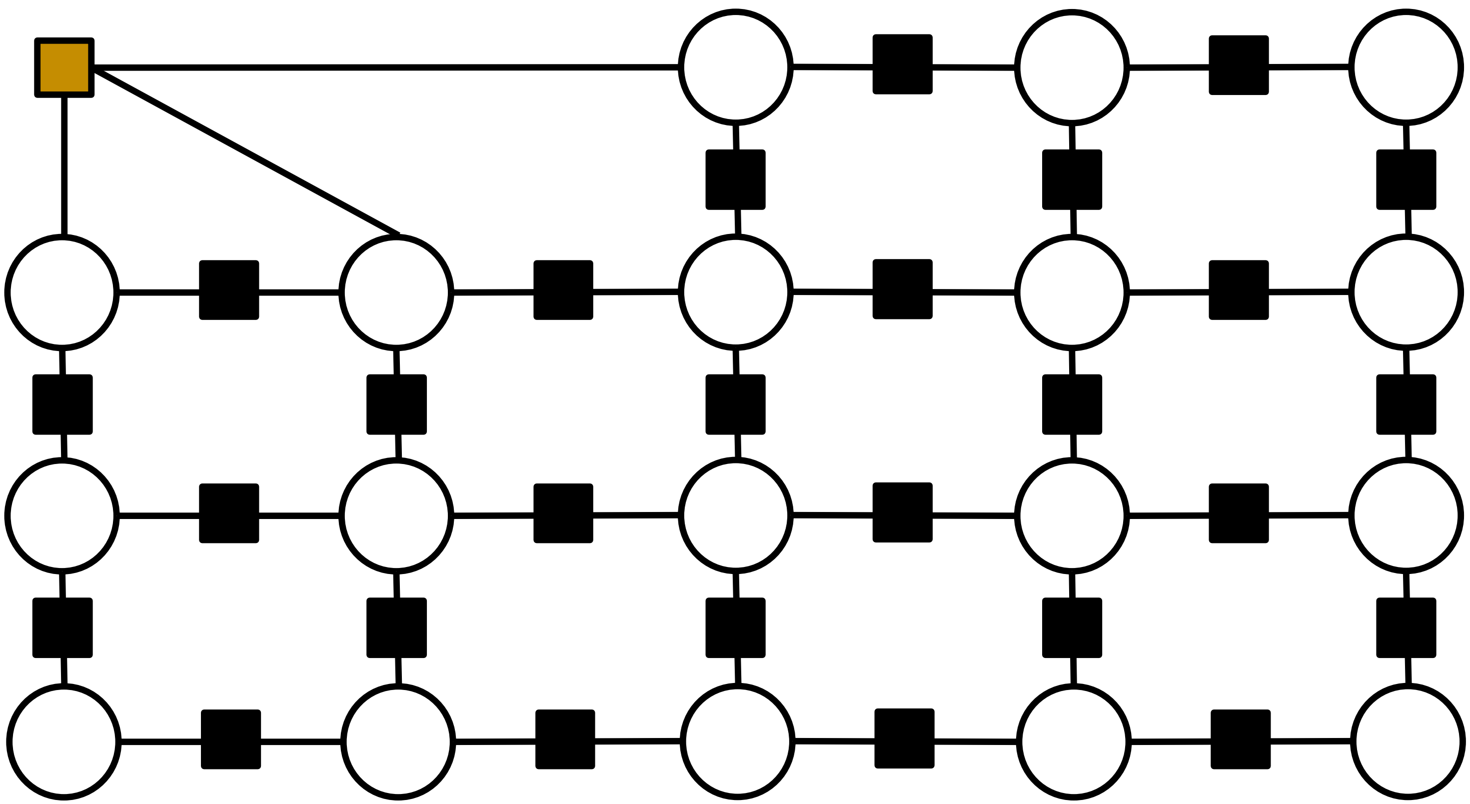
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



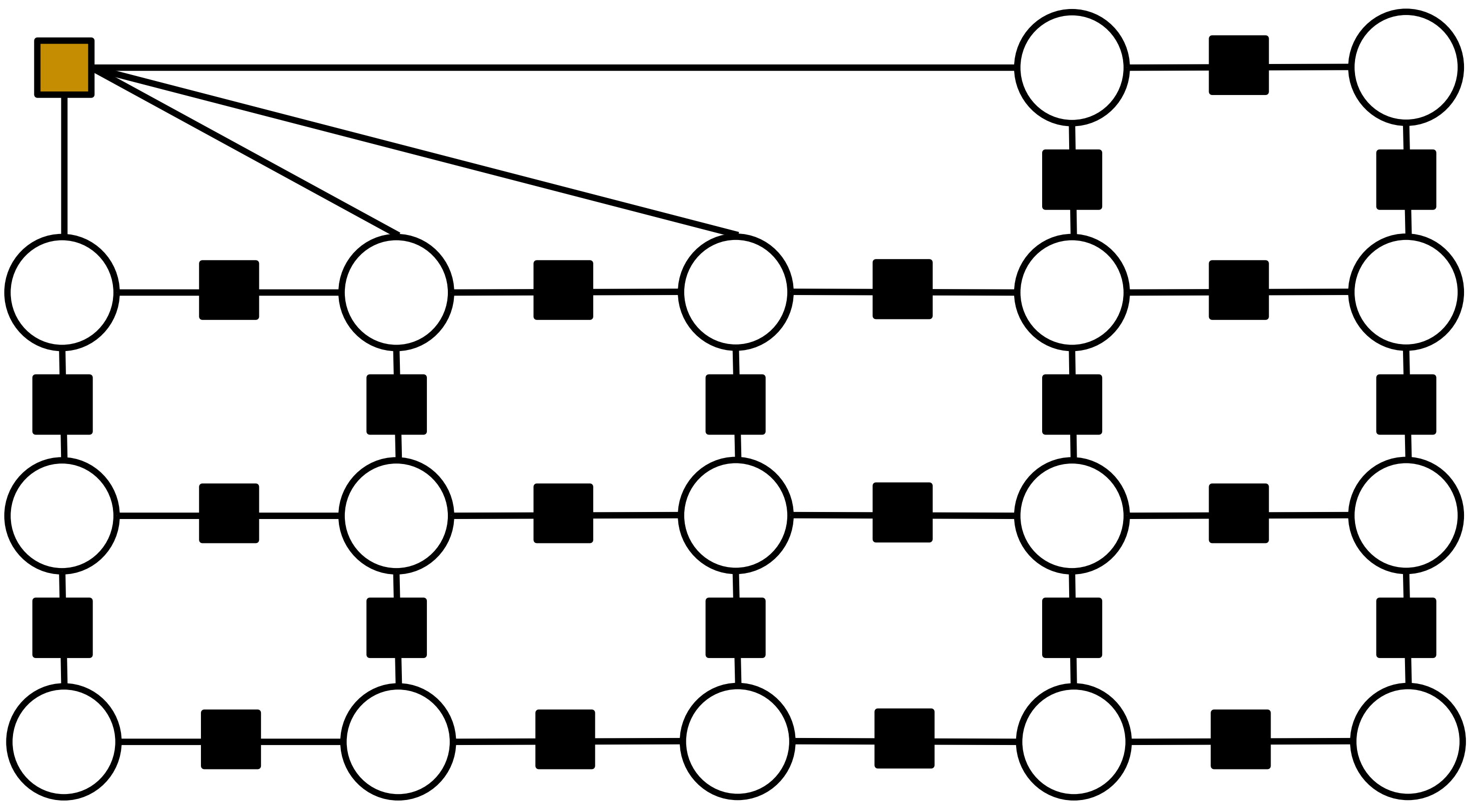
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



