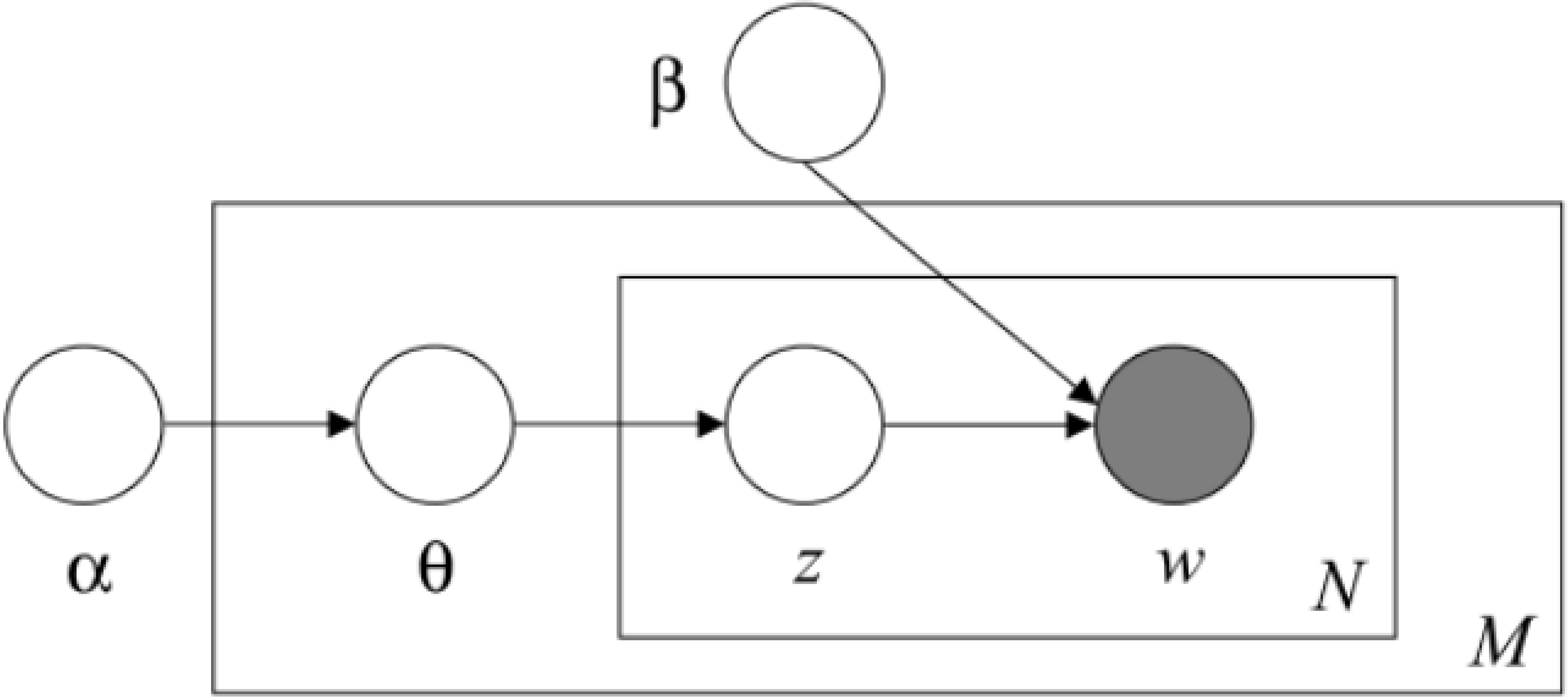
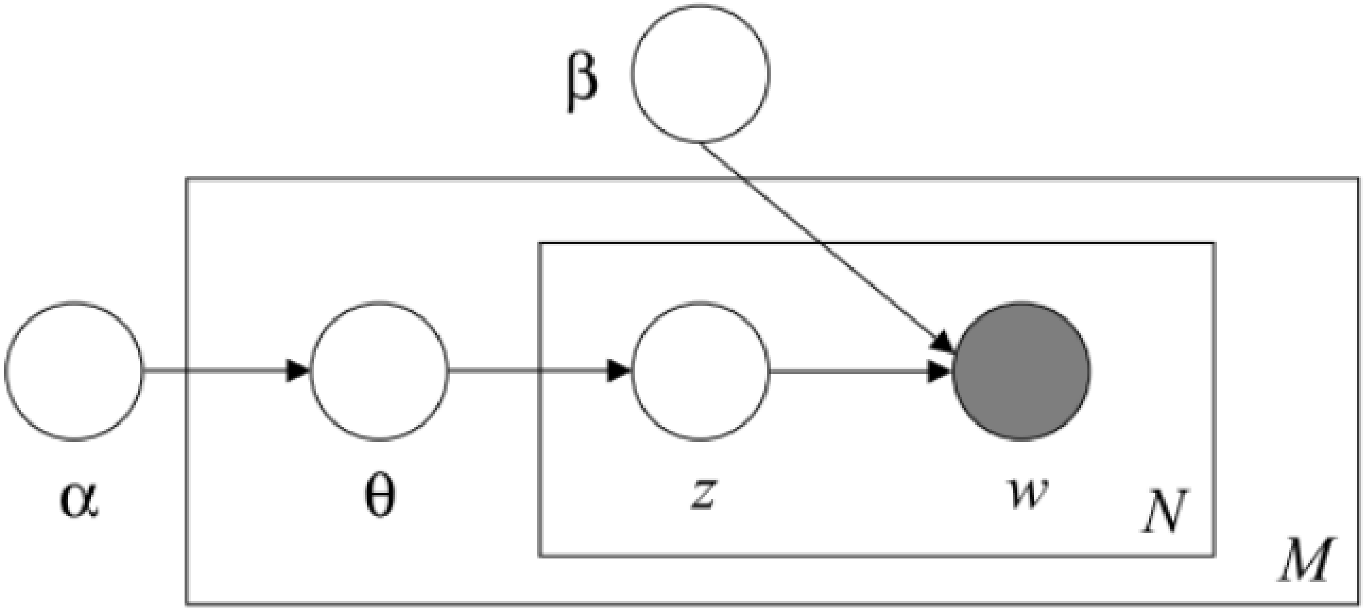
