

Mean Field Approximation

Kayhan Batmanghelich

Slides Credit (Partially adopted from):

- CSC 412 (UofT): Zemel & Urtasun
- Shakir Mohamed (DeepMind)

Inferential Problems

$$\underbrace{\text{Posterior}}_{p(z|y)} = \frac{\overbrace{\text{Likelihood}}_{p(y|z)} \overbrace{\text{Prior}}_{p(z)}}{\underbrace{\int p(y, z) dz}_{\text{Marginal likelihood/ Model evidence}}}$$

$$P(y \in A | x)$$
$$\equiv E[\mathbb{1}(y \in A) | x]$$

Most inference problems will be one of:

Marginalisation

$$p(y) = \int p(y, \theta) d\theta$$

Expectation

$$\mathbb{E}[f(y) | x] = \int f(y) p(y | x) dy$$

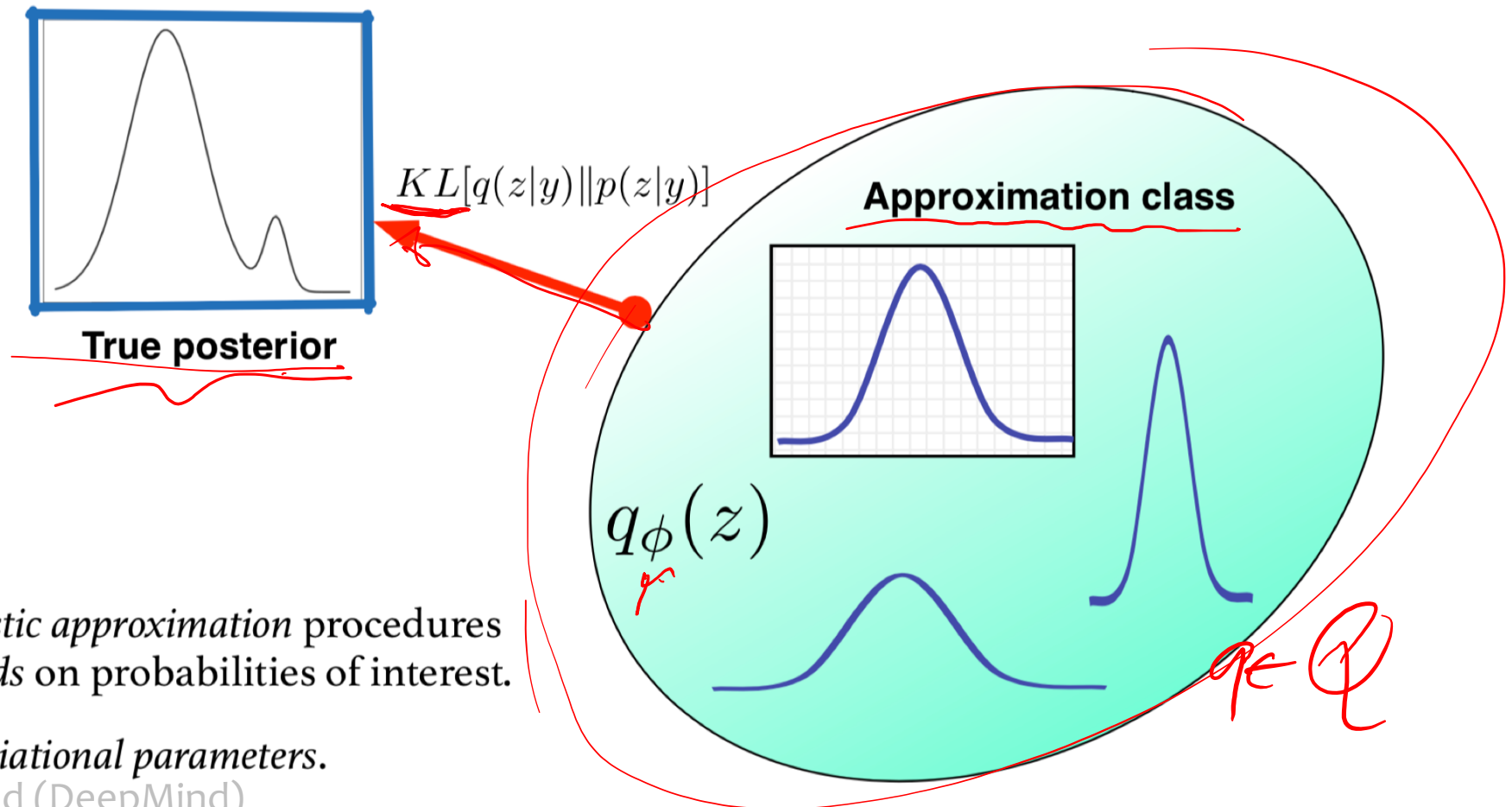
Prediction

$$p(y_{t+1}) = \int p(\overline{y_{t+1}} | y_t) p(y_t) dy_t$$

Variational Methods

Variational Principle

General family of methods for approximating complicated densities by a simpler class of densities.



Deterministic approximation procedures with *bounds* on probabilities of interest.

Fit the *variational parameters*.

Variational Calculus

Called a variational method because it derives from the
Calculus of Variations

Functions:

- Variables as input, output is a value.
- Full and partial derivatives $\frac{df}{dx}$
- E.g., Maximise likelihood $p(x|\theta)$ w.r.t. parameters θ

We exploit both types of derivatives in variational inference.