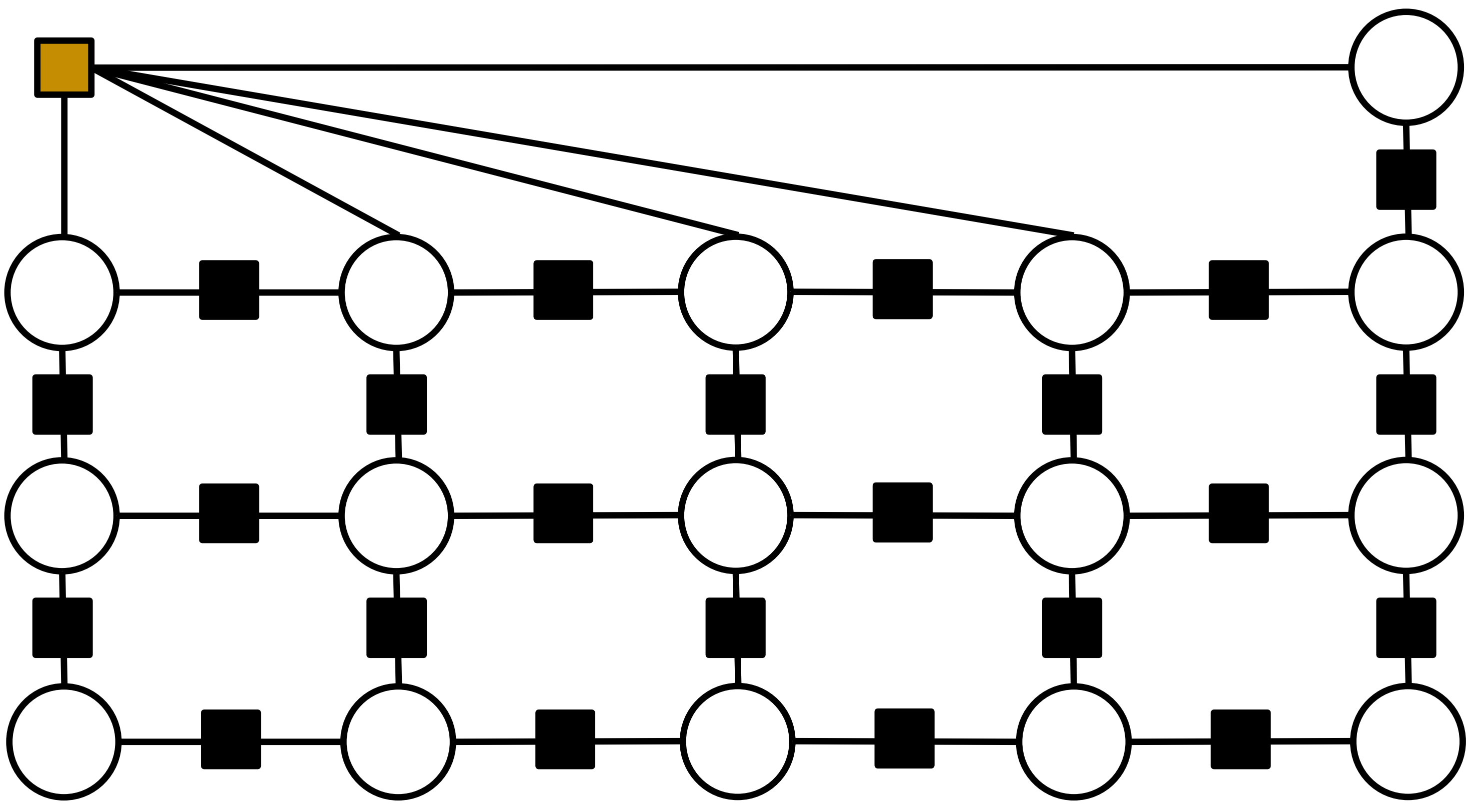
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



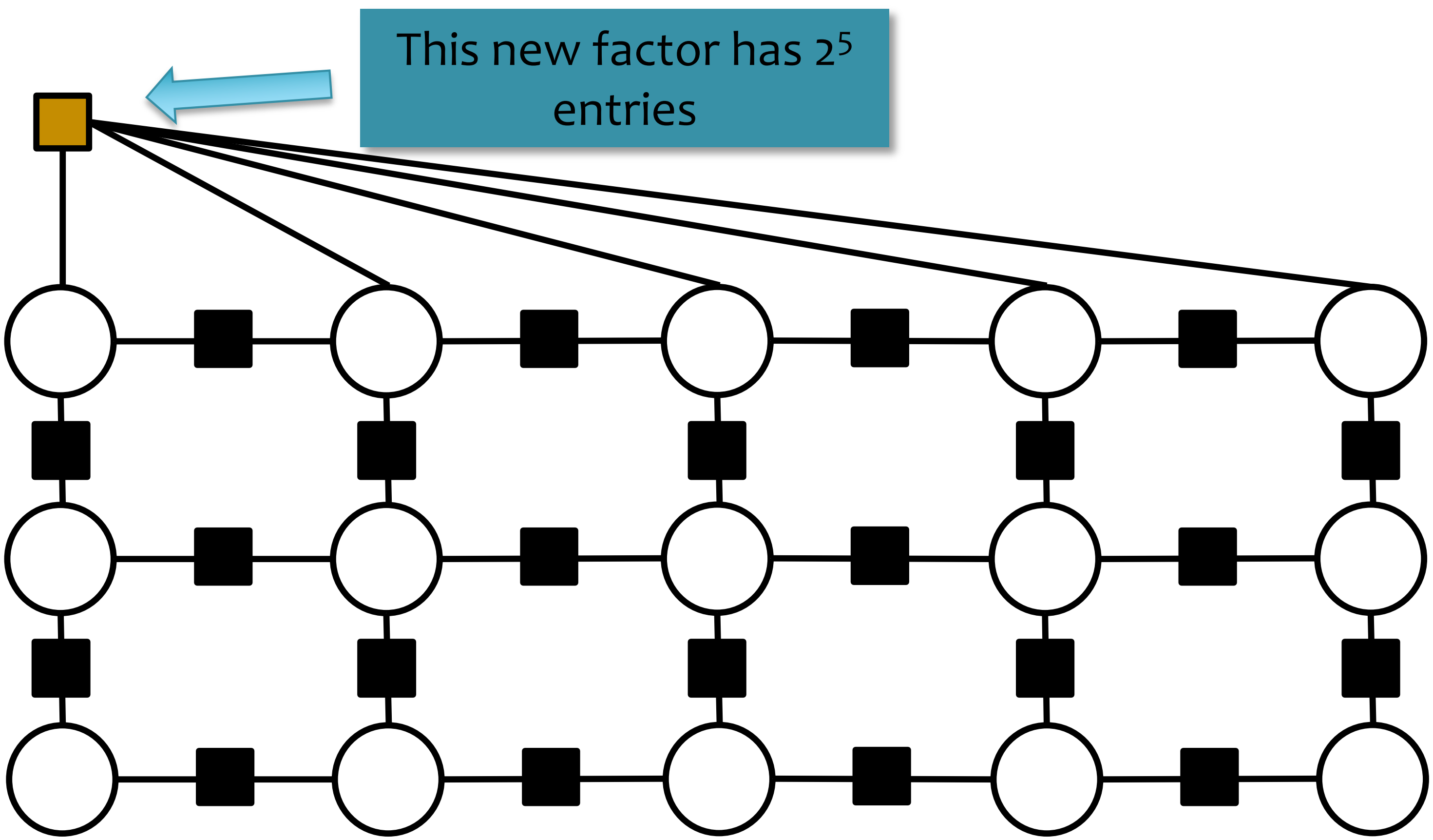
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



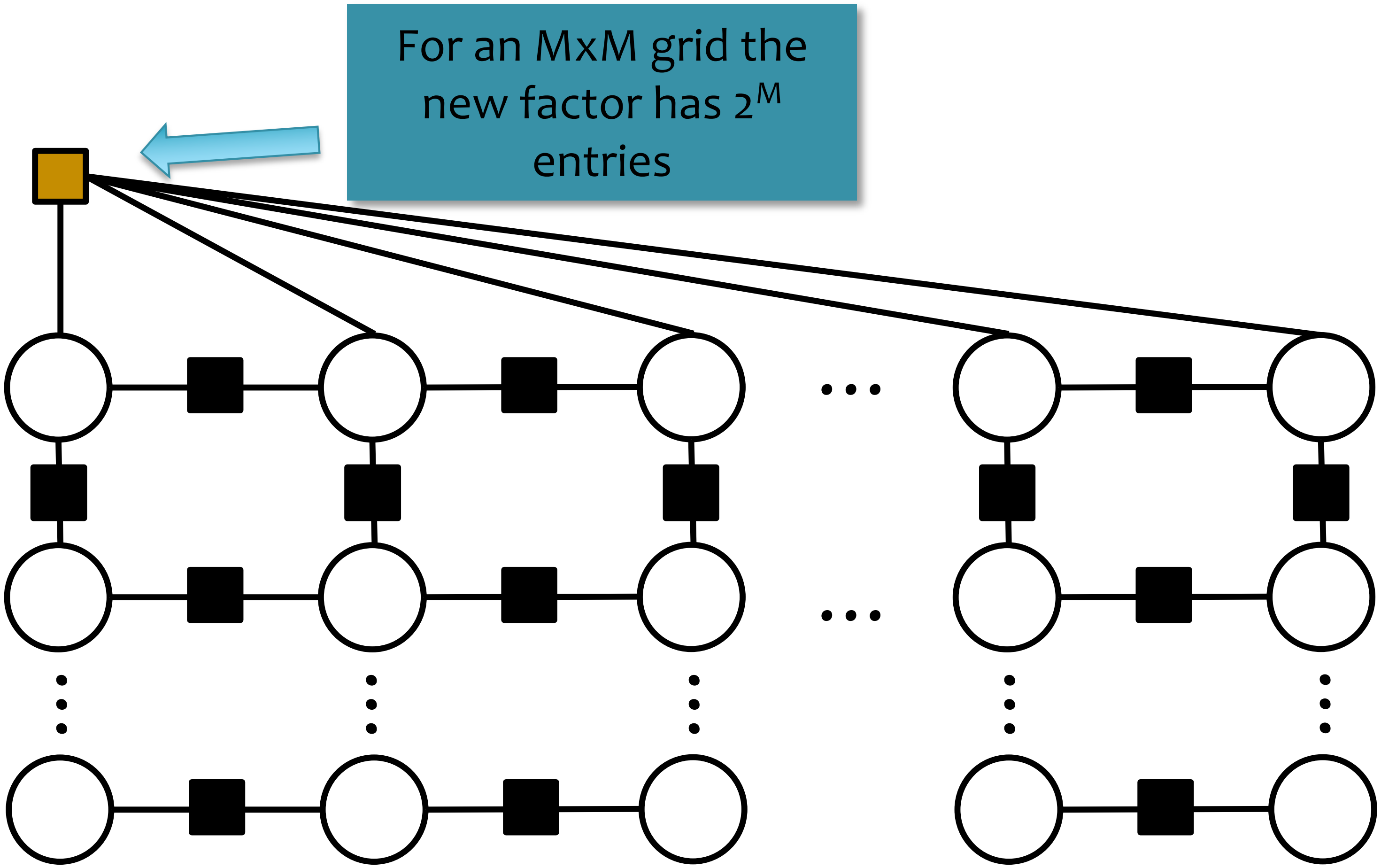
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

