Topic models generally refer to unsupervised methods aiming to detect groups of words (topics) that best characterise a document. In this lecture, we are interested in studying LDA. the most representative topic model that uses VI in optimising the parameters.

## Latent Dirichlet Allocation (LDA)

In LDA, we would like to model the generating process for each document $w$ in a corpus of $M$ documents $D = \{w\}_{n=1}^{M}$ as follows,

1. Choose the number of word in $w$: $N \sim \text{Poisson}(\epsilon)$
2. Choose the probabilities of topics in $w$: $\theta \sim \text{Dir}(\alpha)$ . Note that $\sum_{i=1}^{K} \theta_i = 1$
3. For each word in $w$:
    - Choose a topic $z_n \sim \text{Mult}(z_n|\theta)$. Note that $p(z_n = i|\theta) = \theta_i$
    - Choose a word $w_n \sim \text{Mult}(w_n|z_n, \beta)$ - a multinomial distribution conditioned on the topic. Hence, there are parameters $\{\beta_{ij}\}_{i=1,j=1}^{K,V}$ where $K$ and $V$ are the number of topics and vocabulary in the corpus. Note that $p(w_n = j|z_n = i, \beta) = \beta_{ij}$ and $\sum_{j=1}^{V} \beta_{ij} = 1$

The graphical model of LDA looks like

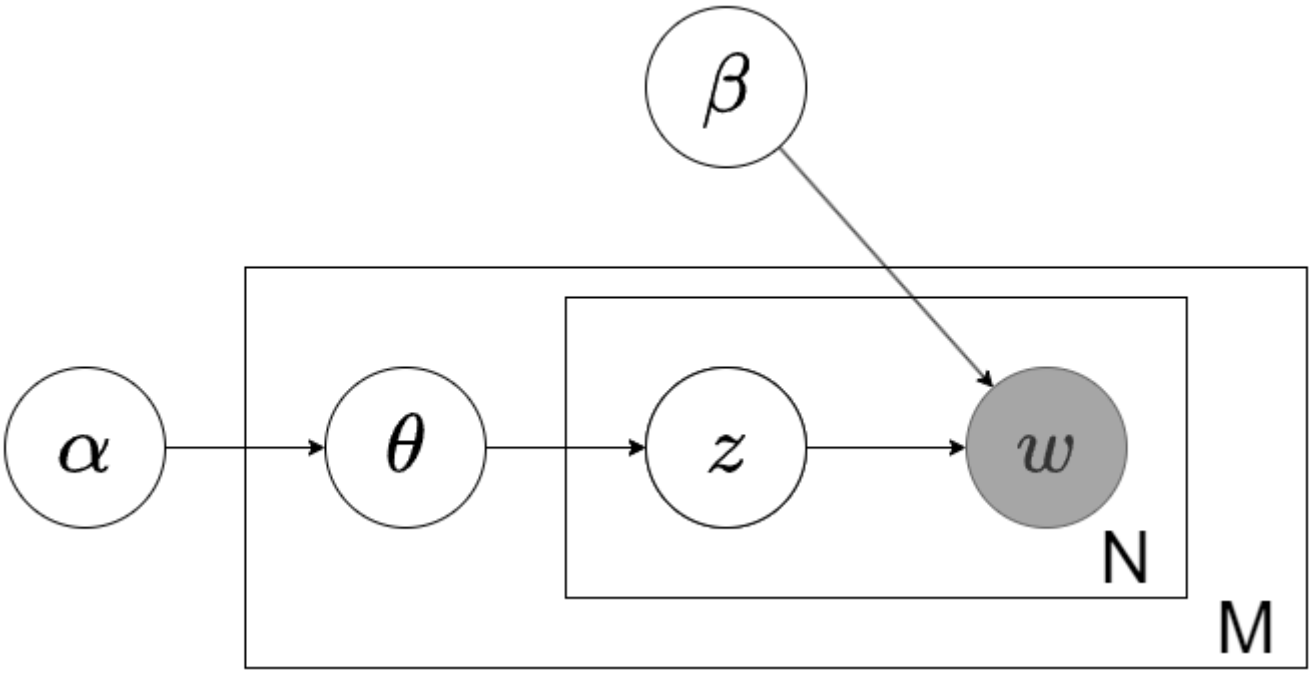


Figure 1: LDA graphical model

To gasp a sense of how complex this model is, we can derive the likelihood of one document given parameters $\alpha$ and $\beta$ (assuming constant $\epsilon$) as follows,