Functionals:

- Functions as input, output is a value.
- Functional derivatives $\frac{\delta F}{\delta f}$
- E.g., Maximise the entropy $H[p(x)]$ w.r.t. $p(x)$

Variational Calculus

$$\int \delta(x) dx = 1$$

$$f(\theta)$$

$$\frac{\delta f}{\delta \theta}$$

Two basic rules

• Functional derivative:

$$\frac{\delta f(x)}{\delta f(x')} = \delta(x - x')$$

• Commutative rule:

$$\frac{\delta}{\delta f(x')} \frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x} \frac{\delta f(x)}{\delta f(x')}$$

$$\frac{\delta f(x)}{\delta f(x')}$$

Simple Example: Maximize the entropy w.r.t $p(x)$

$$\max_{p(x) \in \mathcal{P}} H[p(x)]$$

$$\frac{\delta H[p(x)]}{\delta p(x)} = -\frac{\delta}{\delta p(x)} \int p(x) \log p(x) dx$$

$$= -\int p(x) \frac{1}{p(x)} \delta(x - x') dx' - \int \log p(x) \delta(x - x') dx'$$

$$= -1 - \log p(x)$$

Variational Methods

Unknown Data

- **Goal:** Approximate a difficult distribution $p(x|e)$ with a new distribution $q(x) \in \mathcal{Q}$
 - $p(x|e)$ and $q(x)$ should be “close”
 - Computation on $q(x)$ should be easy
- How should we measure distance between distributions?
- The Kullback-Leibler divergence (KL-divergence) between two distribution p and q is defined as

$$D(p||q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

- It measures the expected number of **extra bits (nats)** required to describe samples from $p(x)$ using a code based on q instead of p
- $D(q||q) \geq 0$ for all p, q with equality if and only if $p = q$
- The KL-divergence is asymmetric

$$p(x|e) = \frac{p(x,e)}{p(e)} = \frac{p(x,e)}{\int p(x,e) dx}$$

Variational Inference

$$p(x|e) \propto \frac{p(x,e)}{\tilde{p}(x)}$$

$$\frac{p(x|e)}{p(x)} = \frac{\tilde{p}(x)}{Z}$$

Let's look at the unnormalized distribution

Give me an example of a normalizer

$$J(q) = \sum_x q(x) \log \frac{q(x)}{\tilde{p}(x)}$$

$$= \sum_x q(x) \log \frac{q(x)}{Z \cdot p(x)}$$

$$J(q) = \cancel{KL(q||p)} - \log Z$$

$$J(q) \leq -\log Z$$

$$= \sum_x q(x) \log \frac{q(x)}{p(x)} - \log Z$$

$$= \underbrace{KL(q||p)}_{\text{Non-negative}} - \log Z$$

$$J(q) \geq -\log Z$$

$$-J(q) \leq \log Z$$

Since Z is constant, by minimizing J(q), we will force q to become close to p

Let's repeat that again ...

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} KL(q(\mathbf{z}) || p(\mathbf{z} | \mathcal{D}))$$

$$KL(q(\mathbf{z}) || p(\mathbf{z} | \mathcal{D})) = \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{z} | \mathcal{D})]$$

$$KL(q(\mathbf{z}) || p(\mathbf{z} | \mathcal{D})) = \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{z}, \mathcal{D})] + \log p(\mathcal{D})$$

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(\mathbf{z}, \mathcal{D})] - \mathbb{E}_q[\log q(\mathbf{z})]$$

$$KL(q(\mathbf{z}) || p(\mathbf{z} | \mathcal{D}))$$

$$\text{ELBO}(q)$$

$$\log p(\mathcal{D})$$



$$\frac{p(\mathbf{z}, \mathcal{D})}{p(\mathcal{D})}$$

evidence

Alternative Interpretations

$\min_{q \in \mathcal{Q}} J(q)$
 $\max_{q \in \mathcal{Q}} -J(q)$
 $\max_{q \in \mathcal{Q}} H(q) - \mathbb{E}_q[\cdot]$

$\tilde{p}(x) = p(x)$
 $\tilde{p}(x) \geq 0$

$$\begin{aligned}
 J(q) &= \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{\tilde{p}(\mathbf{x})} \\
 &= \mathbb{E}_q[\log q(\mathbf{x})] + \mathbb{E}_q[-\log \tilde{p}(\mathbf{x})] = \underbrace{-\mathbb{H}(q)}_{\text{entropy}} + \underbrace{\mathbb{E}_q[E(\mathbf{x})]}_{\text{Energy}} \\
 &\quad \text{variational free energy or Helmholtz free energy}
 \end{aligned}$$

View 1: Minimize expected energy while maximizing the entropy

$J(q) = \mathbb{E}_q[\log q(\mathbf{x}) - \log p(\mathbf{x}) p(\mathcal{D})]$
 $= \mathbb{E}_q[\log q(\mathbf{x}) - \log p(\mathbf{x}) - \log p(\mathcal{D})]$
 $= \mathbb{E}_q[-\log p(\mathcal{D})] + KL(q||p)$

$\tilde{p}(x) = p(x)$
 $\tilde{p}(x) \geq 0$

$ELBO$
 $J(q)$
 $\log p(\mathcal{D})$

View 2: Expected "Evidence" plus a penalty term that measures how far apart the two distributions are

$$KL(q||P) = \mathbb{E}_q \left[\log \frac{q}{P} \right]$$

$$KL(P||q)$$

Forward or Reverse KL

Which direction of KL divergence

- Suppose p is the true distribution

$$D(p||q) = \sum_{\mathbf{x}} \boxed{p(\mathbf{x})} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})} \right]$$

p are typically intractable
How can I sample from it?

- What about the reverse direction

$$D(q||p) = \sum_{\mathbf{x}} \boxed{q(\mathbf{x})} \log \frac{q(\mathbf{x})}{\boxed{p(\mathbf{x})}}$$

How I don't know how to evaluate it?

$KL(q||p)$

$KL(p||q)$

$$\mathbb{E}_{\mathbf{x} \sim q} \left[\log \frac{q(\mathbf{x})}{\frac{p(\mathbf{x})}{Z}} \right]$$

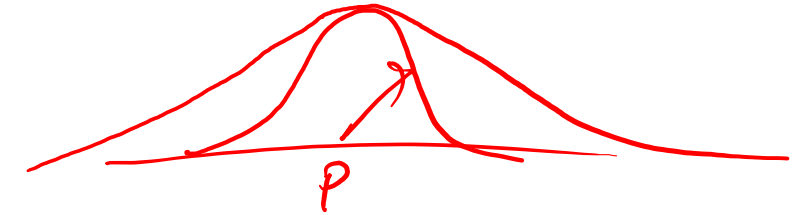
$$A := \{x \mid p(x) = 0\}$$

$$B := \{x \mid q(x) = 0\}$$

$$A \subset B$$

Which Direction of KL?

$$B \subset A$$



Information Projection

$$\rightarrow KL(q||p) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \underbrace{\frac{q(\mathbf{x})}{p(\mathbf{x})}}_0$$

- This is infinite if $p(x) = 0$ and $q(x) > 0$.
- Thus we must ensure that if $p(x) = 0$ then $q(x) = 0$.
- Thus the reverse KL is **zero forcing** and q will **under-estimate** the support of p .