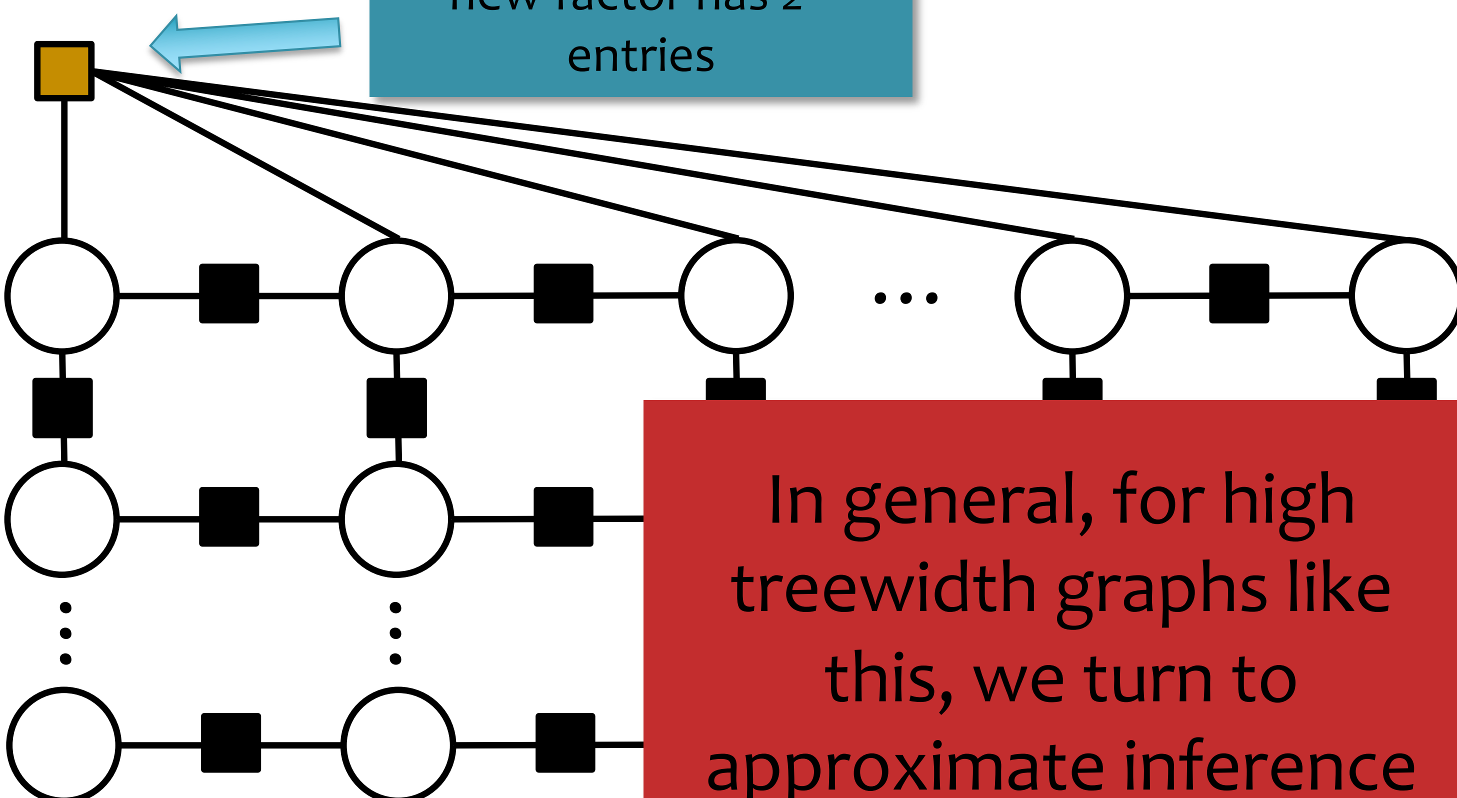
- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?

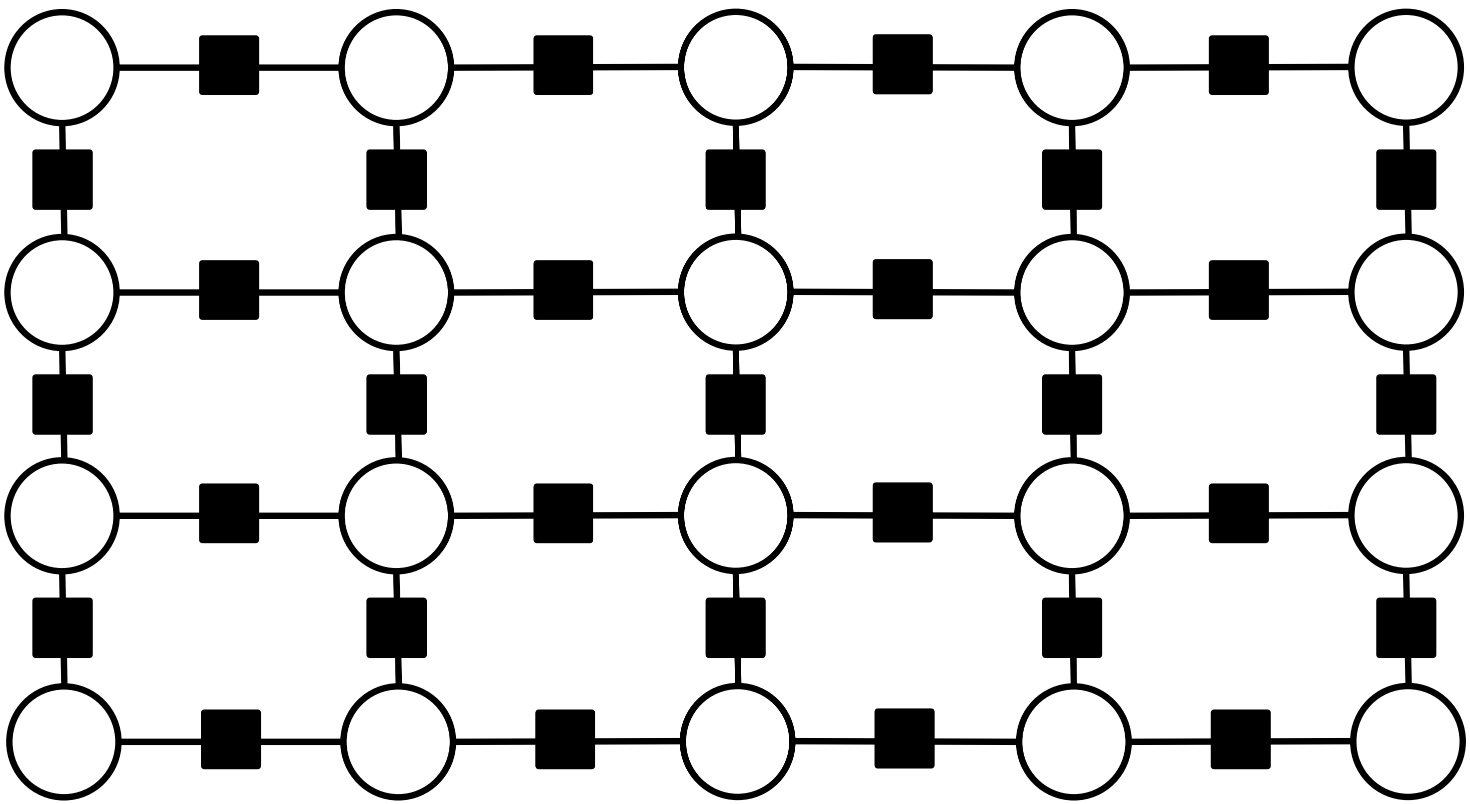
For an MxM grid the new factor has 2^M entries



In general, for high treewidth graphs like this, we turn to approximate inference (which we'll cover soon!)

Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?
- Can we instead run belief propagation to do exact inference?

