CS598JHM: Advanced NLP (Spring 2013)
http://courses.engr.illinois.edu/cs598jhm/

Lecture 8: Variational inference

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center
Office hours: by appointment

The EM algorithm again

C. Bishop (2006) Pattern Recognition and Machine Learning

The EM algorithm

Assume a model $p(\mathbf{X}, \mathbf{Z} \mid \theta)$ over $\mathbf{X}$ and $\mathbf{Z}$ in which maximizing the *complete* log-likelihood $\ln p(\mathbf{X}, \mathbf{Z} \mid \theta)$ of a complete data set $(\mathbf{X}, \mathbf{Z})$ is easy.

However, if we are only given an incomplete data set $\mathbf{X}$, maximizing the *marginal* log-likelihood $\ln p(\mathbf{X} \mid \theta)$ is difficult because it contains a summation inside the logarithm:

$$\ln p(\mathbf{X}|\theta) \quad = \quad \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

We instead maximize the *expected* complete log-likelihood $Q(\theta, \theta')$ according to $\mathbf{Z}$'s posterior $p(\mathbf{Z} \mid \mathbf{X}, \theta')$ under the old parameters θ'

$$\begin{aligned} Q(\theta, \theta') &= E_{\mathbf{Z}|\mathbf{X},\theta'}[\ \ln p(\mathbf{X}, \mathbf{Z} \mid \theta)\ ] \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta') \ln p(\mathbf{X}, \mathbf{Z} \mid \theta) \end{aligned}$$

The EM algorithm

1. Initialization: Choose initial parameters θ^{old}

2. Expectation step: Evaluate $p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$

   (posterior of $\mathbf{Z}$ under current parameters θ^{old})

3. Maximization step: Find $\theta^{\text{new}} = \text{argmax}_\theta \, Q(\theta, \theta^{\text{old}})$

   Maximize expected complete log-likelihood under $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$

$$\arg\max_\theta Q(\theta, \theta^{\text{old}}) = \arg\max_\theta E_{\mathbf{Z}|\mathbf{X},\theta^{\text{old}}}[\ \ln p(\mathbf{X}, \mathbf{Z} \mid \theta)\ ]$$

$$= \arg\max_\theta \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta)$$

4. Stop, or set $\theta^{\text{old}} := \theta^{\text{new}}$ and go to 2.

Another view of $\ln p(\mathbf{X}|\, \theta)$

By the product rule: $\ln p(\mathbf{X}, \mathbf{Z}|\, \theta) = \ln p(\mathbf{Z} \mid \mathbf{X}, \theta) + \ln p(\mathbf{X} \mid \theta)$

Define a functional (a function from functions to scalars) $\mathcal{L}(q, \theta)$ of $q(\mathbf{Z})$ that depends on the joint likelihood $p(\mathbf{X}, \mathbf{Z}|\, \theta)$

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

The KL-divergence of $q(\mathbf{Z})$ and $p(\mathbf{Z} \mid \mathbf{X}, \theta)$, the posterior of $\mathbf{Z}$:

$$KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})}$$

Thus for *any* distribution $q(\mathbf{Z})$ over the latent variables $\mathbf{Z}$:

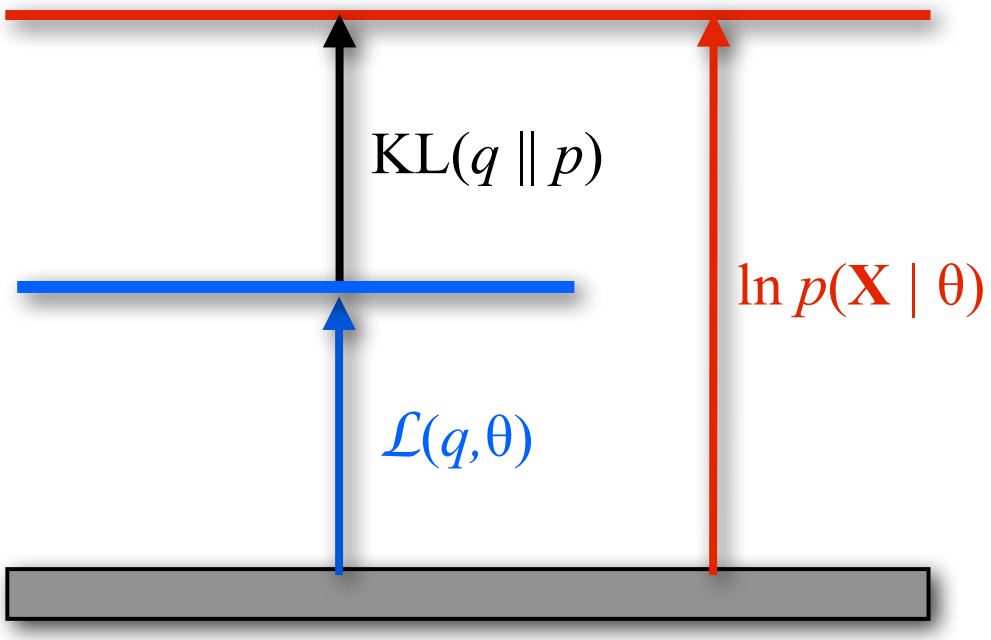
$$\ln p(\mathbf{X} \mid \theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q \,||\, p)$$
$$= \mathcal{L}(q(\mathbf{Z}), p(\mathbf{X}, \mathbf{Z} \mid \theta)) + \mathrm{KL}(q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{X}, \theta))$$

$$\ln p(\mathbf{X} \mid \theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q \,\|\, p)$$

$\mathcal{L}(q,\theta)$ is a lower bound on the marginal likelihood $\ln p(\mathbf{X} \mid \theta)$:

$$\mathcal{L}(q,\theta) = \ln p(\mathbf{X} \mid \theta) - \mathrm{KL}(q \,\|\, p)$$

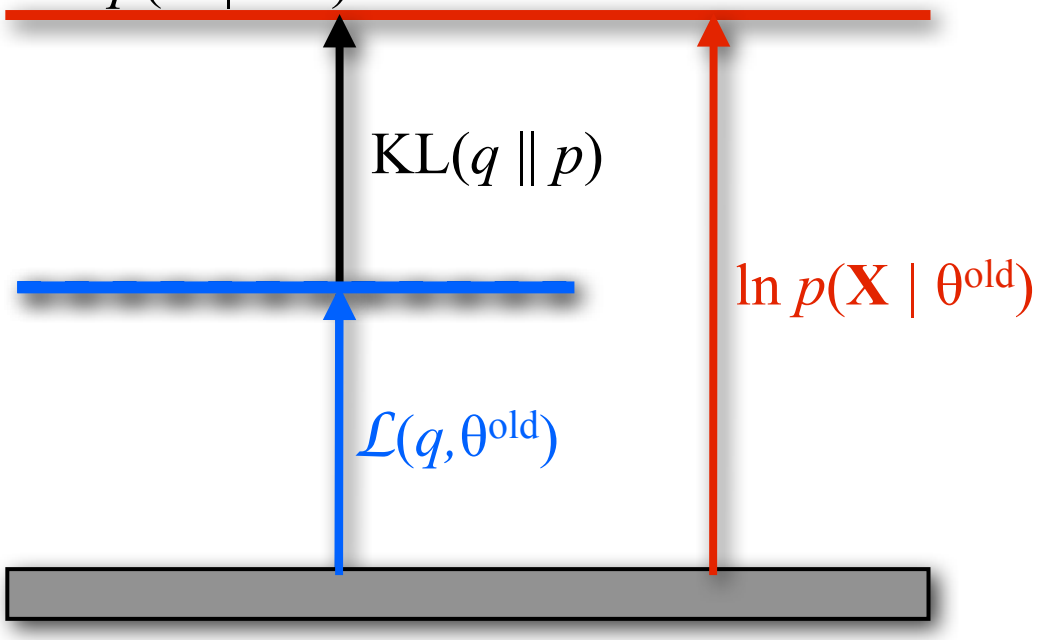
Thus, maximizing $\mathcal{L}(q,\theta)$ will also increase $\ln p(\mathbf{X} \mid \theta)$