Moment Projection

$$KL(p||q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

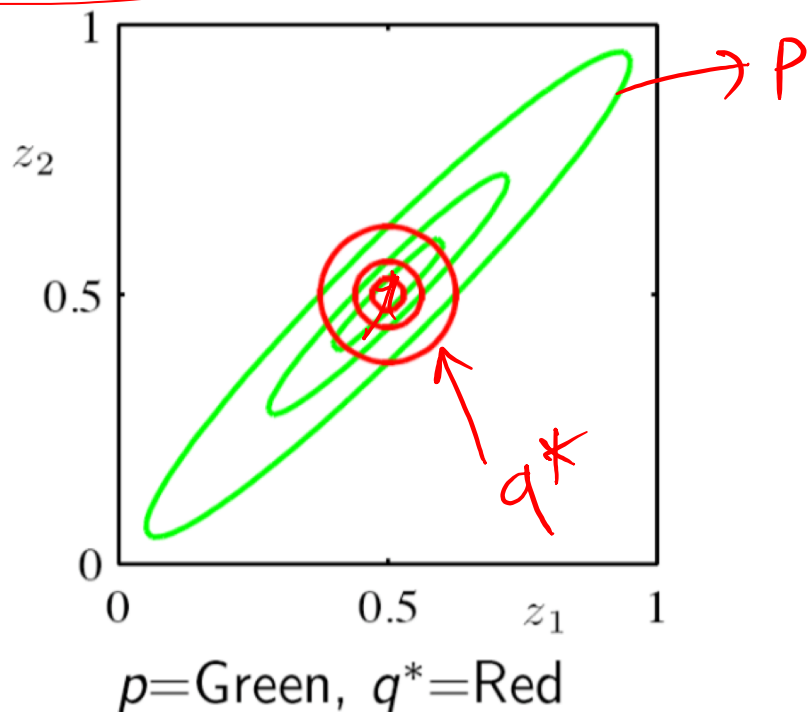
- This is infinite if $q(x) = 0$ and $p(x) > 0$. This is **zero avoiding**, and the forward KL **over-estimates** the support of p .

$$Q = N(0, \sigma^2 I)$$

Example: Single Gaussian

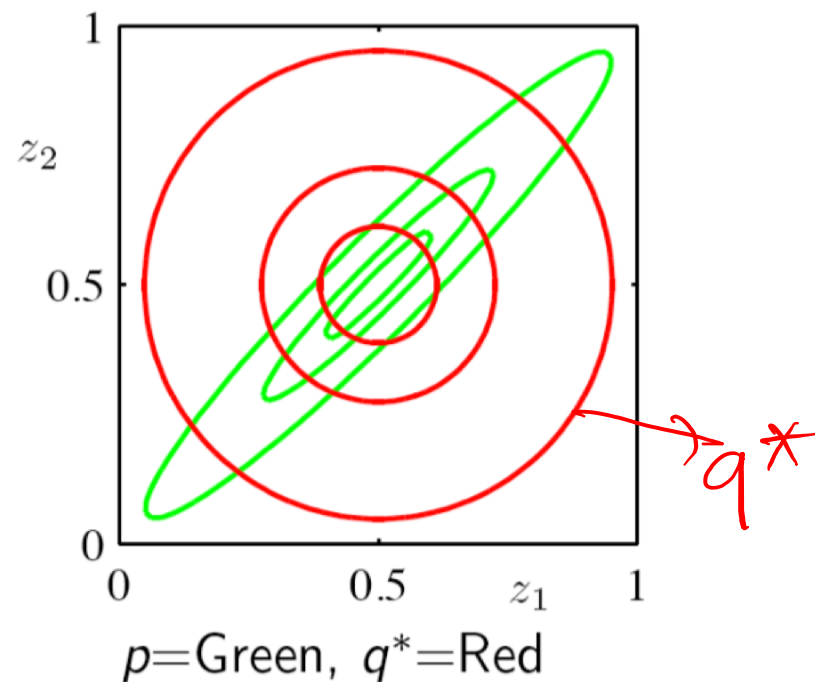
Information Projection

$$KL(q||p) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$



Moment Projection

$$KL(p||q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

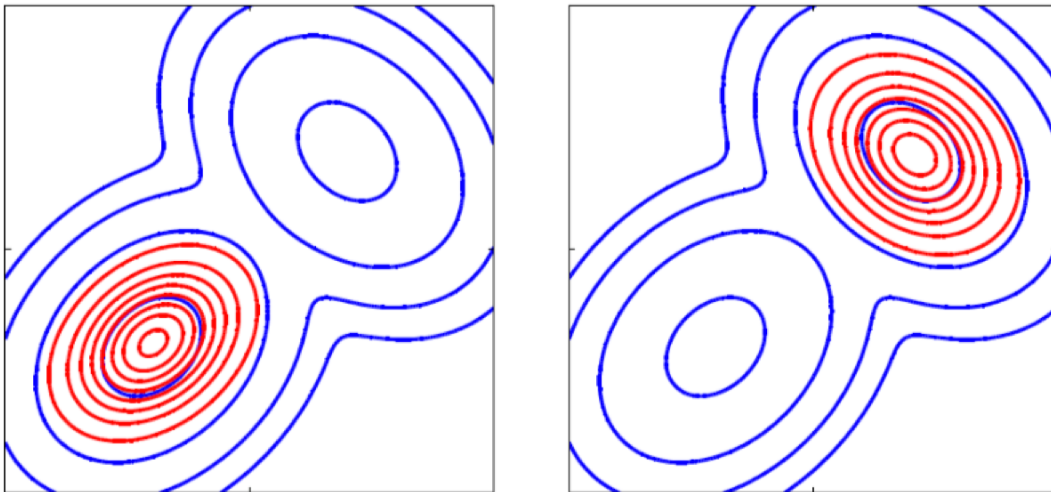


$q \in \mathcal{N}(\mu, \Sigma)$

Example: Mixture of Gaussians

Information Projection

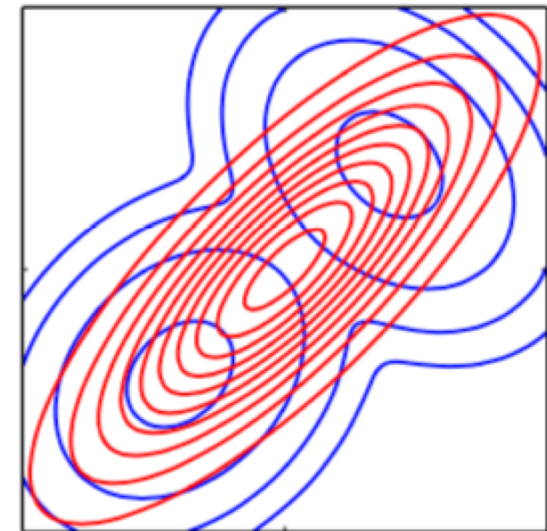
$$KL(q||p) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$



p =Blue, q^* =Red (two local minima!)