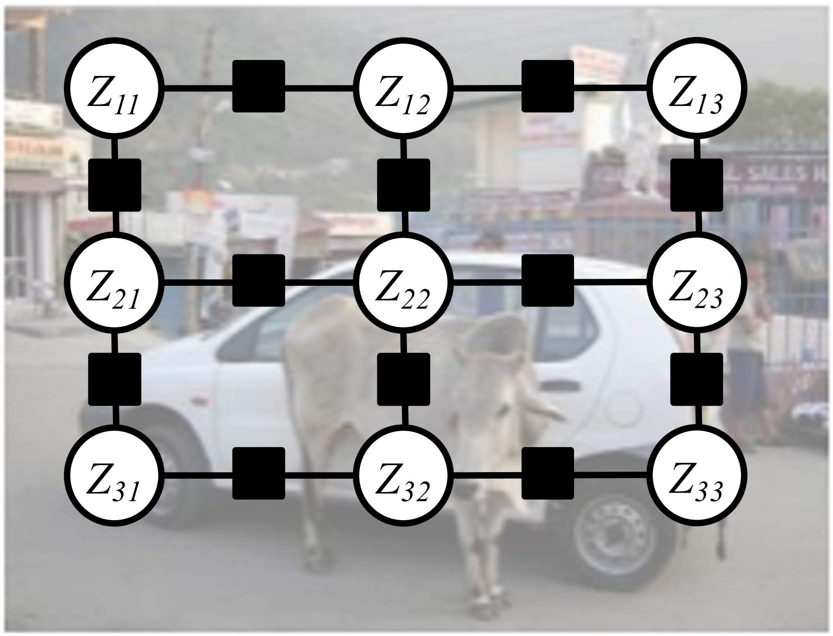


HIGH-LEVEL INTRO TO VARIATIONAL INFERENCE

Variational Inference

Problem:

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

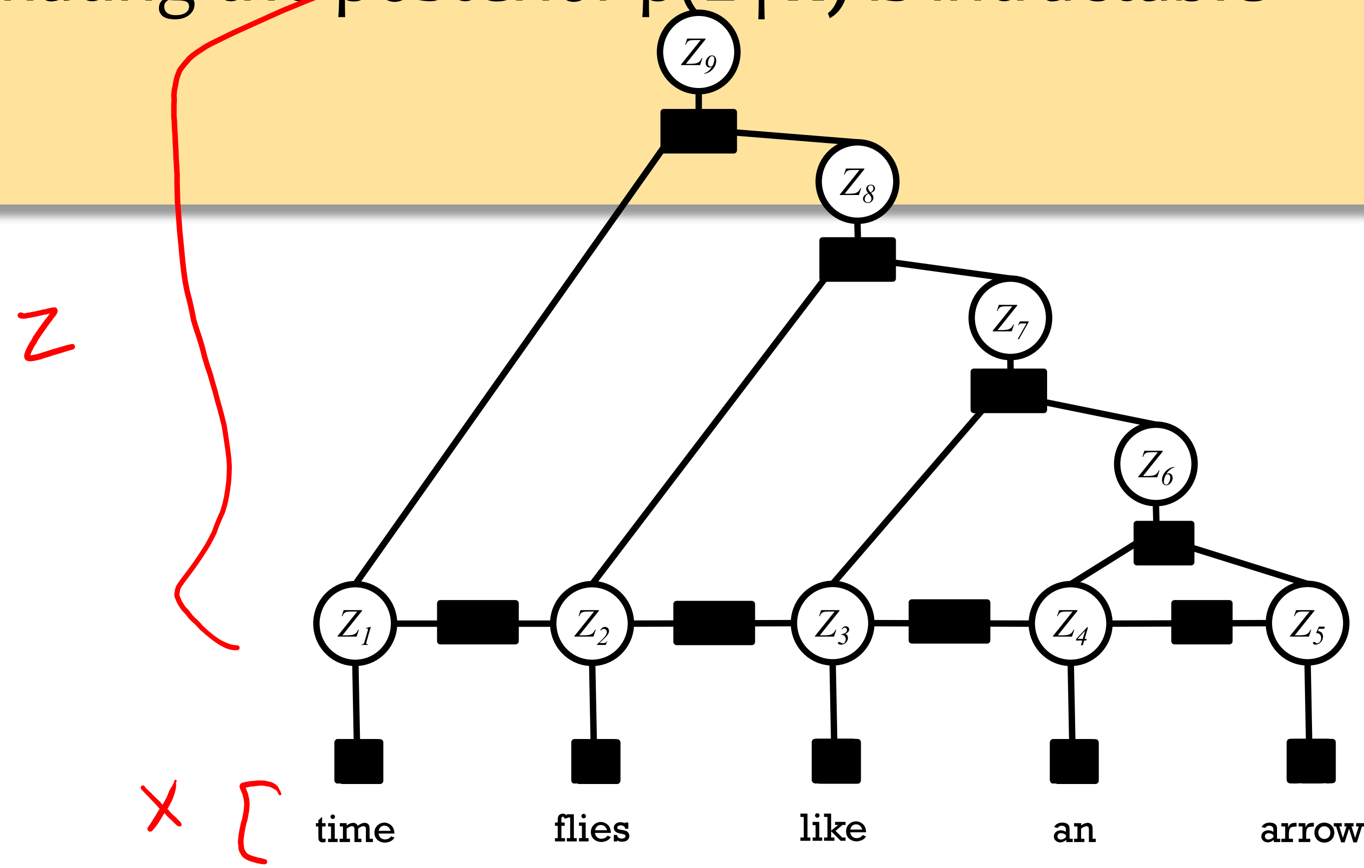


Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Problem:

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

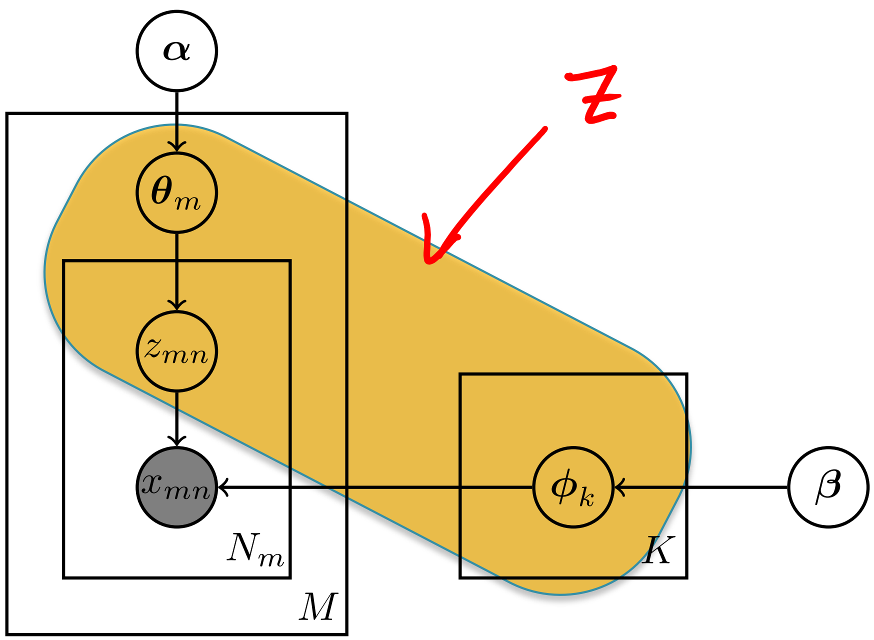


Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

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

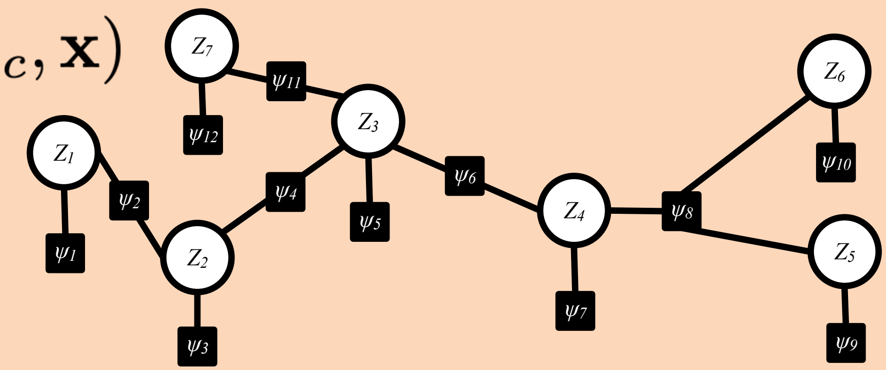
Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