E-step: $q^{\text{new}}(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$

$$q^{\text{new}}(\mathbf{Z}) = \text{argmax}_q \, \mathcal{L}(q(\mathbf{Z}), \theta^{\text{old}})$$
$$= \text{argmax}_q \, [\, \ln p(\mathbf{X} \mid \theta^{\text{old}}) - \text{KL}(\, q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})\,)\,]$$
$$= \text{argmin}_q \, \text{KL}(\, q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})\,)$$
$$= p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$$

$$\mathcal{L}(q^{\text{new}}(\mathbf{Z}), \theta^{\text{old}}) = \ln p(\mathbf{X} \mid \theta^{\text{old}})$$



M-step: $\theta^{new} = \text{argmax}_\theta \mathcal{L}(q(\mathbf{Z}), \theta)$

$$\theta^{new} = \text{argmax}_\theta \mathcal{L}(q(\mathbf{Z}), \theta)$$

$$= \text{argmax}_\theta \mathcal{L}(p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}), \theta)$$

$$= \text{argmax}_\theta \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \ln p(\mathbf{Z} \mid \mathbf{X}, \theta^{old})$$

$$= \text{argmax}_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z})$$

$$= \text{argmax}_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} \text{H}(q(\mathbf{Z}))$$

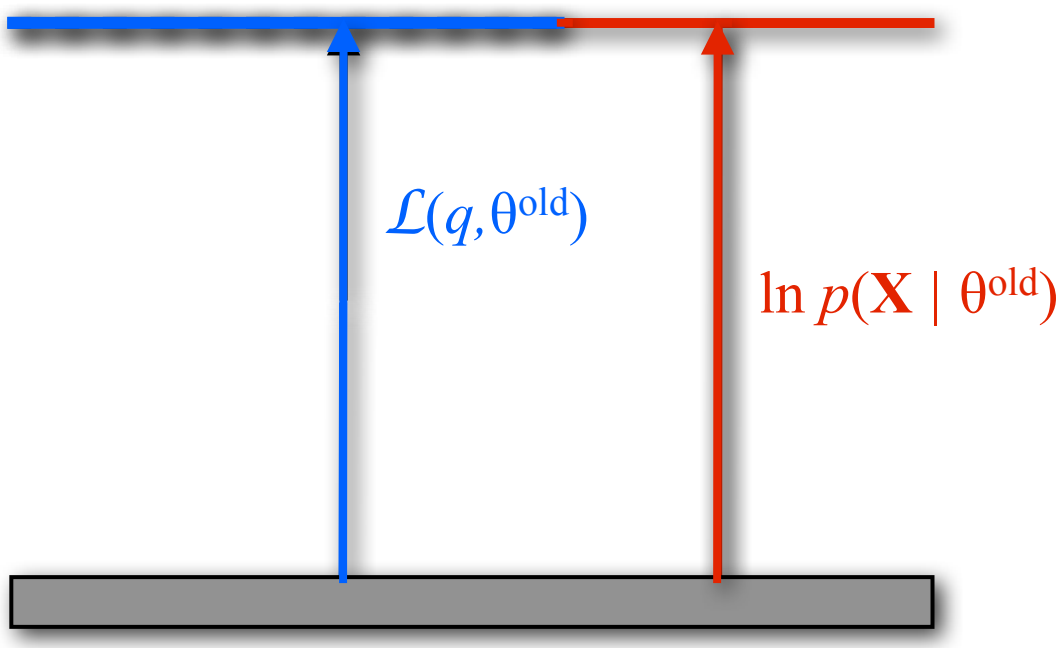
$$= \text{argmax}_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta)$$

$$= \text{argmax}_\theta \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z} \mid \theta)$$