Moment Projection

$$KL(p||q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

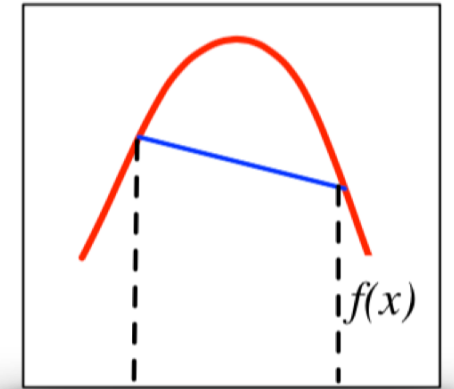


Let's apply this technique in a context

Review: Jensen Inequality

An important result from convex analysis:

$$\text{For concave functions } f(.) \\ f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$$



Logarithms are strictly *concave* allowing us to use Jensen's inequality.

$$\log \int p(x) g(x) dx \geq \int p(x) \log g(x) dx$$

$$\log \mathbb{E}[g(x)] \geq \mathbb{E}[\log g(x)]$$

$$p(z|y) = \frac{p(y|z)p(z)}{p(y)}$$

Let's Take a Look at an Integration

Integral problem

$$\log p(y) = \log \int p(y|z)p(z) dz = \mathbb{E}_{z \sim p(z)} [p(y|z)]$$

$$\log p(y) = \log \int p(y|z)p(z) \frac{q(z)}{q(z)} dz = \mathbb{E}_{z \sim q} [g(\cdot)]$$

Jensen's inequality

$$\log \int p(x)g(x)dx \geq \int p(x) \log g(x)dx$$

$$\begin{aligned} \log p(y) &\geq \int q(z) \log \left(p(y|z) \frac{p(z)}{q(z)} \right) dz \\ &= \int q(z) \log p(y|z) - \int q(z) \log \frac{q(z)}{p(z)} \end{aligned}$$

Variational lower bound

$$= \mathbb{E}_{q(z)} [\log p(y|z)] - KL[q(z) || p(z)]$$

Interpreting the Lower Bound (ELBO)

KL
ELBO

$$\mathcal{F}(y, q) = \mathbb{E}_{q(z)} [\log p(y|z)] - KL [q(z) || p(z)]$$

Data

Approximate
Posterior

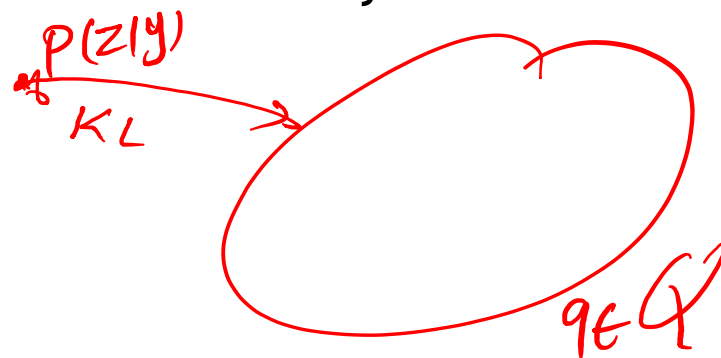
Reconstruction

Penalty

Approximate posterior distribution $q(z)$: Best match to true posterior $p(z|y)$, one of the unknown inferential quantities of interest to us.

Reconstruction Cost: The expected log-likelihood measure how well samples from $q(z)$ are able to explain the data y .

Penalty: Ensures the the explanation of the data $q(z)$ doesn't deviate too far from your beliefs $p(z)$. A mechanism for realising Okham's razor.



$q(z)$

Interpreting the Lower Bound (ELBO) $p(x|e)$

$$\mathcal{F}(y, q) = \mathbb{E}_{q(z)} [\log p(y|z)] - KL [q(z) || p(z)]$$

Some comments on q :

- **Integration is now optimisation:** optimise for $q(z)$ directly.
 - I write $q(z)$ to simplify the notation, but it depends on the data, $q(z|y)$.
 - *Easy convergence assessment* since we wait until the free energy (loss) reaches convergence.
- **Variational parameters:** parameters of $q(z)$
 - E.g., if a Gaussian, variational parameters are mean and variance.
 - Optimisation allows us to *tighten the bound* and get as close as possible to the true marginal likelihood.

$$\mathcal{F}(y, q) = \mathbb{E}_{q(z)} [\log p(y|z)] - KL [q(z) || p(z)]$$



Approximate
Posterior

How to implement it?
What is q exactly?

$p(x) = \tilde{p}(x) / Z$ Free-form and Fixed-form

Free-form: variational method solves for the exact distribution setting the functional derivative to zero.

q Proper $\int q(x) dx = 1$ $q(x) \geq 0$

$$\frac{\delta \mathcal{F}(y, q)}{\delta q(z)} = 0 \quad \text{s.t.} \quad \int q(z) dz = 1$$
$$q(z) \propto p(z) \exp(\log p(y|z, \theta))$$

Great! The optimal solution is the true posterior distribution.

But solving for the normalisation is our original problem.

fixed

~~Free-form~~: variational method specifies an explicit form of the q -distribution.