TYPES OF VARIATIONAL APPROXIMATIONS

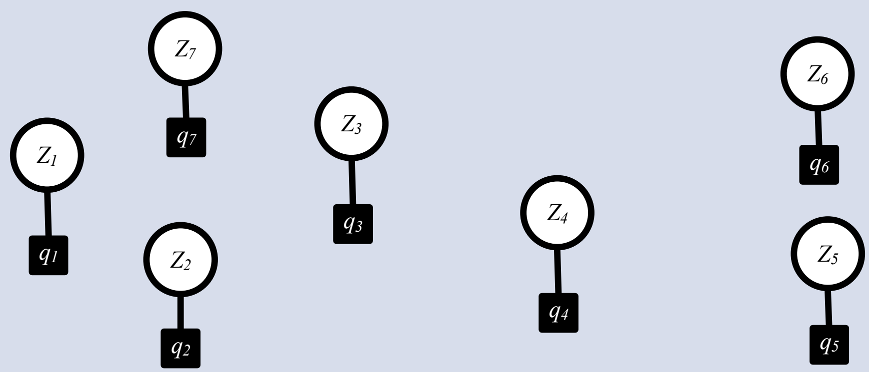
Mean Field Approximation

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

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



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

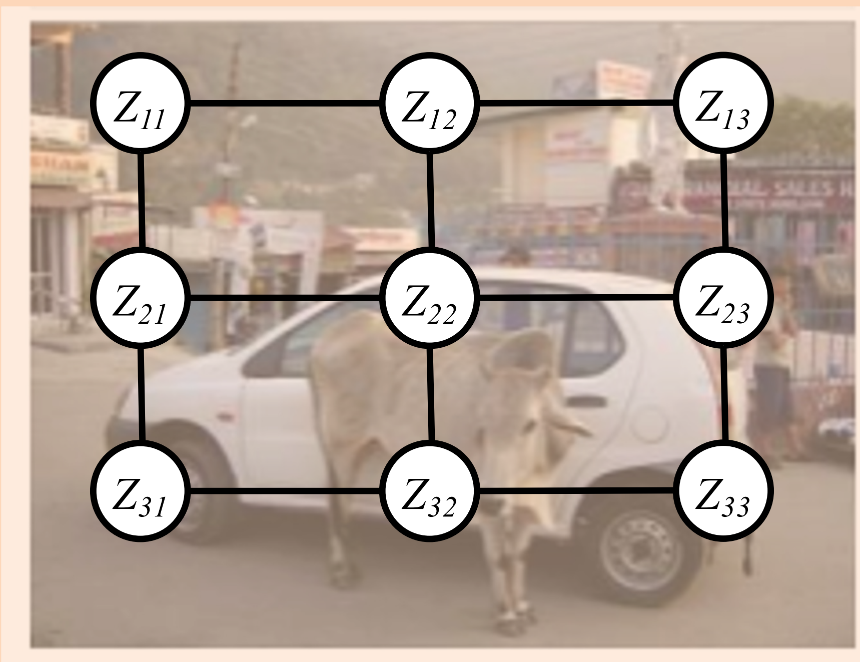


Mean Field Approximation

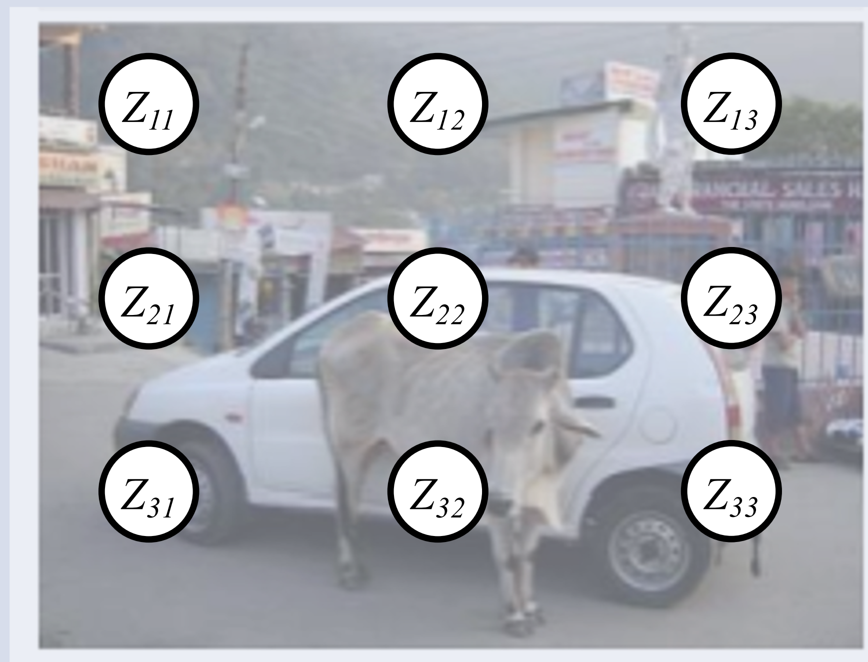
The **mean field approximation** assumes our variational approximation $q_\theta(\mathbf{z})$ treats each variable as independent

Ising Model

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



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



Structured Mean Field

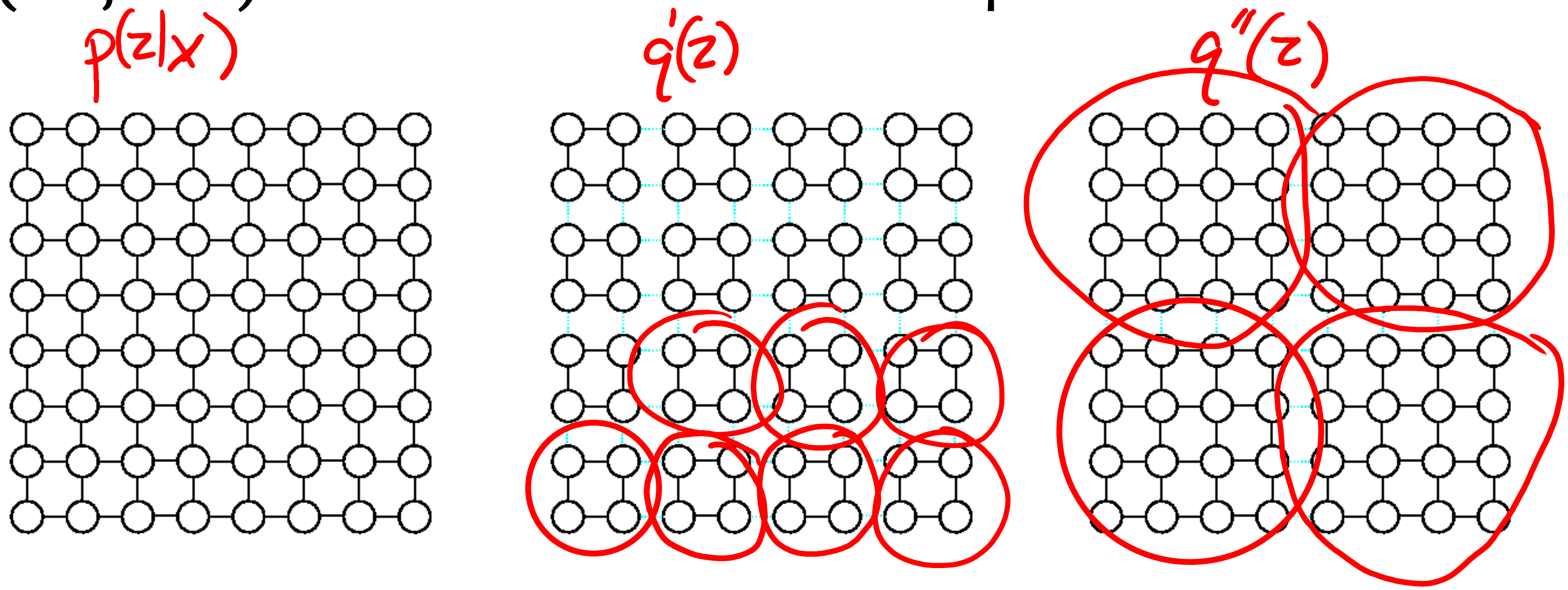
- If q is not a mean-field approximation, but decomposes over "blocks" of variables, then we have the **Structured Mean Field algorithm**

V.I.

- Connection to related algorithms:
  - This is analogous to Blocked Gibbs Sampling
  - This is analogous to Generalized Belief Propagation
  - The names here (Structured, Blocked, Generalized) are different b/c they were invented by different people and no-one thought to rename them all "Blocked"