$$= \text{argmax}_\theta \text{Q}(\theta, \theta^{old})$$

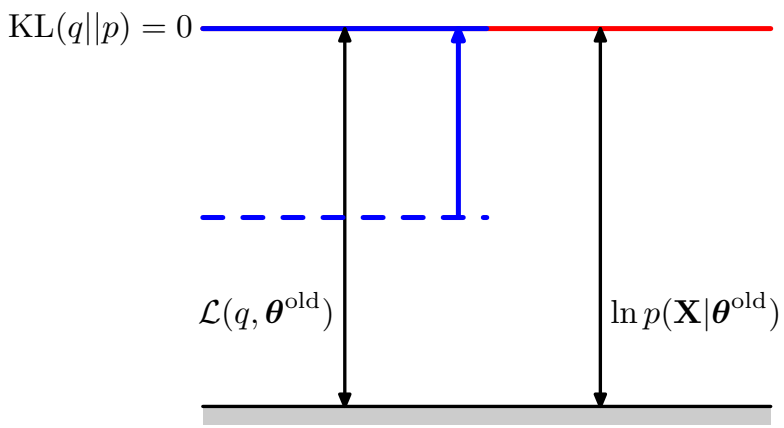
M-step: $\theta^{new} = \text{argmax}_\theta \mathcal{L}(q(\mathbf{Z}), \theta)$

If $\mathcal{L}(q^{new}, \theta^{new}) > \mathcal{L}(q^{new}, \theta^{old})$:
$\text{KL}(q^{new}(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X}, \theta^{new})) > 0$
and $q^{new}(\mathbf{Z}) < p(\mathbf{Z} | \mathbf{X}, \theta^{new})$



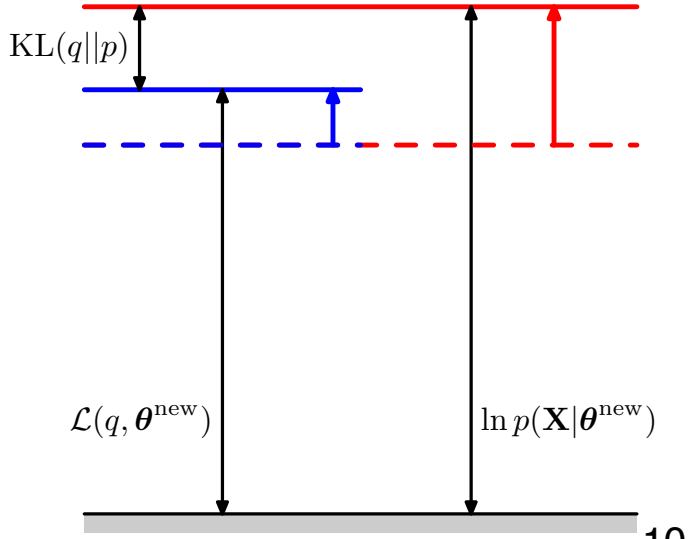
EM again...

$\mathcal{L}(q, \theta)$ is a lower bound on the incomplete log-likelihood $\ln p(\mathbf{X} \mid \theta)$



E-step:
With θ^{old} fixed, return q^{new}
that maximizes $\mathcal{L}(q, \theta^{old})$ wrt. $q(\mathbf{Z})$,
Now $\mathrm{KL}(q^{new} \,||\, p^{old}) = 0$.

M-step:
With q^{new} fixed, return θ^{new}
that maximizes $\mathcal{L}(q^{new}, \theta)$ wrt. θ.
If $\mathcal{L}(\, q^{new}, \theta^{new}) > \mathcal{L}(\, q^{new}, \theta^{old})$:
$\ln p(\, \mathbf{X} \mid \theta^{new}) > \ln p(\, \mathbf{X} \mid \theta^{old})$,
and hence $\mathrm{KL}(\, q^{new} \,||\, p^{new}) > 0$



Variational inference

Variational inference is applicable when you have to compute an *intractable* posterior over latent variables $p(\mathbf{Z}|\mathbf{X})$

Basic idea:
Replace the exact, but intractable posterior $p(\mathbf{Z}|\mathbf{X})$
with a ***tractable*** **approximate posterior** $q(\mathbf{Z}|\mathbf{X}, \mathbf{V})$

$q(\mathbf{Z}|\mathbf{X}, \mathbf{V})$ is from a family of simpler distributions over the latent variables $\mathbf{Z}$ that is defined by a set of **free variational parameters** $\mathbf{V}$

Unlike in EM, $\mathrm{KL}(q \,||\, p) > 0$ for any q, since q only approximates p