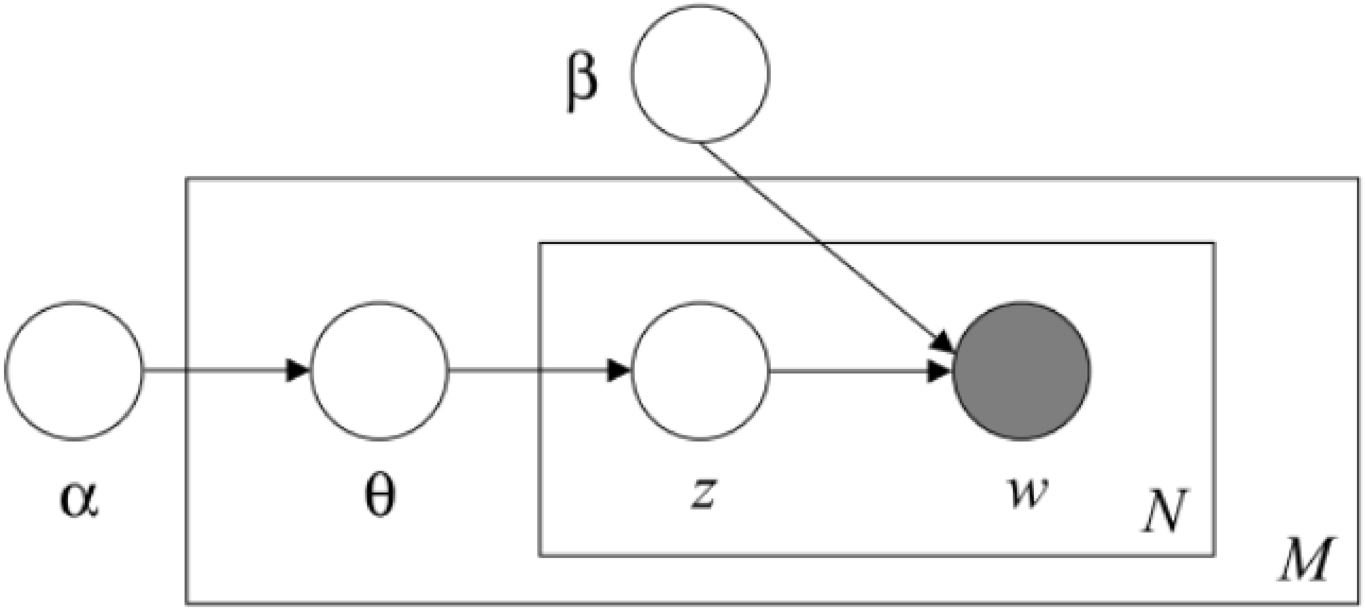
Tutorial 9 Latent Dirichlet Distribution
Joonatan Mänttäri