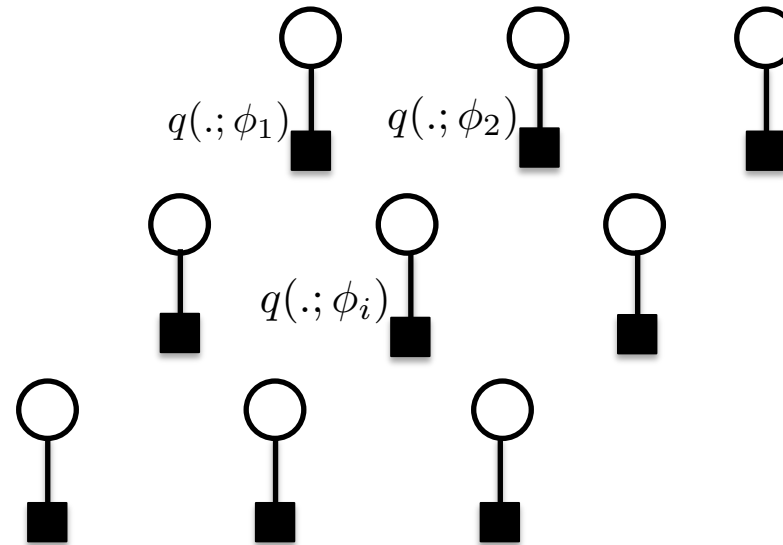
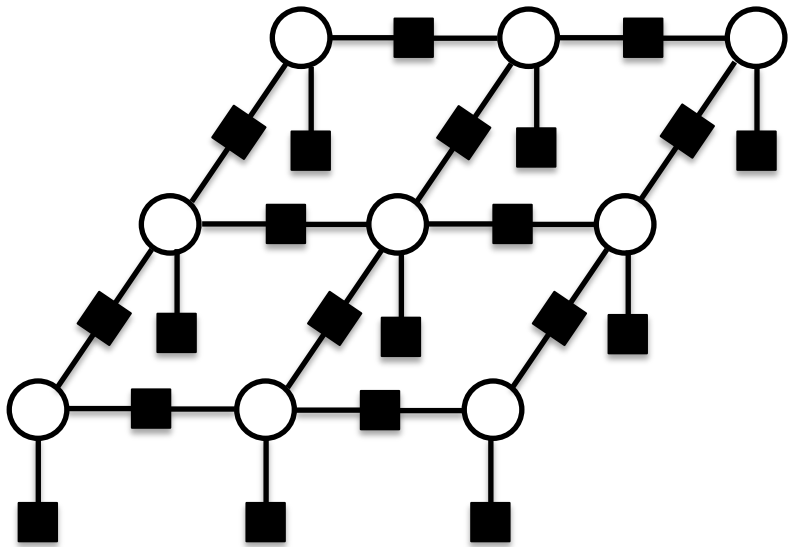
$$q_{\phi}(z) = f(z; \phi)$$

This is ideally a rich class of distributions. Parameters ϕ are called variational parameters.

(Naïve) Mean Field Approach

Very popular approach assuming the posterior is fully factorizable

$$q(\mathbf{x}; \phi) = \prod_i q_i(x_i; \phi_i)$$



Mean Field Approach

Very popular approach assuming the posterior is fully factorizable

$$q(\mathbf{x}; \phi) = \prod_i q_i(x_i; \phi_i) = q_1(x_1; \phi_1) q_2(x_2; \phi_2) \dots$$

Goal: optimizing this cost function over q_i

$$\min_{q_1, \dots, q_D} KL(q || p)$$

Remember that we want to maximize this lower bound:

$$L(q) = -J(q) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})} \leq \log p(\mathcal{D})$$

$$\max_q \mathbb{E}_{\mathbf{x} \sim q} \left[\log \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})} \right]$$

Mean Field Updates

Let's focus on q_j (holding all other terms constant)

$$\begin{aligned}
L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\
&= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\
&= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \left[\sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) - \mathbb{E}_{q_j} [\log \tilde{p}(\mathbf{x})] \right] \\
&\quad \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\sum_{k \neq j} \log q_k(\mathbf{x}_k) + \log q_j(\mathbf{x}_j) \right] \\
&= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}
\end{aligned}$$

$$p(z|D) \propto p(z|D) / \tilde{p}$$

Mean Field Updates

ax
a

Let's focus on q_j (holding all other terms constant)

$$\begin{aligned} L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\ &= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const} \end{aligned}$$

$$L(q_j) = \mathbb{E}_{q_j} \left[\mathbb{E}_{q_{-j}} [\log \tilde{p}(\mathbf{x})] \right] + H(q_j)$$

$$\frac{\delta L(q_j)}{\delta q_j} = 0$$

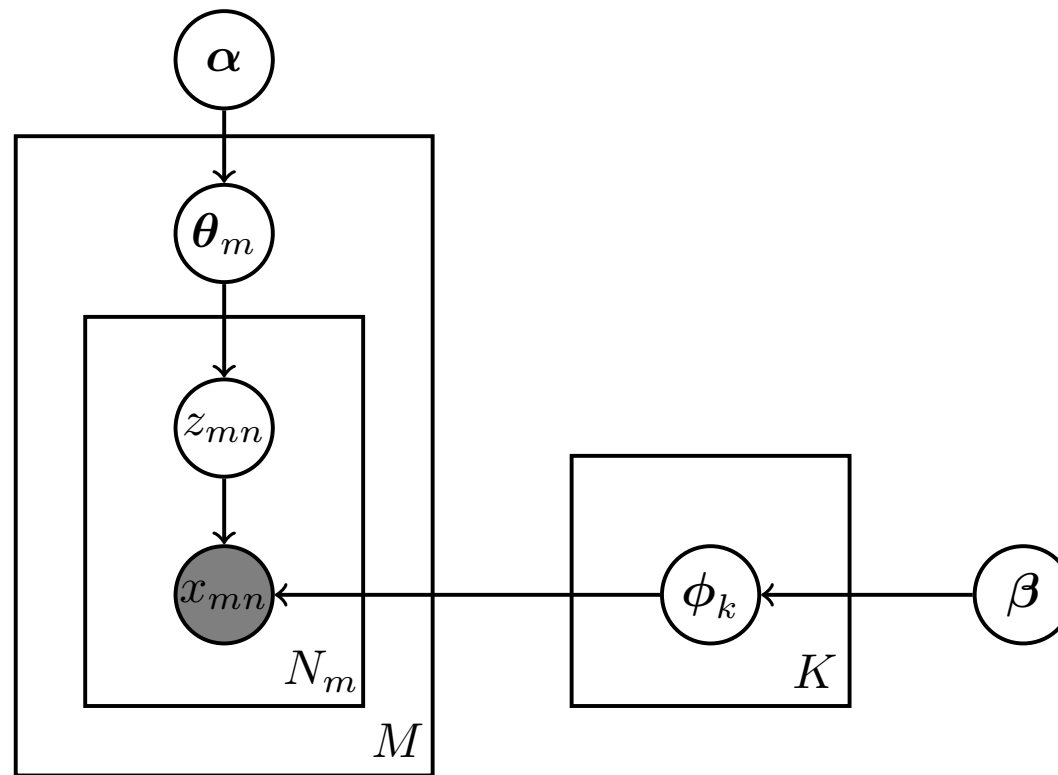
$$\frac{\delta L(q_j)}{\delta q_j} = \mathbb{E}_{q_{-j}} [\log \tilde{p}(\mathbf{x})] - \log q_j - 1 = 0$$