Variational EM

Initialization:
Define initial model θ^{old} and variational distribution $q(\mathbf{Z} \mid \mathbf{X}, \mathbf{V})$

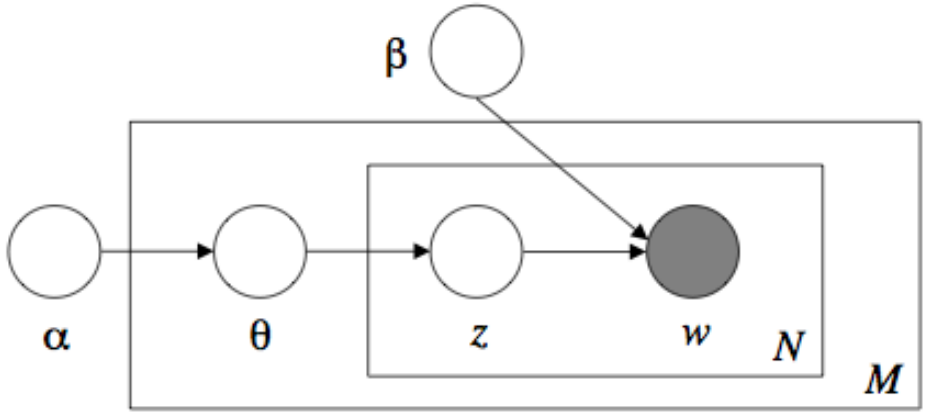
E-step:
Find $\mathbf{V}$ that maximize the variational distribution $q(\mathbf{Z} \mid \mathbf{X}, \mathbf{V})$
Compute the expectation of true posterior $p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$
under the new variational distribution $q(\mathbf{Z} \mid \mathbf{X}, \mathbf{V})$

M-step:
Find model parameters θ^{new} that maximize the expectation of the $p(\mathbf{Z}, \mathbf{X} \mid \theta)$ under the variational posterior $q(\mathbf{Z} \mid \mathbf{X}, \mathbf{V})$

Set $\theta^{\text{old}} := \theta^{\text{new}}$

Variational inference for LDA

Blei, Ng, Jordan (2003)'s LDA



$\boldsymbol{\alpha}$: k-dimensional Dirichlet prior over topic distributions

$\boldsymbol{\theta}_d$: k-dimensional multinomial over topics

β_k: V-dimensional multinomial over words for topic k

β_{ij}: probability of word j in topic i

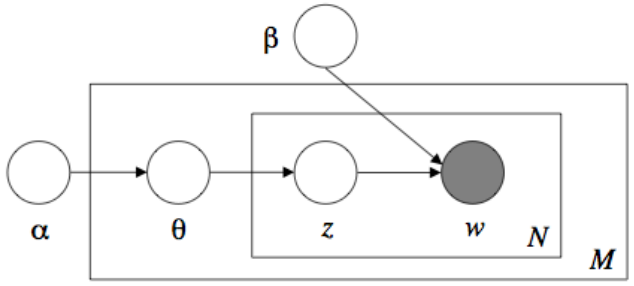
$z_{d,n}$: topic label for word n in document d

(indicator variable: k-dimensional unit vector)

$w_{d,n}$: word n in document d

(indicator variable: V-dimensional unit vector)

Posterior inference for LDA



Given:

- data (M documents $\mathbf{w}_1, \ldots, \mathbf{w}_M$)
- Dirichlet priors $\boldsymbol{\alpha}$ and word multinomials $\boldsymbol{\beta}$

Compute the posterior of the hidden variables:

- document-specific topic multinomials $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$
- word-specific topic labels $z_{1,1}, \ldots, z_{M,N}$

$$p(\boldsymbol{\theta}_{1:D}\ \mathbf{z}_{1,1:M,N}, \mid \mathbf{w}_{1:M},\ \boldsymbol{\alpha},\ \boldsymbol{\beta}_{1:k})$$

The posterior is intractable

$$p(\,\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = p(\,\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \,/\, p(\,\mathbf{w} \mid \alpha, \beta)$$