‘Documents’ are modeled as ‘words’ chosen from a multinomial distribution of ‘words’, w, given a ‘topic’, z, in turn given by a multinomial distribution





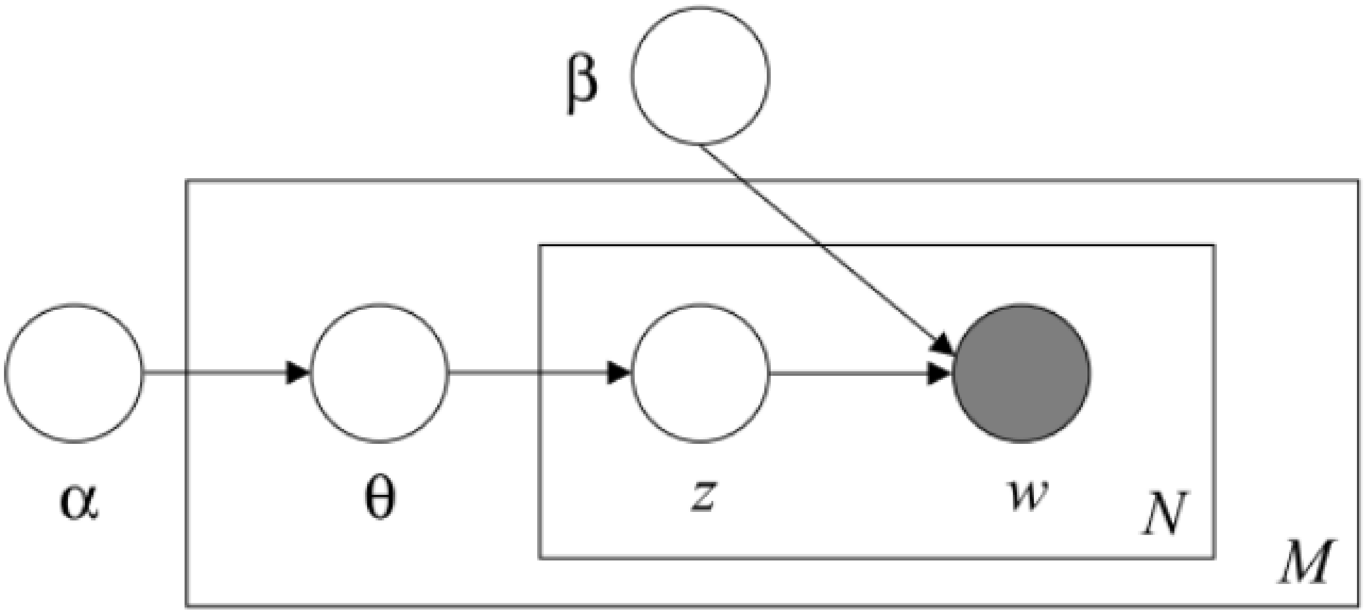
- Choose N ~Poisson(ξ)
- Choose θ ~Dir(α)
- For each of the N words w_n:
- Choose a topic z_n ~Multinomial(θ).
- Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n.



$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1},$$

$$p(\theta,\mathbf{z},\mathbf{w}\,|\,\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n\,|\,\theta)p(w_n\,|\,z_n,\beta),$$

$$p(\mathbf{w}\,|\,\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n\,|\,\theta)p(w_n\,|\,z_n,\beta)\right)d\theta.$$