$$
\begin{aligned}
p(w|\alpha,\beta) &= \int p(w,z,\theta|\alpha,\beta)dzd\theta && (1)\\
&= \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta && (2)\\
&= \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\int\prod_{i=1}^{K}\theta_i^{\alpha_i-1}\left(\prod_{n=1}^{N}\sum_{i=1}^{K}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j}\right)d\theta && (3)
\end{aligned}
$$

where $\Gamma$ is the gamma function coming from Dirichlet distribution of $\theta$ and $w_n^j$ means the $n$-th word in $w$ is the $j$-th word in the vocabulary. Here, the complexity originates from the coupled term $\theta_i\beta_{ij}$ , which cannot be separated during computations including the integral. This reflects a strong dependence between the latent variables in LDA. Another problem is the number of coupled terms is $N^K$, which is similar to that of MoG and thus, is not computable as $N$ and $K$ increase. The likelihood of the whole corpus is even harder to compute, which reads:

$$
\begin{aligned}
p(D|\alpha,\beta) &= \prod_{d=1}^{M} p(w_d|\alpha,\beta) && (4)\\
&= \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta_d)p(w_n|z_n,\beta)\right)d\theta_d && (5)
\end{aligned}
$$

## Mean-field solution for LDA

We simplify the joint latent distribution by a mean-field variational distribution:

$$
q(\theta,z) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)
$$

where $q(\theta|\gamma)$ and $q(z_n|\phi_n)$ are Dirichlet and multinomial distributions, respectively. Obviously, $q$ breaks the dependence between $\theta$ and $z$, yet it is not fully mean-field since it still models the joint distribution $\theta$ as a whole instead of individual distributions for each $\theta_i$. That is a reasonable choice since it reduces the approximation error while we can still find the optimal variational distribution in this form (see below).

As usual, we start with deriving the $\mathrm{ELBO}$ for a single document $w$, which in this case is:

$$\begin{aligned}
\text{ELBO}(\gamma, \phi|\alpha, \beta) &= \mathbb{E}_q[\log p(w, \theta, z|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta, z|\gamma, \phi)] && (6)\\
&= \mathbb{E}_q[\log p(w|z, \beta)] + \mathbb{E}_q[\log p(z|\theta)] + \mathbb{E}_q[\log p(\theta|\alpha)] && (7)\\
&- \mathbb{E}_q[\log q(\theta|\gamma)] - \mathbb{E}_q[\log q(z|\phi)] && (8)
\end{aligned}$$

Let's examine each term one by one. The first term focus on the generating distribution--a conditional multinomial, which reads:

$$\begin{aligned}
\mathbb{E}_q[\log p(w|z, \beta)] &= \mathbb{E}_q[\log \prod_{n=1}^{N}\prod_{j=1}^{V}(\beta_{z_n j})^{w_n^j}] && (9)\\
&= \sum_{i=1}^{K}\sum_{n=1}^{N}\sum_{j=1}^{V} q(z_n = i) w_n^j \log \beta_{ij} && (10)\\
&= \sum_{i=1}^{K}\sum_{n=1}^{N}\sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} && (11)
\end{aligned}$$