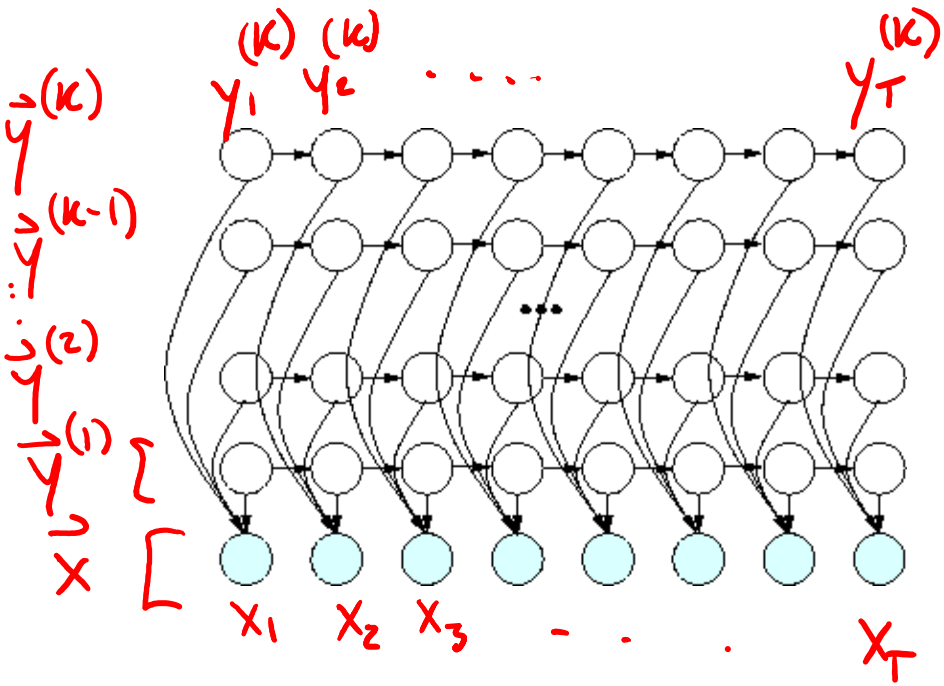
Structured Mean Field

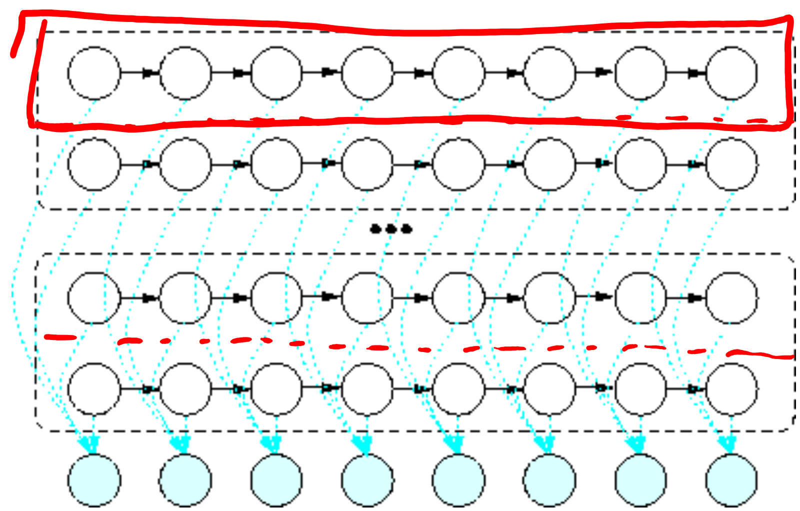
- We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:
- Instead of making all latent variables independent (i.e. naïve mean field, previous example), clusters of (disjoint) latent variables are independent.



Structured Mean Field

- For a factorial HMM, we could decompose into chains





$$\frac{p(y|x)}{Z}$$

$$q(y)$$

Collapsed vs. Uncollapsed V.I.

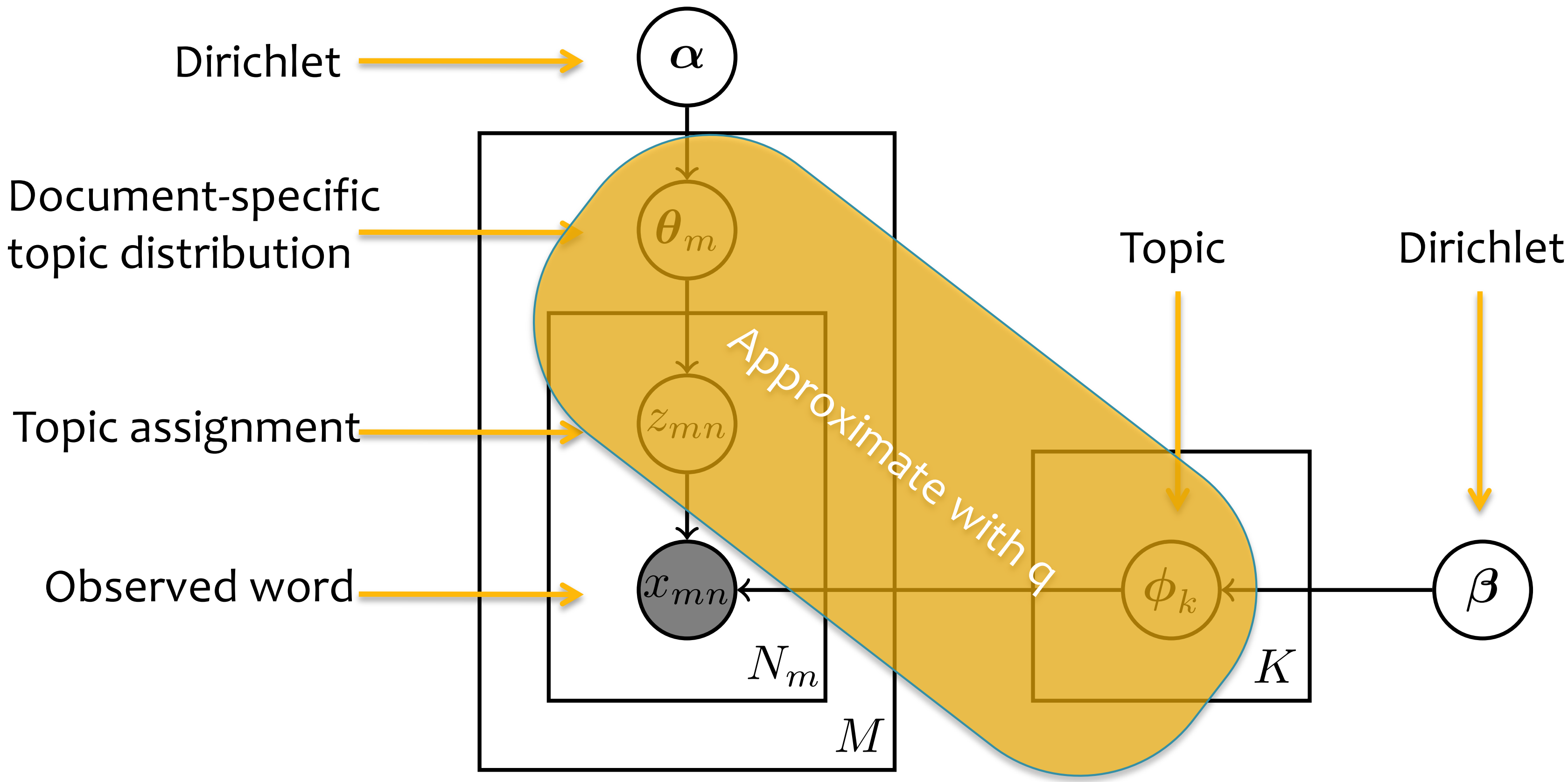
Just as we had collapsed and uncollapsed Gibbs samplers for LDA…

…we can have collapsed and uncollapsed variational inference for LDA

Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)

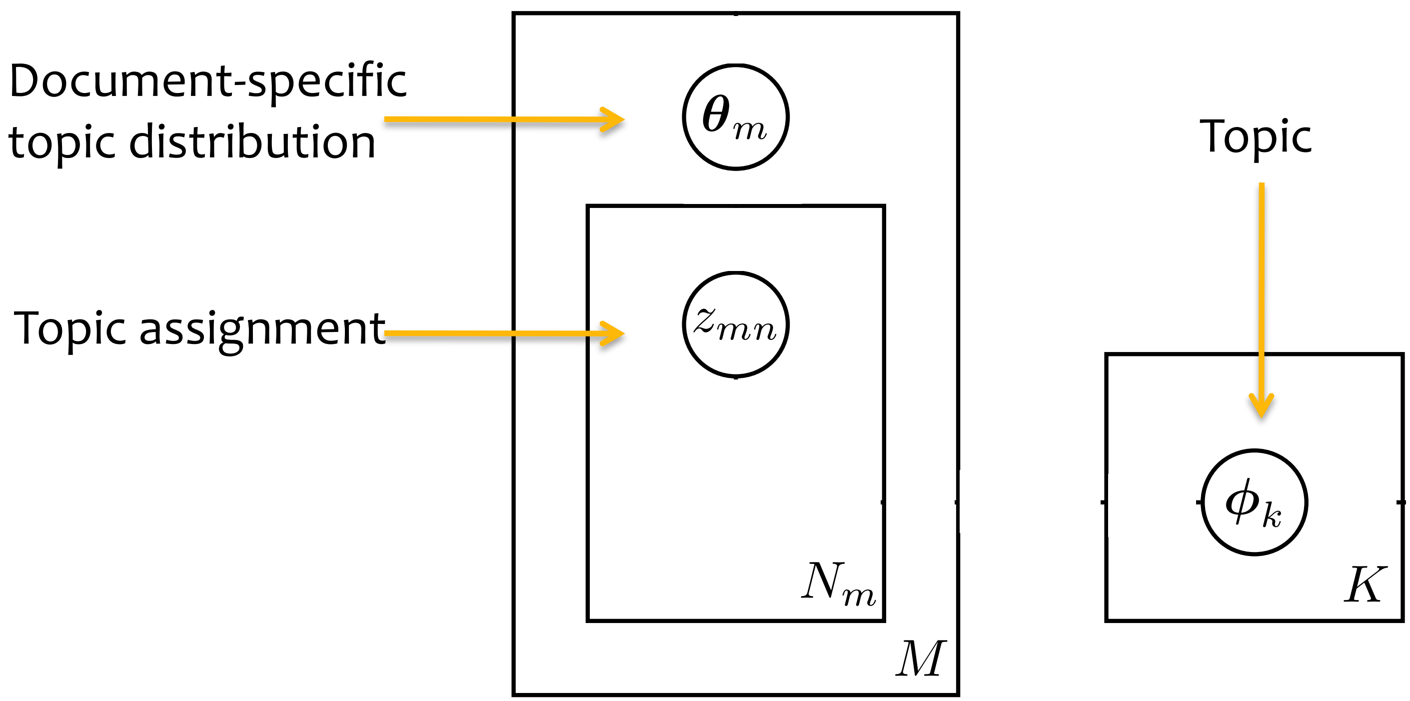
- Uncollapsed Variational Inference, aka. Explicit V.I. (original distribution)



Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)

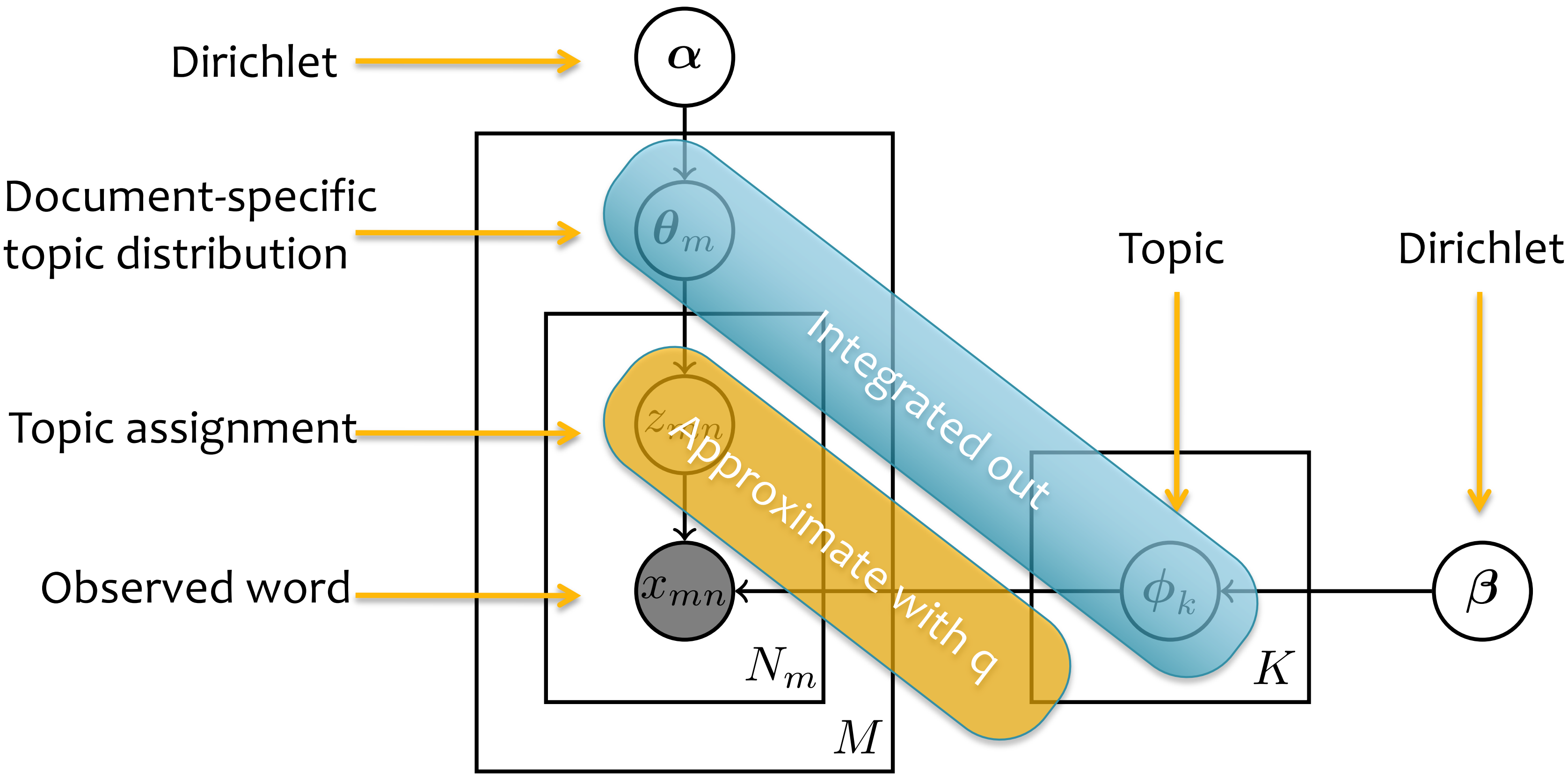
- Uncollapsed Variational Inference, aka. Explicit V.I. (mean field variational approximation)



Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)

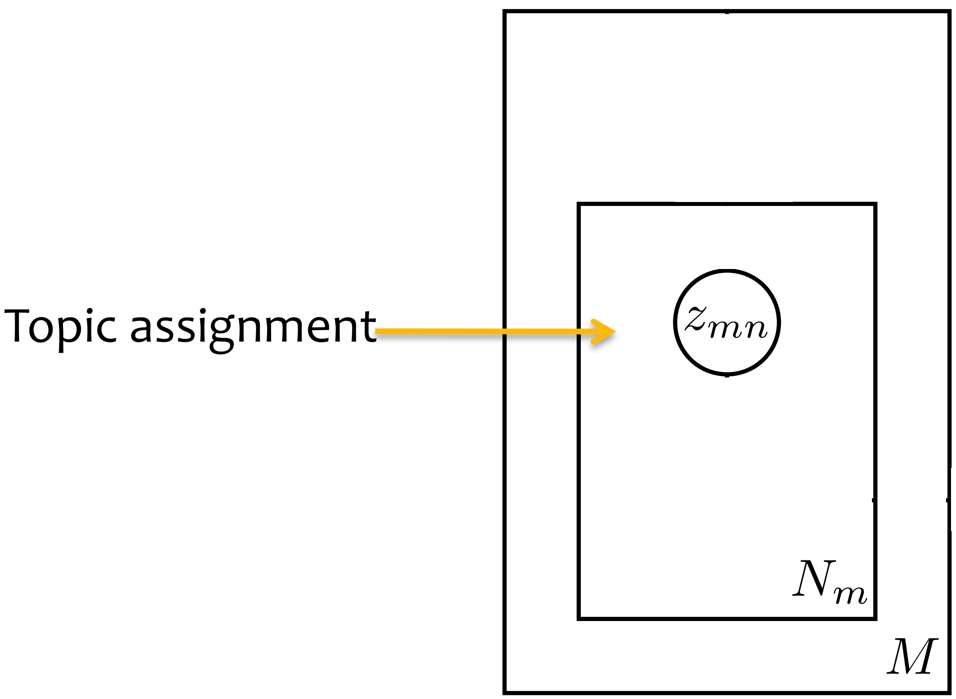
- Collapsed Variational Inference (original distribution)



Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)

- Collapsed Variational Inference
  (mean field variational approximation)



MEAN FIELD VARIATIONAL INFERENCE

Side Note

Contrast of three variational inference techniques:

1. Mean field variational inference
   minimizes KL (q || p)
2. Expectation propagation
   minimizes KL(p || q)
3. Loopy Belief Propagation
   minimizes the Bethe Free Energy

We are focused here on KL(q || p)

KL Divergence

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

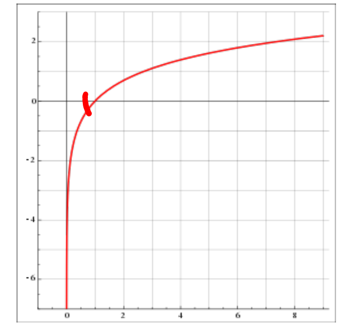
$$\text{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right] = \begin{cases} \sum_x q(x) \log \frac{q(x)}{p(x)} & \text{discrete } x \\ \int_x q(x) \log \frac{q(x)}{p(x)} dx & \text{cont. } x \end{cases}$$

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

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

$$\mathrm{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

KL Divergence



Understanding the Behavior of KL as an objective function

Example 1: Keeping all else constant, consider the effect of a particular x' on KL(q || p)

x'	q(x')	p(x')	q(x') log(q(x')/p(x'))	effect on KL(q || p)
1	0.9	0.9	0	no increase
2	0.9	0.1	1.97	big increase
3	0.1	0.9	-0.21	little decrease
4	0.1	0.1	0	little decrease

KL **does** insist on good approximations for values that have **high** probability in q

KL **does not** insist on good approximations for values that have **low** probability in q

Example 2: Which q distribution minimizes KL(q || p)?

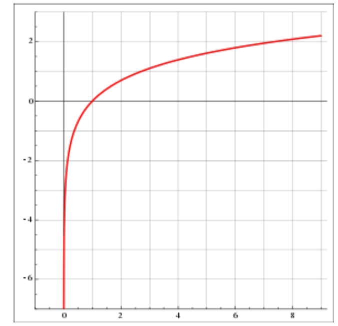
$$\mathbf{p} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad \mathbf{q}^{(1)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad \mathbf{q}^{(2)} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad \mathbf{q}^{(3)} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

Q: If we're minimizing KL, why not return $q^{(3)}$?
A: Because it's not a distribution!

