Latent Dirichlet Allocation

STAT4609 Big Data Analysis

Example Class 7

LIU Chen

16 Mar, 2023

Latent Dirichlet Allocation

Outline

- The foundations of Bayesian Parameter Estimation in the discrete domain.
- Latent Dirichlet Allocation
  - Mixure Modelling
  - Generative Model
  - Inference via Gibbs Sampling
  - The collapsed LDA Gibbs sampler

LDA: intuition

Latent Dirichlet allocation (LDA) by Blei et al. is a probabilistic generative model that can be used to estimate the properties of multinomial observations by unsupervised learning.

The intuition is to find the latent structure of “topics” or “concepts” in a text corpus, which captures the meaning of the text that is imagined to be obscured by “word choice” noise.

It has been empirically showed that the co-occurrence structure of terms in text documents can be used to recover this latent.

LDA: Mixure Model I

In LDA, a word w is generated from a convex combination of topics z. In such a mixture model, the probability that a word w instantiates term t is:

$$p(w = t) = \sum_k p(w = t \mid z = k) p(z = k), \quad \sum_k p(z = k) = 1,$$

where each mixture component $p(w = t \mid z = k)$ is a multinomial distribution over terms (cf. the unigram model above) that corresponds to one of the latent topics $z = k$ of the text corpus. The mixture proportion consists of the topic probabilities $p(z = k)$.

LDA: Mixure Model II

The Main Objectives of LDA inference

LDA goes a step beyond a global topic proportion and conditions the topic probabilities on the document a word belongs to.

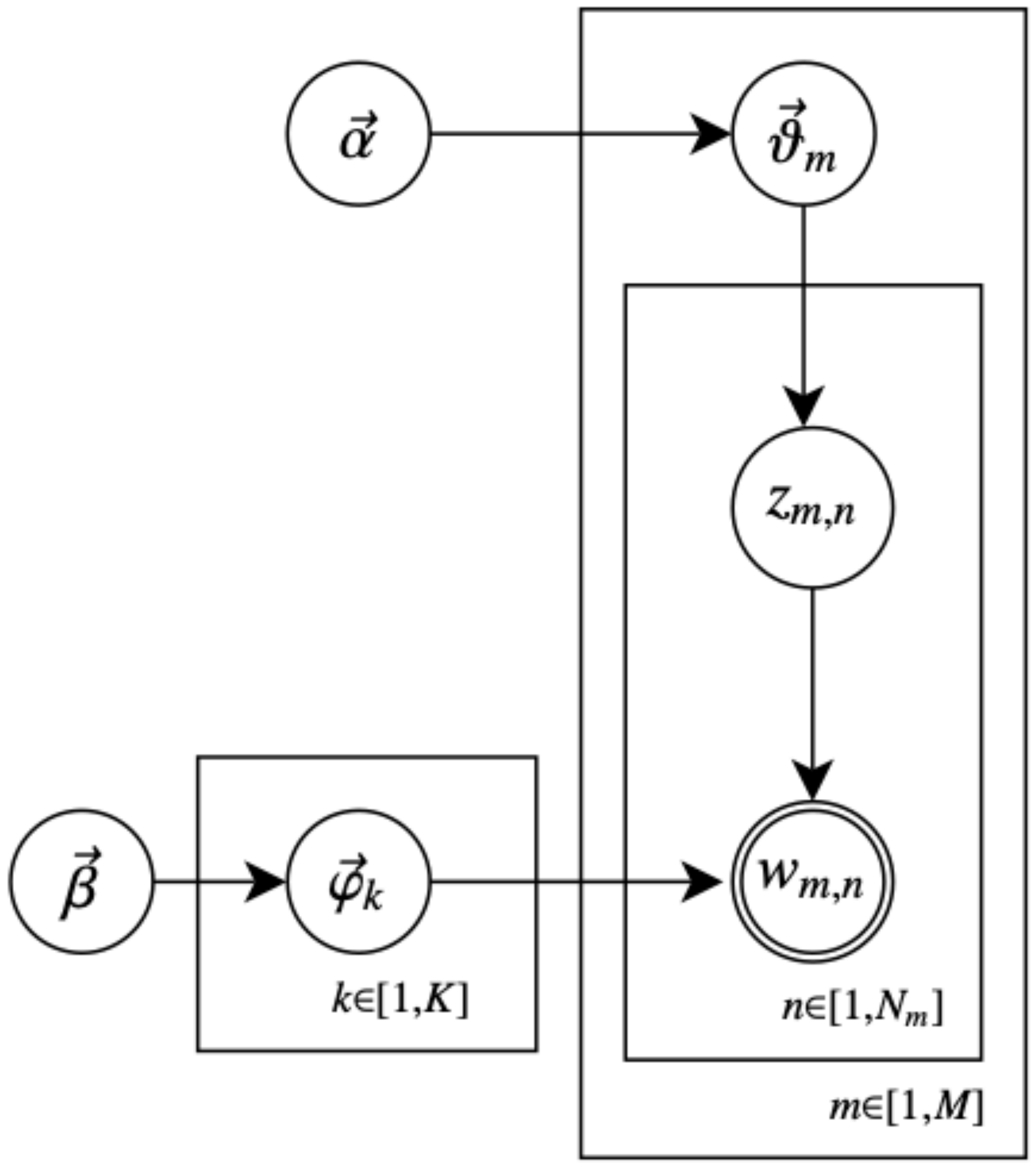
Thus, we can formulate the main objectives of LDA inference:

1. **to find the term distribution** $p(t \mid z = k) = \vec{\varphi}_k$ **for each topic** k**;**
2. **to find the topic distribution** $p(z \mid d = m) = \vec{\vartheta}_m$ **for each document** m**.**

The estimated parameter sets $\underline{\Phi} = \left\{ \vec{\varphi}_k \right\}_{k=1}^{K}$ and $\underline{\Theta} = \left\{ \vec{\vartheta}_m \right\}_{m=1}^{M}$ are the basis for latent-semantic representation of words and documents.

LDA: Generative Model

To derive an inference strategy, we view LDA as a generative process.



Bayesian network of latent Dirichlet allocation.

Consider the Bayesian network of LDA shown on the left. This can be interpreted as follows: **LDA generates a stream of observable words $w_{m,n}$, partitioned into documents $\vec{w}_m$.**

- The topics $\vec{\varphi}_k$ (The probability of words occurring in topic k) are sampled once for the entire corpus.
- For each of these documents, a topic proportion $\vec{\vartheta}_m$ (The distribution of topics in document m) is drawn.
  - For each word position (m, n),
    - sample a topic indicator $z_{m,n} \sim \text{Multinomial}(\vec{\vartheta}_m)$,
    - sample a word $w_{m,n} \sim \text{Multinomial}(\vec{\varphi}_{z_{m,n}})$

LDA: Generative Model

We can mathematically describe the random variables as follows,

- The distribution of word in topic k,

$$\overrightarrow{\varphi}_{k=1...K} \sim \mathrm{Dir}(\overrightarrow{\beta})$$

- The distribution of topics in document d,

$$\overrightarrow{\vartheta}_{m=1...M} \sim \mathrm{Dir}(\overrightarrow{\alpha})$$

- The topic indicator at word position (m, n)

$$z_{m=1...M,n=1...N_m} \sim \mathrm{Multinomial}(\overrightarrow{\vartheta}_m),$$

- identify of word n in document m

$$w_{m=1...M,n=1...N_m} \sim \mathrm{Multinomial}(\overrightarrow{\varphi}_{z_{m,n}})$$

Pseudo Code of Generative Model for LDA

`// topic plate`
for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim \mathrm{Dir}(\vec{\beta})$

`// document plate:`
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim \mathrm{Dir}(\vec{\alpha})$
  sample document length $N_m \sim \mathrm{Poiss}(\xi)$
  `// word plate:`
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n} \sim \mathrm{Mult}(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim \mathrm{Mult}(\vec{\varphi}_{z_{m,n}})$

LDA: Generative Model

Quantities in the model of LDA

M	number of documents to generate (const scalar).
K	number of topics / mixture components (const scalar).
V	number of terms \$t\$ in vocabulary (const scalar).
$\vec{\alpha}$	hyper-parameter on the mixing proportions (K-vector or scalar if symmetric).
$\vec{\beta}$	hyper-parameter on the mixture components (V-vector or scalar if symmetric).
$\vec{\vartheta}_m$	parameter notation for $p(z \mid d = m)$, the topic mixture proportion for document m.
	One proportion for each document, $\underline{\Theta} = \left\{ \vec{\vartheta}_m \right\}_{m=1}^{M}$ ($M \times K$ matrix).
$\vec{\varphi}_k$	parameter notation for $p(t \mid z = k)$, the mixture component of topic k. One component for each topic, $\underline{\Phi} = \left\{ \vec{\varphi}_k \right\}_{k=1}^{K}$($K \times V$ matrix).
N_m	document length (document-specific), here modelled with a Poisson distribution with constant parameter ξ.
$z_{m,n}$	mixture indicator that chooses the topic for the n-th word in document m.
$w_{m,n}$	term indicator for the n-th word in document m.