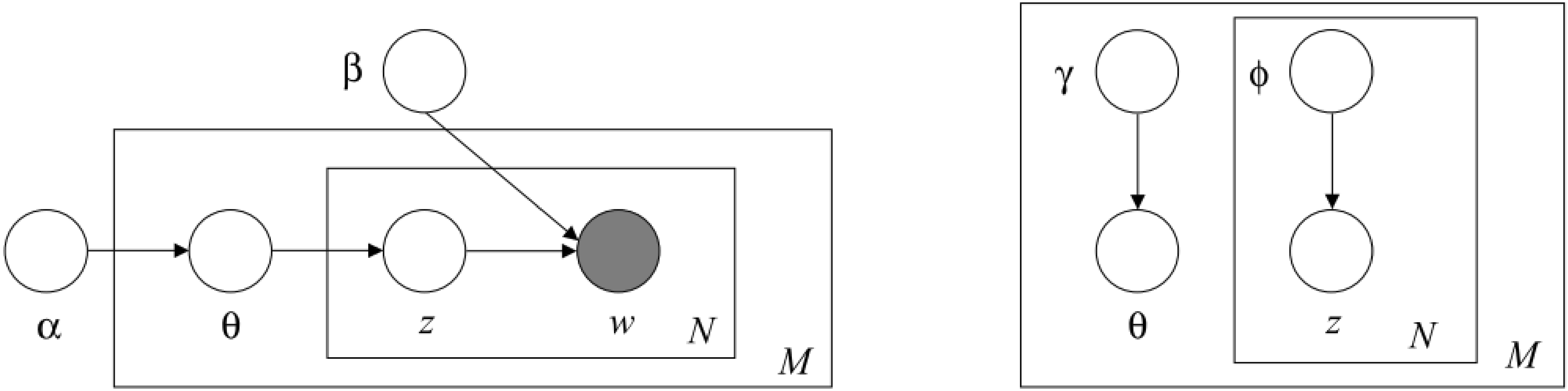


$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

w is a document, ie. a vector of words.
This inference would allow a characterization of that document in terms of topics, **z**.

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta.$$

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

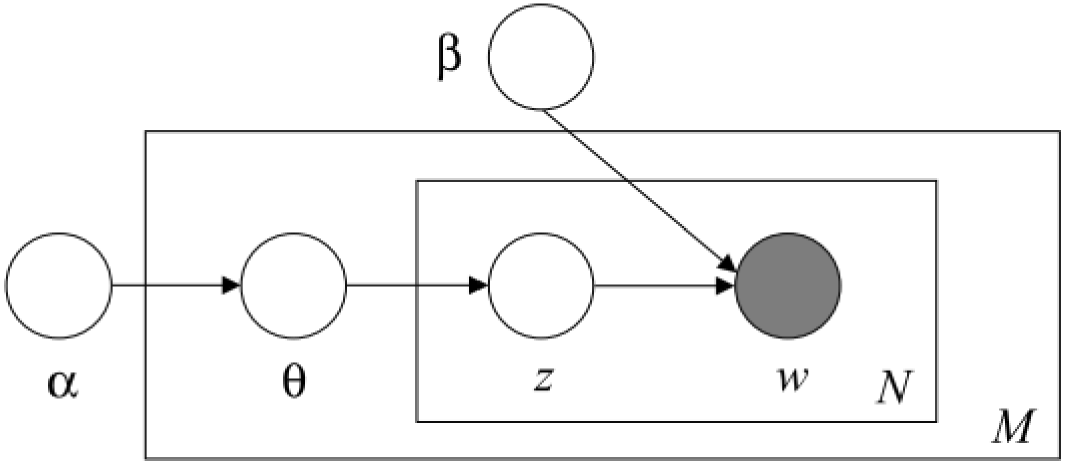


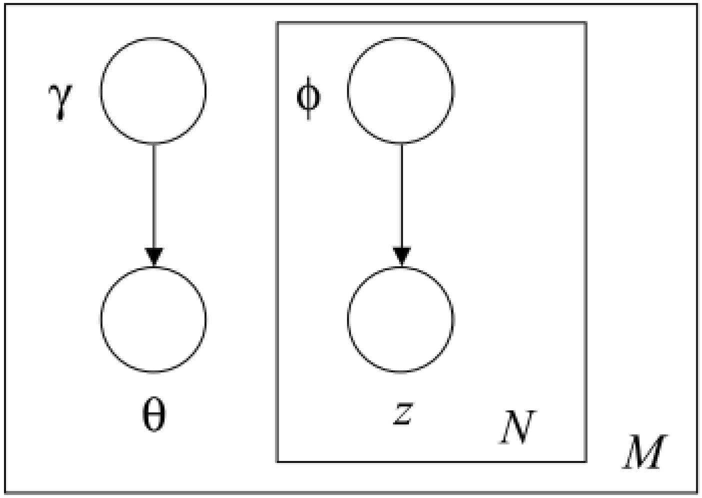
(Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\
&\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta - \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\
&= \mathrm{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \mathbf{z})].
\end{aligned}$$