The second term is:

$$\begin{aligned}
\mathbb{E}_q[\log p(z|\theta)] &= \sum_{n=1}^{N} \mathbb{E}_q \log \theta_{z_n} && (12)\\
&= \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{ni} \mathbb{E}_{\theta_i \sim q} \log \theta_i && (13)\\
&= \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) && (14)
\end{aligned}$$

where $\Psi$ is the digamma function. The third term is just the cross-entropy between 2 Dirichlet distributions:

$$\mathbb{E}_q[\log p(\theta|\alpha)] = \mathbb{E}_{\theta\sim q}[\log \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}] \tag{15}$$

$$= \log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K \mathbb{E}_{\theta\sim q}\left[(\alpha_i - 1)\log \theta_i\right] \tag{16}$$

$$= \log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) \tag{17}$$

$$+ \sum_{i=1}^K (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \tag{18}$$

The fourth term is similar to the third term, yet this time it is just the entropy of a Dirichlet distribution:

$$\mathbb{E}_q[\log q(\theta|\gamma)] = \log \Gamma(\sum_{i=1}^K \gamma_i) - \sum_{i=1}^K \log \Gamma(\gamma_i) \tag{19}$$

$$+ \sum_{i=1}^K (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^K \gamma_j)) \tag{20}$$

The fifth term is basically the entropy of Multinomials:

$$\mathbb{E}_q[\log q(z|\phi)] = \sum_{n=1}^N \mathbb{E}_{z_n\sim\phi_n} q(z_n = i|\phi_n) \tag{21}$$

$$= \sum_{n=1}^N \sum_{i=1}^K \phi_{ni} \log \phi_{ni} \tag{22}$$

Note that all of the terms are easy to compute, and so the $\mathrm{ELBO}$, Now, we need to find the optimal $q$.

**Optimal variational multinomial**

Remember we assume that $p(z) = \prod_{n=1}^N z_n$, so we can directly use the mean-field solution in Lecture 3:

$$q^*(z_n) \propto \exp\{\mathbb{E}_{q_{-z_n}}[\log p(w, \theta, z)]\} \tag{23}$$
$$\Rightarrow \mathbb{E}_{q_{z_n}} q^*(z_n) \propto \mathbb{E}_q[\log p(w, \theta, z)] \tag{24}$$
$$\Rightarrow \mathbb{E}_{q_{z_n}} q^*(z_n) \propto \exp\left\{\sum_{i=1}^{K} \phi_{ni} \sum_{j=1}^{V} w_n^j \log \beta_{ij} + (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)\right\} \tag{25}$$
$$\Rightarrow \phi_{ni} = q^*(z_n = i) \propto \exp\left\{\sum_{j=1}^{V} w_n^j \log \beta_{ij} + (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)\right\} \tag{26}$$

We can also redo the optimisation from scratch by writing the $\mathrm{ELBO}$ as a function of $\{\phi_{ni}\}_{i=1}^K$ and find the maximisers $\{\phi_{ni}\}_{i=1}^K$

such that $\sum_{i=1}^K \phi_{ni} = 1$. Again we can use Lagrange multipliers to solve the constrained optimisation, creating the following Lagrangian:

$$L(\{\phi_{ni}\}_{i=1}^K) = \sum_{i=1}^{K}\sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} + \sum_{i=1}^{K} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) \tag{27}$$
$$- \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni} + \lambda_n(\sum_{j=1}^{K} \phi_{nj} - 1) \tag{28}$$

Solving this problem will lead to the same solution as Eq. (23).

**Optimal variational Dirichlet**

As $q(\theta|\gamma)$ does not follow mean-field assumption, we need to directly maximise $\mathrm{ELBO}$ w.r.t $\gamma_i$ (note there is generally no constraint). Let's first write $\mathrm{ELBO}$ as a function of $\gamma_i$:

$$L(\gamma_i) = \sum_{n=1}^{N} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) + (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) \tag{29}$$
$$- \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \log \Gamma(\gamma_i) - (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) \tag{30}$$
$$= (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j))(\sum_{i=1}^{N} \phi_{ni} + \alpha_i - \gamma_i) - \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \log \Gamma(\gamma_i) \tag{31}$$
$$\Rightarrow \frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i)(\sum_{i=1}^{N} \phi_{ni} + \alpha_i - \gamma_i) - \Psi'(\sum_{j=1}^{K} \gamma_j)(\sum_{i=1}^{N} \phi_{ni} + \alpha_i - \gamma_i) \tag{32}$$