LDA: Likelihoods

Looking at the topology of the Bayesian network, we can specify the complete-data likelihood of a document m, i.e., **the joint distribution of all known and hidden variables**, given the hyperparameters:

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi} | \vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n} | \vec{\varphi}_{z_{m,n}}) p(z_{m,n} | \vec{\vartheta}_m)}_{\text{word plate}} \cdot p(\vec{\vartheta}_m | \vec{\alpha})}^{\text{document plate (1 document)}} \cdot \underbrace{p(\underline{\Phi} | \vec{\beta})}_{\text{topic plate}} . \qquad (56)$$

LDA: Likelihoods

The complete data likelihood of a document

Looking at the topology of the Bayesian network, we can specify the complete-data likelihood of a document m, i.e., **the joint distribution of all known and hidden variables**, given the hyperparameters:

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi} | \vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n} | \vec{\varphi}_{z_{m,n}}) p(z_{m,n} | \vec{\vartheta}_m)}_{\text{word plate}} \cdot p(\vec{\vartheta}_m | \vec{\alpha})}^{\text{document plate (1 document)}} \cdot \underbrace{p(\underline{\Phi} | \vec{\beta})}_{\text{topic plate}} . \qquad (56)$$

LDA: Likelihoods

The likelihood of a word given LDA parameters

To specify this distribution is simple and useful as a basis for other derivations. So the probability that a word $w_{m,n}$ instantiates a particular term t given the LDA parameters is obtained by marginalising $z_{m,n}$ from the word plate and omitting the parameter distributions:

$$p\left(w_{m,n} = t \mid \vec{\vartheta}_m, \underline{\Phi}\right) = \sum_{k=1}^{K} p\left(w_{m,n} = t \mid \vec{\varphi}_k\right) p\left(z_{m,n} = k \mid \vec{\vartheta}_m\right)$$

LDA: Likelihoods

The likelihood of a word given LDA parameters

The likelihoods of a document $\vec{w}_m$ and of the corpus $\mathscr{W} = \left\{\vec{w}_m\right\}_{m=1}^{M}$ are just the joint likelihoods of the independent events of the token observations $w_{m,n}$:

$$p(\boldsymbol{W} \mid \underline{\boldsymbol{\Theta}}, \underline{\Phi}) = \prod_{m=1}^{M} p\left(\vec{w}_m \mid \vec{\boldsymbol{\vartheta}}_m, \underline{\Phi}\right) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p\left(w_{m,n} \mid \vec{\boldsymbol{\vartheta}}_m, \underline{\Phi}\right)$$

Inference via Gibbs sampling

Why do we choose Gibbs sampling for LDA?

Although latent Dirichlet allocation is still a relatively simple model, exact inference is generally intractable. The solution to this is to use approximate inference algorithms, such as mean-field variational expectation maximisation, expectation propagation, and Gibbs sampling.

Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation and often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA. Therefore we select this approach and present a derivation that is more detailed than the original one by Griffiths and Steyvers

Inference via Gibbs sampling

proximate high-dimensional probability distribution

MCMC methods can emulate high-dimensional probability distributions $p(\vec{x})$ by the stationary behaviour of a Markov chain.

This means that one sample is generated for each transition in the chain after a stationary state of the chain has been reached, which happens after a so-called "burn-in period" that eliminates the influence of initialisation parameters.

Gibbs sampling is a special case of MCMC where the dimensions x_i of the distribution are sampled alternately one at a time, conditioned on the values of all other dimensions, which we denote $\vec{x}_{\neg i}$. The algorithm works as follows:

1. choose dimension i (random or by permutation)

2. sample x_i from $p\left(x_i \mid \vec{x}_{\neg i}\right)$

To build a Gibbs sampler, the univariate conditionals (or full conditionals) $p\left(x_i \mid \vec{x}_{\neg i}\right)$ must be found, which is possible using:

$$p\left(x_i \mid \vec{x}_{\neg i}\right) = \frac{p(\vec{x})}{p\left(\vec{x}_{\neg i}\right)} = \frac{p(\vec{x})}{\int p(\vec{x})\mathrm{d}x_i} \text{ with } \vec{x} = \left\{x_i, \vec{x}_{\neg i}\right\}$$

Inference via Gibbs sampling

with latent variable, to approximate the posterior, $p(\vec{z} \mid \vec{x})$.