Jenson’s Inequality

$\mathcal{L}(\gamma, \phi; \alpha, \beta)$ This is our ‘Lower Bound’ and it depends implicitly on some data set, a document of words **w.**





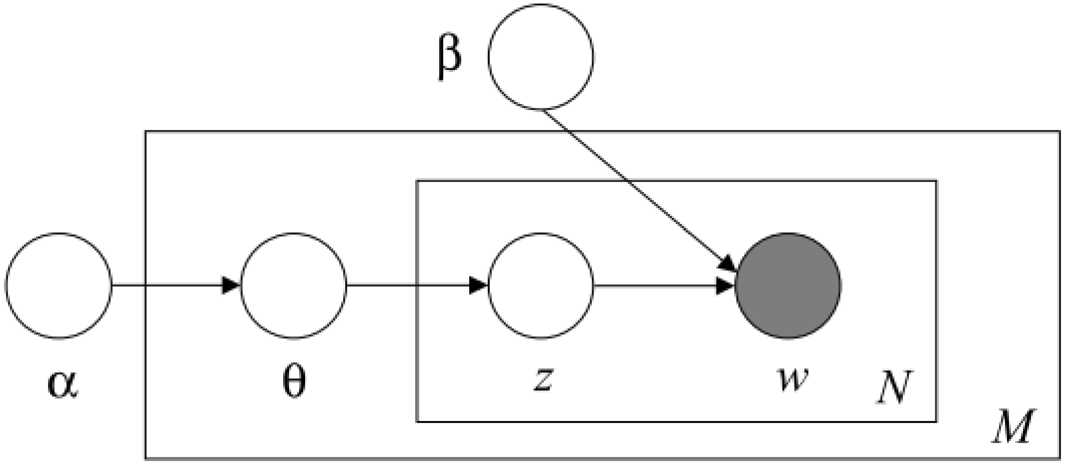
$$\log p(\mathbf{w} \mid \alpha, \beta) = \mathcal{L}(\gamma, \phi; \alpha, \beta) + \mathbf{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).$$ [1]

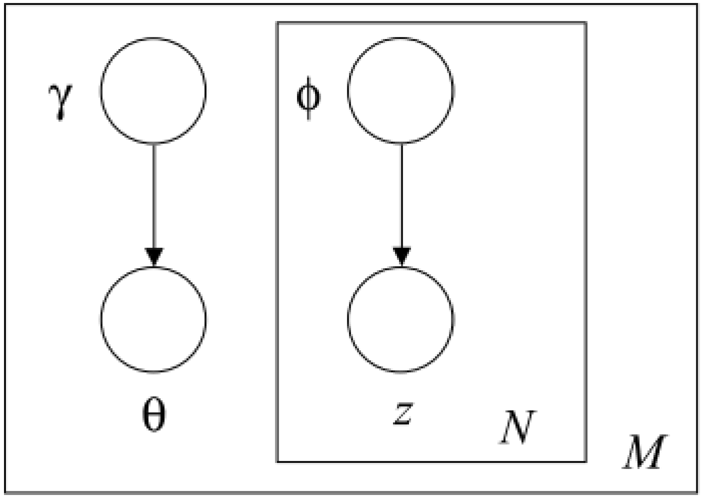
So maximizing the lower bound with respect to our q parameters will make q match p best. This is actually called the (variational) E step. It looks at a given document and fits it topic and word distributions independently using q. It is done per document, M times.

We also want to find p parameters that maximize p for our document. Which we do by maxing the lower bound. This is the **M** step. Done over whole corpus once.