$p(\,\mathbf{w} \mid \alpha, \beta)$ requires marginalizing over the hidden variables $\boldsymbol{\theta}$, $\mathbf{z}$:

$$p(\,\mathbf{w} \mid \alpha, \beta) = \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\,\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{w} \mid \alpha, \beta\,)\, d\boldsymbol{\theta}\, d\mathbf{z}$$

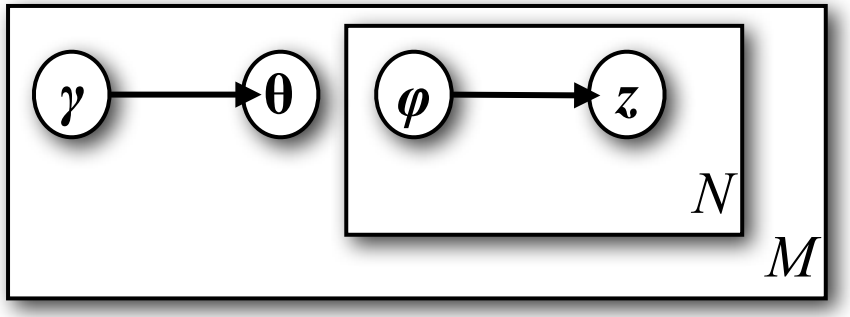
This is intractable, because $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are coupled when we sum over the latent topics.

$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{K} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

$$= \prod_n \sum_i p(w_n \mid \beta_i)\, p(\beta_i \mid \theta_i)$$

An approximate LDA model

$$q(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\varphi}) = q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \prod_n q(z_n \mid \varphi_n)$$



This introduces two sets of variational parameters:

- $\boldsymbol{\gamma}$: document-specific topic Dirichlets
- $\boldsymbol{\varphi}$: word-specific topic multinomials

Probability of topic z given document d: $q(\theta_d \mid \gamma_d)$

Each document has its own Dirichlet prior over topics γ_d

Probability of topic assignment to word $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$

Each word $\text{word}[d][n]$ has its own multinomial over topics $\varphi_{d,n}$

Variational EM

Rewrite the marginal log-likelihood $\log p(\mathbf{w} \mid \alpha, \beta)$ as

$$\log p(\mathbf{w} \mid \alpha, \beta) = \mathcal{L}(\gamma, \varphi; \alpha, \beta) + \mathrm{KL}(q(\theta, \mathbf{z} \mid \gamma, \varphi) \,\|\, p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \alpha, \beta))$$

with

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\varphi}; \alpha, \beta) &= \mathrm{E}_q[\log p(\boldsymbol{\theta} \mid \alpha)] + \mathrm{E}_q[\log p(\mathbf{z} \mid \boldsymbol{\theta})] + \mathrm{E}_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] \\ &\quad - \mathrm{E}_q[\log q(\boldsymbol{\theta} \mid \boldsymbol{\gamma})] - \mathrm{E}_q[\log q(\mathbf{z} \mid \boldsymbol{\varphi})] \\ &= \mathrm{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - \mathrm{E}_q[\log q(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\varphi})]\end{aligned}$$

Recall standard EM

$$\begin{aligned}\mathcal{L}(p(\mathbf{Z} \mid \mathbf{X}, \theta^{\mathrm{old}}), \theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{\mathrm{old}}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) \\ &\quad - \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{\mathrm{old}}) \log p(\mathbf{Z} \mid \mathbf{X}, \theta^{\mathrm{old}}) \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z}) \\ &= \mathrm{E}_q[\log p(\mathbf{X}, \mathbf{Z} \mid \theta)] - \mathrm{E}_q[\log q(\mathbf{Z})]\end{aligned}$$

Variational EM for LDA

Goal: Find α, β which maximize the marginal log-likelihood

$$l(\alpha, \beta) = \sum_d \log p(\boldsymbol{w}_d \mid \alpha, \beta)$$

E-step: Compute approximate variational posterior $q(\theta, \mathbf{z} \mid \gamma^*, \varphi^*)$