Finding the root of this equation, despite its complexity, we can easily have a local optimal:

$$\gamma_i = \sum_{i=1}^{N} \phi_{ni} + \alpha_i$$

The above formulations are derived for single document. To account for the whole corpus, we need to estimate variational parameters for all documents as $\phi_{dni}$ and $\gamma_{di}$ using the corresponding words in the documents $w_{di}^j$.

## Model's Parameter Estimation

After we find the update rule for variational parameters, we continue with the model's parameters to complete the loop of variational EM. The process is quite similar to optimising the variational parameters where we maximise the $\mathrm{ELBO}$ w.r.t $\beta$ and $\alpha$.

### Optimal Conditional Multinomial

Let's rewrite the $\mathrm{ELBO}$ for all documents as a function of $\beta$ and form the Lagrangian

$$L(\beta) = \sum_{d=1}^{M}\sum_{i=1}^{K}\sum_{n=1}^{N}\sum_{j=1}^{V} \phi_{dni} w_{dn}^j \log \beta_{ij} + \sum_{i=1}^{K} \lambda_i (\sum_{j=1}^{V} \beta_{ij} - 1) \tag{33}$$

$$\Rightarrow \frac{\partial L}{\partial \beta_{ij}} = \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dni} w_{dn}^j / \beta_{ij} + \sum_{i=1}^{K} \lambda_i \tag{34}$$

Setting it to zero leads to

$$\beta_{ij} \propto \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dni} w_{dn}^j \tag{35}$$

The meaning turns out to be simple. To estimate the probability of vocab $j$ belonging to topic $i$, we count the number of $j$-th vocab appears in the corpus weighted by the variational probability of topic $i$ assigned to the appeared word.

### Optimal Dirichlet

Again, we rewrite the $\mathrm{ELBO}$ for all documents as a function of $\alpha$:

$$L(\alpha) = \sum_{d=1}^{M} \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \tag{36}$$

$$+ \sum_{d=1}^{M} \sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{K} \gamma_{dj})) \tag{37}$$

$$\Rightarrow \frac{\partial L}{\partial \alpha_i} = M(\Psi(\sum_{j=1}^{K} \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^{M} \Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{K} \gamma_{dj}) \tag{38}$$

Unfortunately, unlike the case of variational Dirichlet, there is no easy local optimal (i.e. the root of Eq. (34)). The equation involves other parameters $\alpha_j$ and thus can only be solved by iterative methods. In practice, we can ignore optimising for $\alpha$ and treat them as hyperparameters needed tuning.

## Smoothed LDA

One problem with the current LDA is when a test document $w'$ contains a novel word $v$ that are not existing in the training corpus (it can be in the vocabulary set). In this case, $\beta_{iv} = 0$ , $p(w') = 0$ (just assuming that we can compute it using Eq. (3)) and thus, we cannot infer the topics for this document $p(z|w')$. Even we approximate it with the variational distribution, we cannot estimate $\phi_{ni}$ using Eq. (23) at $w_n = v$.