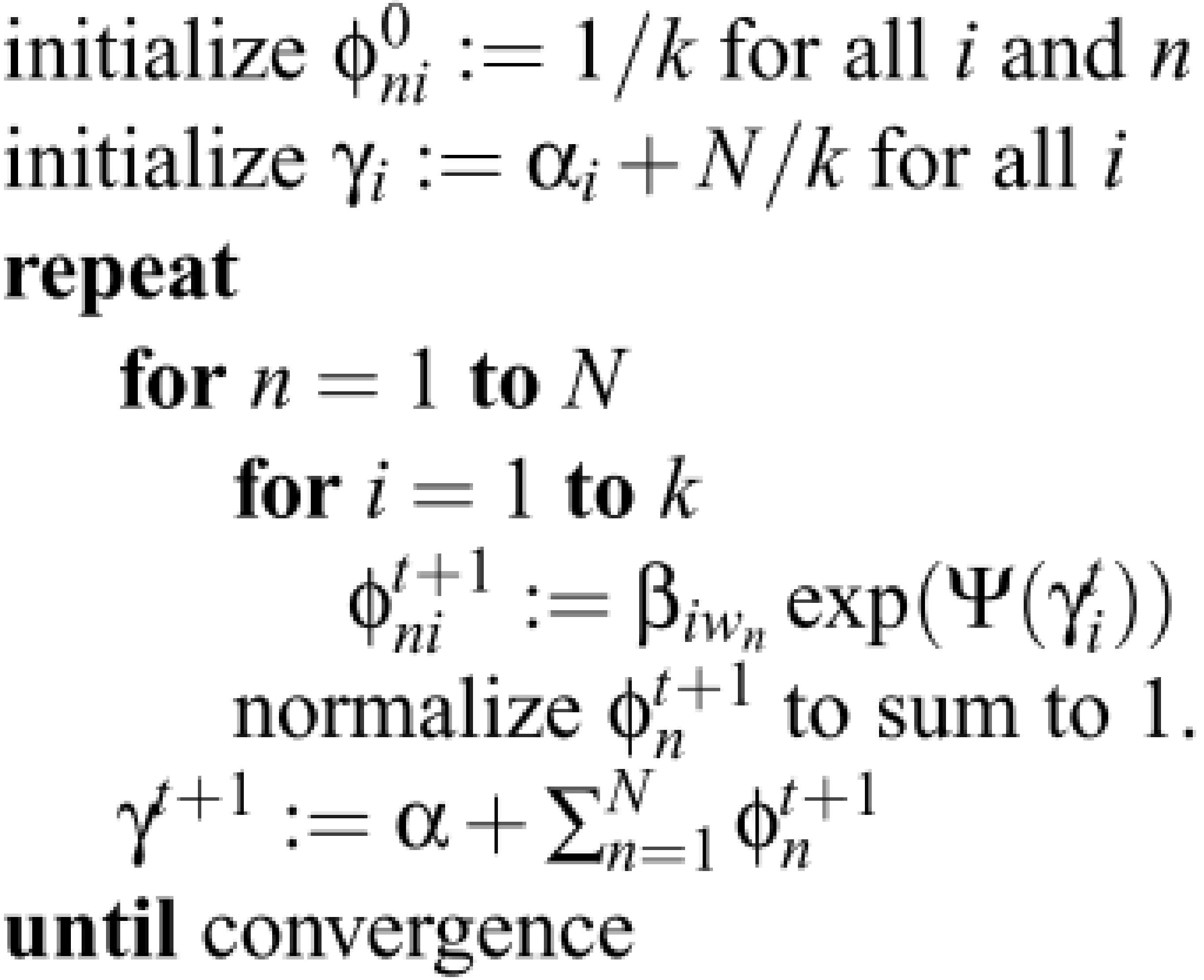
$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}$$

This part is understandable as choosing the word distribution for the topics as the sum over documents of the topic's ratio in a document times the number of times the word is in the document. Alpha will need Newton-Raphson to compute.





The E-Step:



```
initialize $\phi_{ni}^0 := 1/k$ for all $i$ and $n$
initialize $\gamma_i := \alpha_i + N/k$ for all $i$
repeat
    for $n = 1$ to $N$
        for $i = 1$ to $k$
            $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
        normalize $\phi_n^{t+1}$ to sum to 1.
    $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi_n^{t+1}$
until convergence
```

This part is understandable as choosing the 'document's' topic distribution by assigning the topics to the words in the document according to the current 'corpus' model along with the document's model of topic distribution.