$$q_j^* \propto \exp\{\mathbb{E}_{q_{-j}} [\log \tilde{p}(\mathbf{x})]\}$$

Case study: Latent Dirichlet Allocation

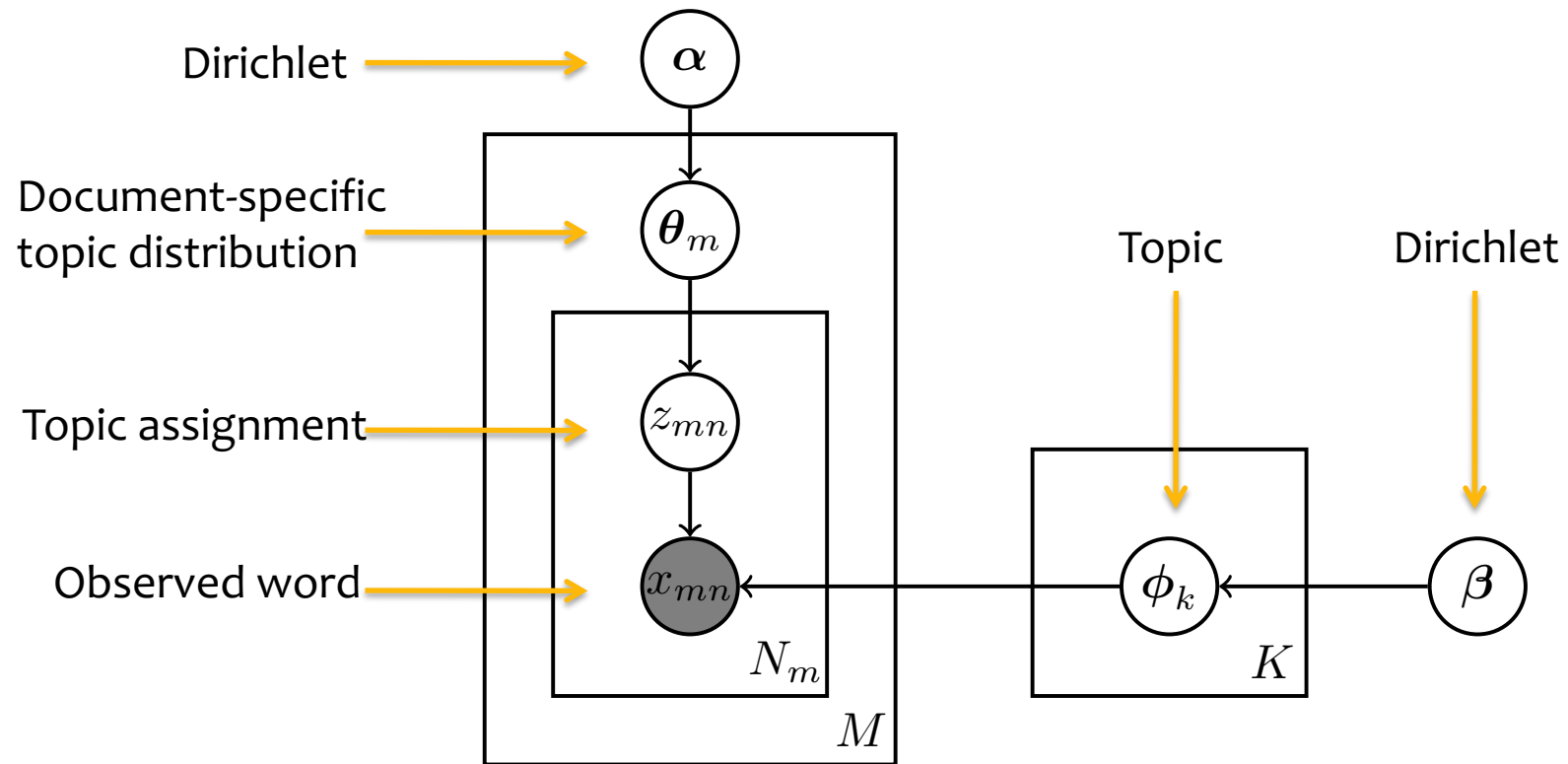
Latent Dirichlet Allocation

- Plate Diagram



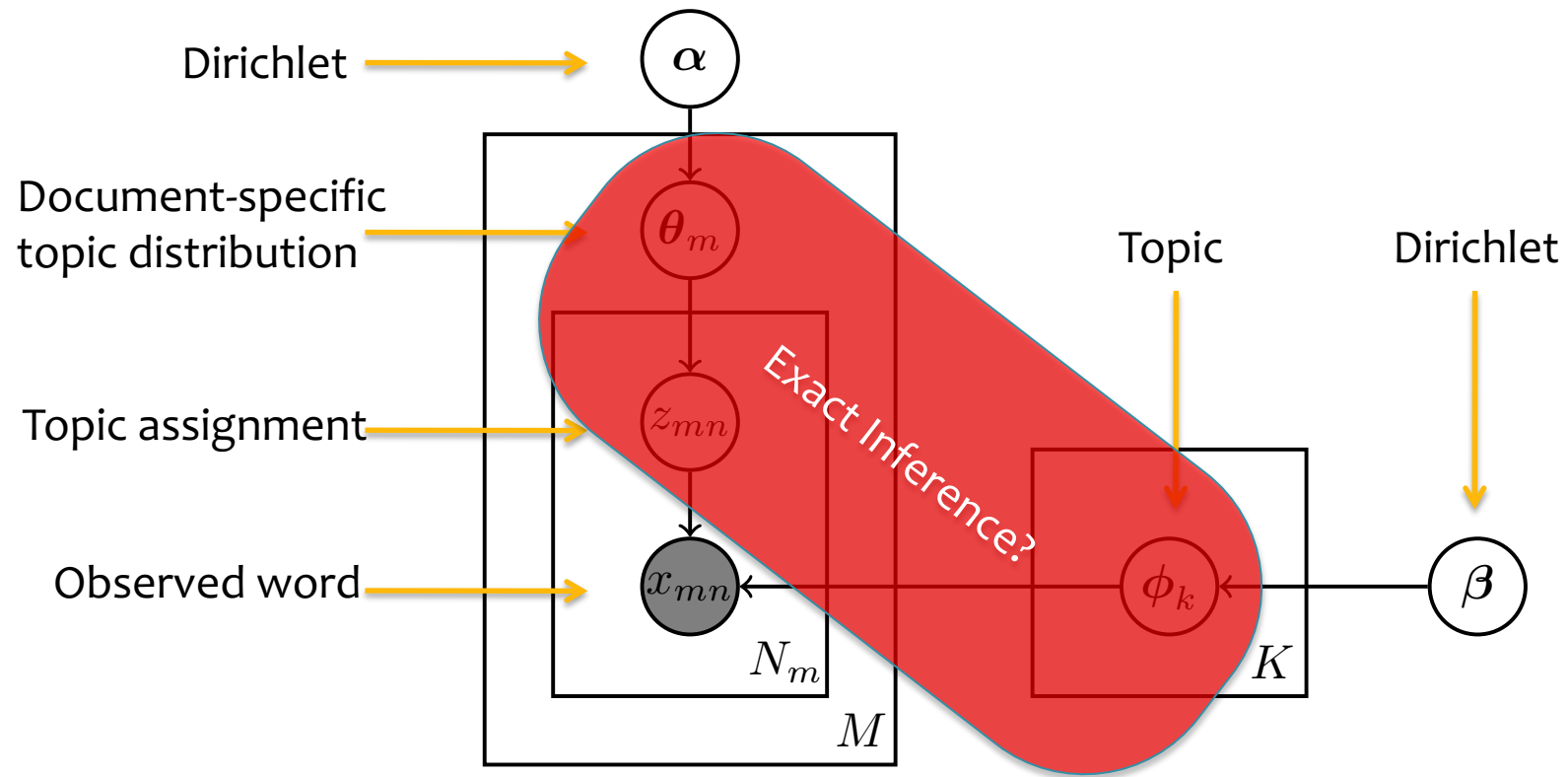
Latent Dirichlet Allocation

- Plate Diagram



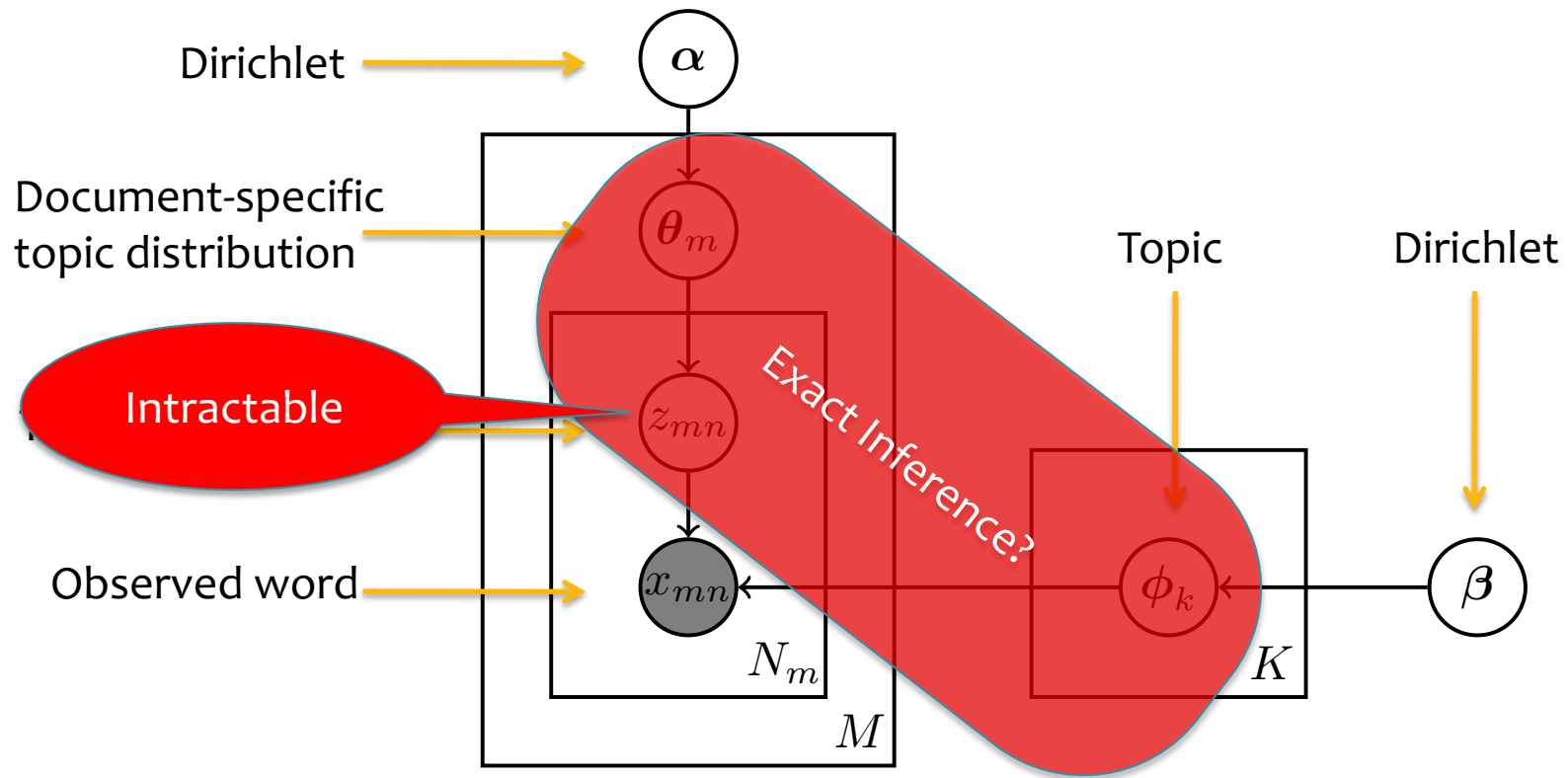
LDA Inference

- Bayesian Approach



LDA Inference

- Bayesian Approach



Inference

- Joint distribution

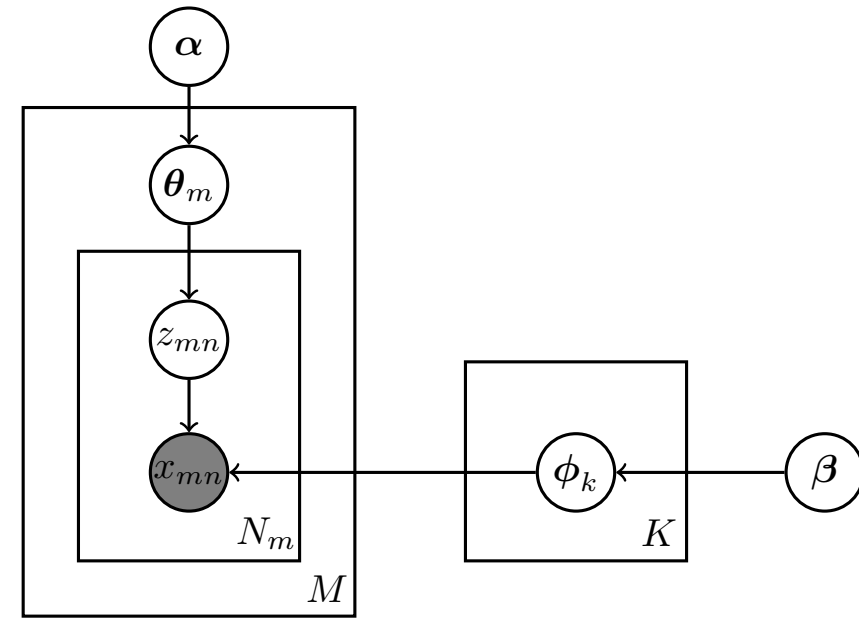
$$p(\cdot) = p(\alpha)p(\beta) \prod_m^M p(\theta_m|\alpha) \prod_{n=1}^{N_m} p(x_{mn}|z_{nm}, \{\phi_k\}_{k=1}^K) p(z_{nm}|\theta_m) \prod_k^K p(\phi_k|\beta)$$

- Latent variables

$$\{\phi_k\}_{k=1}^K, \{z_{nm}\}, \{\theta_m\}$$

- Posterior distribution

$$q(\cdot) = \prod_{k=1}^K p(\phi_k) \prod_{m=1}^M p(\theta_m) \prod_{n=1}^{N_m} p(z_{nm})$$



Let's work out one of the updates....

$$q(\theta_m) \propto \exp \left[\mathbb{E}_{\Pi_n q(z_{nm})} [\log p(\theta_m | \alpha)] + \sum_n \log p(z_{nm} | \theta_m) \right]$$