Find optimal (data-specific) variational parameters $(\gamma_d^*, \varphi_d^*)$

$$(\gamma^*, \varphi^*) = \arg\min_{(\gamma^*, \varphi^*)} \mathrm{KL}\left(q(\theta, \mathbf{z} \mid \gamma, \varphi) \,\|\, p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) \right)$$

to compute $q(\theta, \mathbf{z} \mid \gamma^*, \varphi^*)$

We will do this in two stages: first update $\boldsymbol{\varphi}^*$, then $\boldsymbol{\gamma}^*$

M-step: Update α, β to maximize lower bound on $l(\alpha, \beta)$

(Find MLE estimates of α, β, using the expected sufficient statistics for each document under the approximate posterior q):

Updates: probability of word j in topic i: $\beta_{ij} \propto \sum_d \sum_n \varphi_{dni} w_{dn}^j$

($\beta_{ij} \propto$ sum of φ_{dni} for all words w_{dn} that are equal to word j)

α: computed numerically

Inside the E-step (1): compute φ_d*

φ are word-specific multinomials over topics:

For each document d: $\varphi_d^* = (\varphi_1, ..., \varphi_N)$

with each $\varphi_n = (\varphi_{n1}, ..., \varphi_{nK})$ a multinomial over topics for w_n

$\varphi_{ni}^* = q(z_n = i \mid w_n)$, the variational posterior of topic i for word w_n:

$$\varphi_{ni}^* \propto \beta_{iw_n} \exp(E_q[\log(\theta_i) \mid \gamma])$$

$$= \beta_{iw_n} \exp(\Psi(\gamma_i) - \Psi(\sum_j \gamma_j))$$

Ψ (digamma): first derivative of Γ

This update is like using Bayes rule

$$p(z_n = i \mid w_i) \propto p(w_n \mid z_n)\, p(z_n = i)$$

$$= \beta_{iwn}\ \theta_i$$

except $p(z_n = i)$ is approximated by $\exp(E_{q(\theta|\gamma)}[\log p(z_n = i)])$, which, without the $E_{q(\theta|\gamma)}[...]$, would just be $\exp(\log p(z_n = i)) = p(z_n = i)$

Inside E-step (2): compute γ^*

$\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_M)$ with $\boldsymbol{\gamma}_d = (\gamma_{d1}, \ldots, \gamma_{dK})$ document-specific posterior Dirichlet over the topics

γ_i: posterior probability of θ_i, the probability of topic i in $d = w_1 \ldots w_N$, given the updated variational parameters $\varphi_1^* \ldots \varphi_n^*$
(d index omitted):

$$\gamma_i = \alpha_i + \sum_n \varphi_{ni} = q(\theta_i \mid \varphi_1^* \ldots \varphi_n^*, \boldsymbol{\alpha})$$

Since φ_n is a multinomial, $\varphi_{ni} = E[z_n = i \mid \varphi_{ni}]$ is the expected frequency i of topic for word w_n under the variational distribution

Relation between true and variational parameters:

True posterior = Dir(hyperparameter + observed frequencies)

Variational posterior = Dir(hyperparameter + *expectation* of observed frequencies)

Summary: Variational EM

Define a tractable model q with variational parameters γ, φ that defines a lower bound on the log-likelihood of the actual model

(E) Update γ,φ to get new approx. posterior of latent vars $q(\theta, \mathbf{z} \mid \gamma, \varphi)$

For each document, find optimal values of its variational parameters γ^* (Dirichlet prior for document-specific topic multinomial) and φ^* (multinomials for topic assignment of each word)

Variational inference inside E-step:

Find γ^*, φ^* that minimize the KL-divergence of q to the actual model p:

$$(\gamma^*, \varphi^*) = \arg\min_{(\gamma^*, \varphi^*)} D(\, q(\theta, \mathbf{z} \mid \gamma, \varphi) \;\|\; p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)\,)$$

(M) Update α, β to get new $p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)$

= increase the lower bound on log-likelihood

Use expected sufficient statistics (=counts) under the new approximate posterior $q(\theta, \mathbf{z} \mid \gamma^*, \varphi^*)$ to compute 'MLE' estimates of α, β (this maximizes the lower bound on log-likelihood, or the *approximate expected* complete log-likelihood)