$\mathbf{q}^{(2)}$ *minimizes KL*

$$\mathrm{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

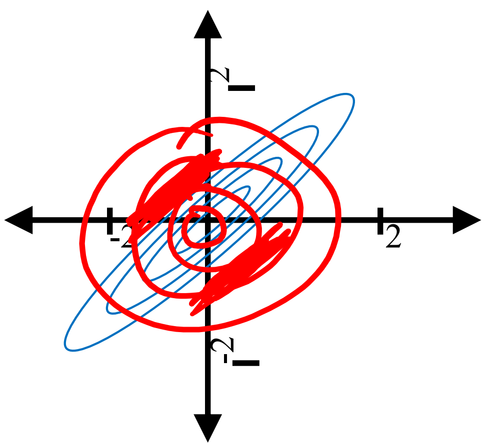
KL Divergence



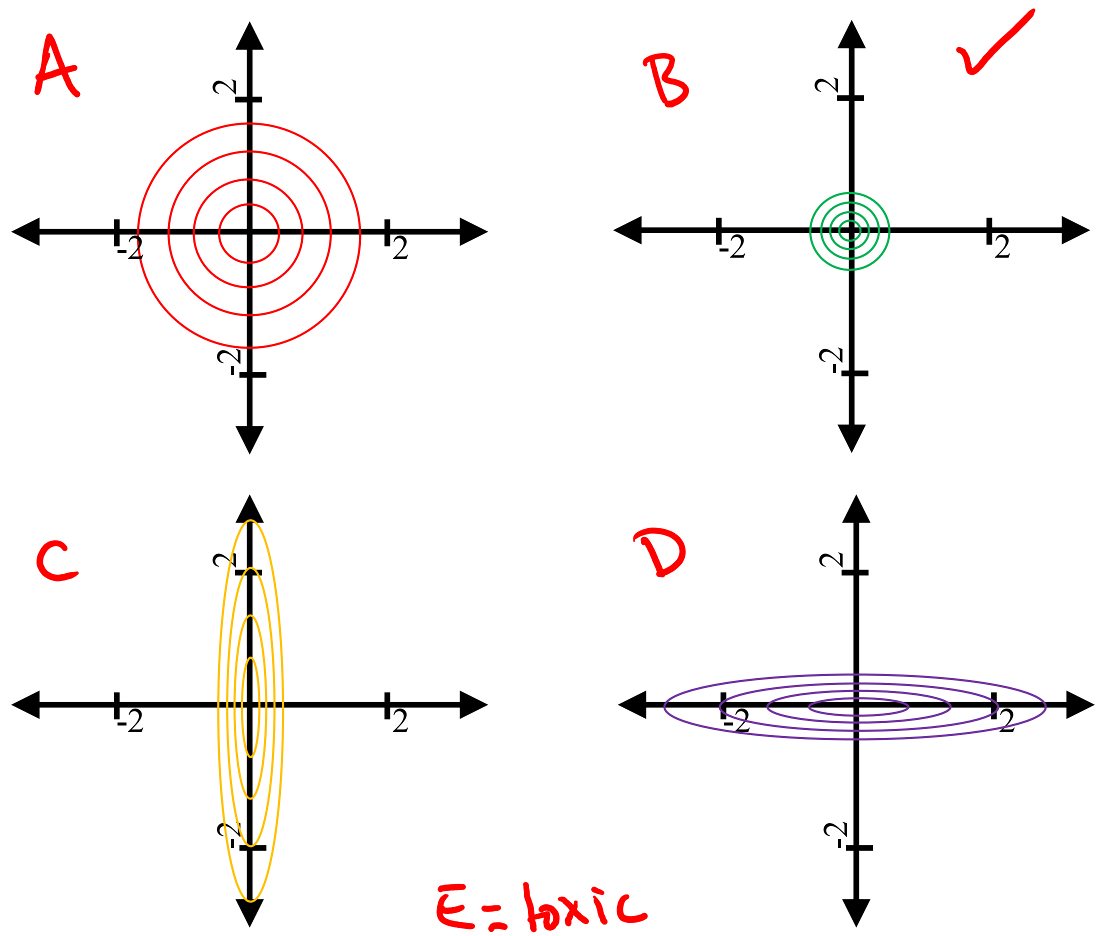
Understanding the Behavior of KL as an objective function

Example 3: Which q distribution minimizes KL(q || p)?

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu} = [0, 0]^T, \boldsymbol{\Sigma})$$



$$q(x_1, x_2) = \mathcal{N}_1(x_1 \mid \mu_1, \sigma_1^2)\mathcal{N}_2(x_2 \mid \mu_2, \sigma_2^2)$$



E = toxic

Two Cases for Intractability

- *Case 1*:

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

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

we assume p(x) is intractable

- *Case 2*:

  give **factor graph** and potentials

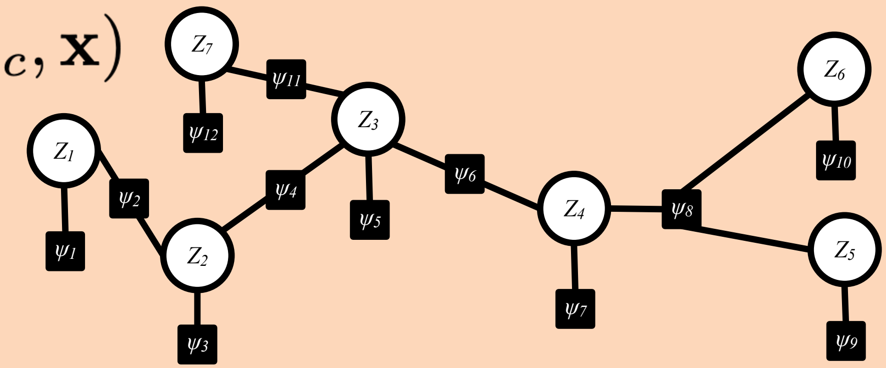
$$\Rightarrow p(z \mid x) = \frac{\tilde{p}(x, z)}{Z(x)}$$

we assume Z(x) is intractable

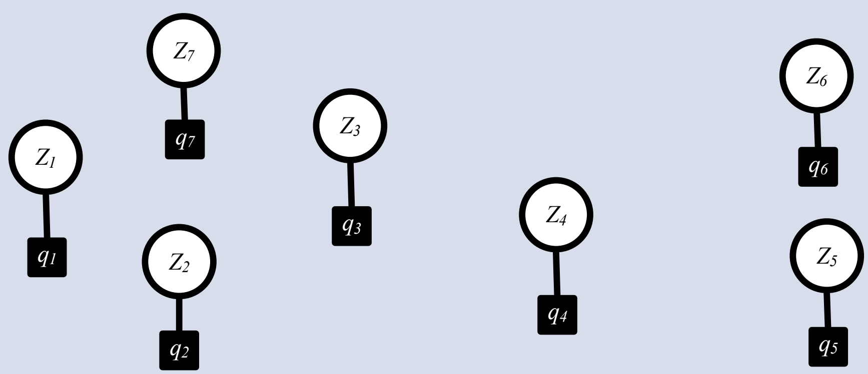
Mean Field Approximation

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

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

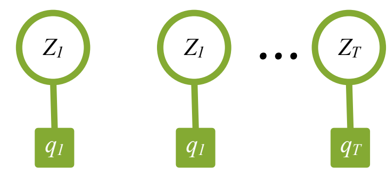


$$q_\theta(\mathbf{z}) = \prod_{t=1}^{T} q_t(z_t)$$



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)

$$\hat{q}(\mathbf{z}) = \operatorname*{argmin}_{q(\mathbf{z}) \in \mathcal{Q}} \mathrm{KL}(q(\mathbf{z}) || p(\mathbf{z} \mid \mathbf{x}))$$

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

equivalent

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

Optimizing KL Divergence

- *Question*: How do we minimize KL?

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

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

Why we can't compute KL...

$$\mathrm{KL}(q(\mathbf{z}) \,\|\, p(\mathbf{z} \mid \mathbf{x})) = E_{q(\mathbf{z})}\left[\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right)\right]$$

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

$$= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}[\log p(\mathbf{z} \mid \mathbf{x})]$$

$$= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}\boxed{[\log p(\mathbf{x}, \mathbf{z})] + E_{q(\mathbf{z})}[\log p(\mathbf{x})]}$$

$$= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] + \underbrace{\log p(\mathbf{x})}$$

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

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

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

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

Why we can't compute KL...

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

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?

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

- *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…

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

dropping the
intractable term
gives the ELBO

Optimizing KL Divergence

- *Question*: How do we minimize KL?

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

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

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

The ELBO for a UGM

Here is why…

$= \dfrac{\tilde{p}(z \mid x)}{Z(x)}$

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

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 term is the entropy of q_θ and the entropy will be high if q_θ spreads its probability mass evenly

ELBO as lower bound

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

<u>Takeaway</u>: 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)