One solution is to treat $\beta_{ij}$ as a random variable drawing from a prior Dirichlet distribution. For simplicity, the prior is shared across $\beta_{i,1:V}$ and modelled as an exchangeable Dirichlet with single parameter $\eta$; hence, $\beta_i \sim \text{Dir}(\beta_i|\eta)$. The graphical model now becomes:

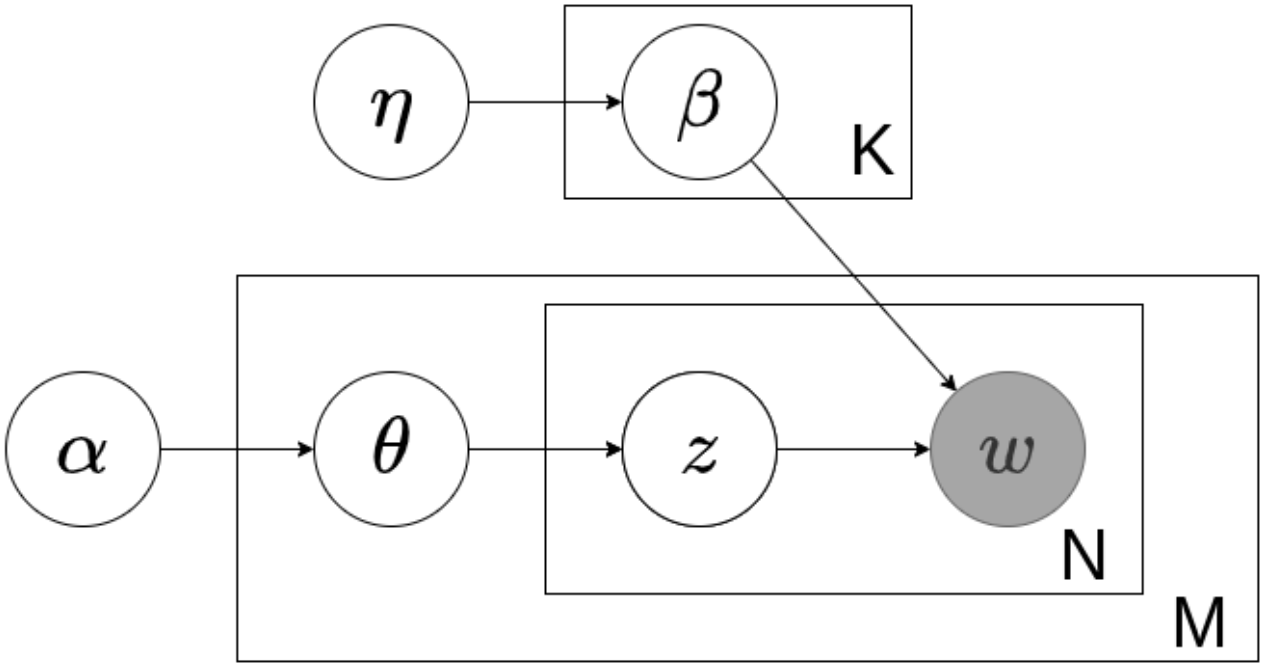


Figure 2: Smoothed LDA graphical model

To efficiently optimise this LVM, we again formulate a mean-field variational distribution of the latent variables, which in this case reads

$$q(\beta, z, \theta) = \prod_{i=1}^{K} \mathrm{Dir}(\beta_i|\lambda_i) \prod_{d=1}^{M} [q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{nd}|\phi_{dn})] \tag{39}$$

where $\lambda_i$ is the new variational parameter. Introducing the prior on $\beta$ modifies the $\mathrm{ELBO}$ in Eq. (6-7) with two additional terms $\mathbb{E}_q[\log p(\beta|\eta)]$ and $\mathbb{E}_q[\log q(\beta|\lambda)]$. The modification does not change the $\phi^*_{dn}$ and $\gamma^*_d$. Maximising the $\mathrm{ELBO}$ w.r.t $\lambda_i$ yields:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N} \phi^*_{dni} w^j_{dn} \tag{40}$$

To find the optimal update for $\eta$, we also use the variational EM procedure. Unfortunately, maximize the $\mathrm{ELBO}$ w.r.t $\eta$ requires iterative methods just as the case of $\alpha$, which does not have a close-form solution.

## Exercises

1. Draw the graphical model of the LDA under mean-field assumption.
2. Prove Eq. (35)
3. Prove Eq. (40)