Remember this:

$$q_j^* \propto \exp \{ \mathbb{E}_{q_{-j}} [\log \tilde{p}(\mathbf{x})] \}$$

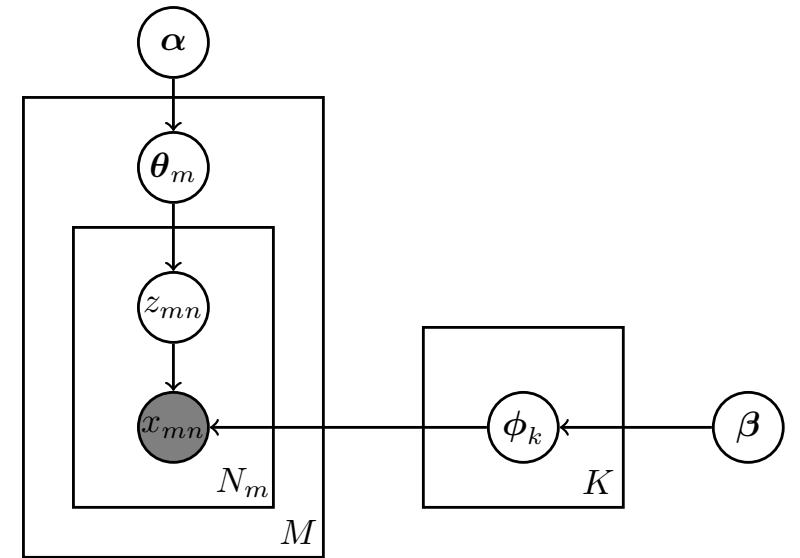
In LDA:

Dirichlet: $p(\theta_m | \alpha) \propto \exp \left[\sum_{k=1}^K (\alpha_k - 1) \log \theta_{mk} \right]$

Categorical: $p(z_{mn} | \theta_m) \propto \exp \left[\sum_{k=1}^K 1(z_{mn} = k) \log \theta_{mk} \right]$

We Obtain:

$$q(\theta_m) \propto \exp \left[\sum_{k=1}^K \left(\sum_{n=1}^N q(z_{mn} = k) + \alpha_k - 1 \right) \log \theta_{mk} \right]$$



Advantages and Disadvantages

Disadvantages:

- An **approximate posterior** only - not always
- **Difficulty in optimisation** — can get stuck in guaranteed to find exact posterior in the limit. local minima.
- Typically **under-estimates the variance** of the posterior and can bias maximum likelihood parameter estimates.

Limited theory and guarantees for variational methods.

Advantages:

- Applicable to almost **all probabilistic models**: non-linear, non-conjugate, high-dimensional, directed and undirected.
- Can be **faster to converge** than competing methods.
- Easy **convergence assessment**.
- **Numerically stable**.
- Can be used on **modern computing architectures** (CPUs and GPUs).
- Principled and scalable approach for **model selection**.

Mean field vs LBP

- LBP minimizes the **Bethe** energy while MF maximizes the **ELBO**.
- LBP is **exact** for trees whereas MF is not, suggesting LBP will in general.
- LBP optimizes over **node** and **edge marginals**, whereas naïve MF only optimizes over **node marginals**, again suggesting LBP will be more accurate.
- MF objective has many more local optima than the LBP objective, so optimizing the MF objective seems to be harder.
- MF tends to be more **overconfident** than BP
- the advantage of MF is that it gives a lower bound on the partition function while for LBP we don't know the relationship.
- MF is **easier** to extend to other distributions besides discrete and Gaussian.

