Part I:
Reminder of Dirichlet and Multinomial distributions
and inference for them

1. Dirichlet distributions

\[ \mu = (\mu_1, \ldots, \mu_k)^T \]  
Constraint \( \sum_{i=1}^{k} \mu_i = 1 \)

\[ \alpha = (\alpha_1, \ldots, \alpha_k)^T \quad \alpha_0 = \sum_{i=1}^{n} \alpha_i \]

\[ Pr(\mu|\alpha) = \text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \prod \mu_i^{\alpha_i-1} \]

2. Multinomial distribution over \( x \in \{1, \ldots, k\} \) where value \( i \) is denoted in binary notation as \( i \)'th unit vector.

\[ Pr(x|\mu) = \prod_{i=1}^{k} \mu_i^{x_i} \]

3. We observe \( N \) values \( D = x_1, \ldots, x_N \) and \( x_n \in \{1, \ldots, k\} \) is denoted in binary notation

4. Likelihood \( L = Pr(D|\mu) = \prod_{n=1}^{N} \prod_{i=1}^{k} \mu_i^{x_{n,i}} = \prod_{i=1}^{k} \mu_i^{m_i} \)

   where \( m_i \) counts occurrences of value \( i \).

5. The posterior \( Pr(\mu|D) \propto \prod_{i=1}^{k} \mu_i^{m_i+\alpha_i-1} \)

\[ Pr(\mu|D) = \text{Dir}(\mu|\alpha + m) \]  
where \( m = (m_1, \ldots, m_k)^T \) and \( \sum \alpha_i + m_i = \alpha_0 + N \)

6. (from assignment 1) The predictive distribution

\[ Pr(x_\ell = 1|D) = \int \text{Pr}(\mu|D) Pr(x_\ell = 1|\mu) d\mu = \ldots = \frac{\alpha_\ell + m_\ell}{\alpha_0 + N} \]

7. (from assignment 1) The evidence function

\[ Pr(D|\alpha) = \int Pr(\mu|\alpha) Pr(D|\mu) d\mu = \ldots = \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \prod \frac{\Gamma(\alpha_i + m_i)}{\Gamma(\alpha_0 + N)} \]
Part II: 
The LDA model and Gibbs sampling for it

1. $K$ Topics $\{1, \ldots, K\}$
   Vocabulary $\{1, \ldots, V\}$
   Topic $z \in \{1, \ldots, K\}$ associated with every word in every document.
   Every word is an element in the Vocabulary $i \in \{1, \ldots, V\}$

2. The model
   $d$ denotes a document
   $\theta_{k|d} = Pr(z = k|d)$
   $\phi_{ik|k} = Pr(\text{word} = i|z = k)$
   Prior: $[\prod_d \text{Dir}(\theta_{d|\alpha})] [\prod_k \text{Dir}(\phi_k|\beta)]$

3. The complete data includes all topic assignments $Z$ and all the words $D$.
   We only observe $D$ but for now calculate quantities for the complete data.

4. Notation
   $N_d$ is the number of words in $d$
   We observe a total of $N$ words, $N = \sum N_d$ where
   $N_k$ is the number of times topic $k$ is used over all documents (this is an abuse of notation but
   the meaning should be clear from context)
   $N_{k|d}$ is the number of times topic $k$ is used in document $d$
   $N_{i|k}$ is the number of times word $i$ is chosen when the topic is $k$ over all documents

5. Complete data Likelihood
   $\mathcal{L} = Pr(D, Z|\phi, \theta) = [\prod_d \prod_k \theta_{k|d}^{N_{k|d}}] [\prod_k \prod_i \phi_{ik|k}^{N_{i|k}}]$

6. Complete data Posterior
   $Pr(\theta_{d|D, Z}) = \text{Dir}(\theta_{d|\alpha + N_{k|d}})$
   $Pr(\phi_{k|D, Z}) = \text{Dir}(\phi_{k|\beta + N_{i|k}})$
   Posterior: $[\prod_d \text{Dir}(\theta_{d|\alpha + N_{k|d}})] [\prod_k \text{Dir}(\phi_{k|\beta + N_{i|k}})]$

7. Complete data Predictive Distribution for new word in document $d$
   $Pr(z_k = 1|D, Z, d) = \frac{\alpha_k + N_{k|d}}{\alpha_0 + N_d}$
   $Pr(\text{word} = i|D, Z, d, z_k = 1) = \frac{\beta_i + N_{i|k}}{\beta_0 + N_k}$

8. Complete data Evidence function
   $Pr(D, Z|\alpha, \beta) = [\prod_d \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_k)}] [\prod \frac{\Gamma(\alpha_0 + N_{k|d})}{\Gamma(\alpha_k + N_{k|d})}] [\prod_k \frac{\Gamma(\beta_0)}{\Gamma(\beta_k)}] [\prod \frac{\Gamma(\beta_0 + N_{i|k})}{\Gamma(\beta_k + N_{i|k})}]$

9. To calculate posterior or MAP over $\phi$ and $\theta$ using above we need to know $Z$ which is not observed. For this type of problem we can use the EM algorithm (discussed in next lecture) but this will require us to use some further approximations.
10. Alternatively we can sample from the posterior using Gibbs sampling. Here we marginalize $\phi, \theta$ and sample $Z$ directly.

For Gibbs sampling we want to resample each $z^{j,d}$ – the topic of the $j$’th word in document $d$ – conditioned on the rest of the complete data.

Let $N^-$ (and all quantities above decorated with superscript $-$) be the same as before counting all words except the $j$’th word in document $d$.

Note that the word itself is observed but the topic that generated it is not observed.

$$Pr(z^{j,d}_k = 1|D^-,Z^-, \text{word}^{j,d}_i = i) \propto Pr(z^{j,d}_k = 1|D^-,Z^-)Pr(\text{word}^{j,d}_i = i|z^{j,d}_k = 1)$$

$$= \frac{\alpha_k + N^-_{k|d}}{\alpha_0 + N^-_d} \frac{\beta_i + N^-_{i|k}}{\beta_0 + N^-_k}$$

The reasoning for the step above is that in the probability of the complete data (including $z^{j,d}$ and $Z^-$) all other terms multiplying the two above are identical for all assignments of topic to $z^{j,d}$. Therefore these terms provide the same constant factor multiplying the above over all assignments. The get the concrete probability for sampling we simply calculate the equation above for all $k$ and normalize.

11. After many steps of sampling we get one random assignment to $Z$ sampled from its posterior distribution. This can be used to yield multiple random samples as in standard approaches to Gibbs sampling. The multiple samples can be used to estimate “summary quantities” of the posterior. For example one can approximate the true evidence function $Pr(D|\alpha, \beta) = \sum_Z Pr(D,Z|\alpha, \beta)$ using the samples, and perform model selection by optimizing $\alpha$ and $\beta$ using the approximation.

12. However, we need to be careful when estimating concrete topics because topics may not be stable across samples (for example their id’s can be swapped). Therefore concrete topics are typically taken from a single sample. In particular, given one sample one can treat it as the complete true data and estimate the posterior (or MAP) of $\phi$ and $\theta$ using the equation given above.