For models that contain hidden variables $\vec{z}$, their posterior given the evidence, $p(\vec{z} \mid \vec{x})$, is a distribution commonly wanted. With previous equation, the general formulation of a Gibbs sampler for such latent-variable models becomes:

$$p\left(z_i \mid \vec{z}_{\neg i}, \vec{x}\right) = \frac{p(\vec{z}, \vec{x})}{p\left(\vec{z}_{\neg i}, \vec{x}\right)} = \frac{p(\vec{z}, \vec{x})}{\int_Z p(\vec{z}, \vec{x}) \mathrm{d}z_i}$$

where the integral changes to a sum for discrete variables. With a sufficient number of samples $\tilde{z}_r$, $r \in [1, R]$, the latent-variable posterior can be approximated using:

$$p(\vec{z} \mid \vec{x}) \approx \frac{1}{R} \sum_{r=1}^{R} \delta\left(\vec{z} - \tilde{\vec{z}}_r\right)$$

with the Kronecker delta $\delta(\vec{u}) = \{1 \text{ if } \vec{u} = 0; 0 \text{ otherwise}\}$

LDA: The target of inference

- The target of inference is the distribution $p(\vec{z} \mid \vec{w})$, which is directly proportional to the joint distribution

$$p(\vec{z} \mid \vec{w}) = \frac{p(\vec{z}, \vec{w})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p\left(z_i, w_i\right)}{\prod_{i=1}^{W} \sum_{k=1}^{K} p\left(z_i = k, w_i\right)} \tag{62}$$

where the hyperparameters are omitted.

This distribution covers a large space of discrete random variables, and the difficult part for evaluation is its denominator, which represents a summation over K^W terms. At this point, the Gibbs sampling procedure comes into play. In our setting, the desired Gibbs sampler runs a Markov chain that uses the full conditional $p\left(z_i \mid \vec{z}_{\neg i}, \vec{w}\right)$ in order to simulate $p(\vec{z} \mid \vec{w})$. We can obtain the full conditional via the hidden-variable approach by evaluating Eq. 60, which requires to formulate the joint distribution.

Gibbs sampling algorithm for LDA

Initialisation

Algorithm `LdaGibbs`($\{\vec{w}\}, \alpha, \beta, K$)

Input: word vectors $\{\vec{w}\}$, hyperparameters α, β, topic number K

Global data: count statistics $\{n_m^{(k)}\}$, $\{n_k^{(t)}\}$ and their sums $\{n_m\}$, $\{n_k\}$, memory for full conditional array $p(z_i|\cdot)$

Output: topic associations $\{\vec{z}\}$, multinomial parameters $\underline{\Phi}$ and $\underline{\Theta}$, hyperparameter estimates α, β

```
// initialisation
```

zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$

for all documents $m \in [1, M]$ **do**

  for all words $n \in [1, N_m]$ in document m **do**

    sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$

    increment document–topic count: $n_m^{(k)}$ += 1

    increment document–topic sum: n_m += 1

    increment topic–term count: $n_k^{(t)}$ += 1

    increment topic–term sum: n_k += 1

Gibbs sampling algorithm for LDA

Sampling

```
// Gibbs sampling over burn-in period and sampling period
```
while not finished **do**

  for all documents $m \in [1, M]$ **do**

    for all words $n \in [1, N_m]$ in document m **do**

      `// for the current assignment of` k `to a term` t `for word` $w_{m,n}$`:`

      decrement counts and sums: $n_m^{(k)} -= 1; n_m -= 1; n_k^{(t)} -= 1; n_k -= 1$

      `// multinomial sampling acc. to Eq.` $\boxed{78}$ `(decrements from previous step):`

      sample topic index $\tilde{k} \sim p(z_i | \vec{z}_{\neg i}, \vec{w})$

      `// for the new assignment of` $z_{m,n}$ `to the term` t `for word` $w_{m,n}$`:`

      increment counts and sums: $n_m^{(\tilde{k})} += 1; n_m += 1; n_{\tilde{k}}^{(t)} += 1; n_{\tilde{k}} += 1$

  `// check convergence and read out parameters`

  if converged and L sampling iterations since last read out **then**

    `// the different parameters read outs are averaged.`

    read out parameter set $\underline{\Phi}$ according to Eq. $\boxed{81}$

    read out parameter set $\underline{\Theta}$ according to Eq. $\boxed{82}$

Multinomial parameters

- Multinomial parameters. Finally, we need to obtain the multinomial parameter sets Θ and Φ that correspond to the state of the Markov chain, $\vec{z}$. According to their definitions as multinomial distributions with Dirichlet prior, applying Bayes' rule on the component $z = k$ in Eq. 65 and m in Eq. 69 yields:

$$p\left(\vec{\vartheta}_m \mid \vec{z}_m, \vec{\alpha}\right) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p\left(z_{m,n} \mid \vec{\vartheta}_m\right) \cdot p\left(\vec{\vartheta}_m \mid \vec{\alpha}\right) = \mathrm{Dir}\left(\vec{\vartheta}_m \mid \vec{n}_m + \vec{\alpha}\right) \tag{79}$$

$$p\left(\vec{\varphi}_k \mid \vec{z}, \vec{w}, \vec{\beta}\right) = \frac{1}{Z_{\varphi_k}} \prod_{\{i: z_i = k\}} p\left(w_i \mid \vec{\varphi}_k\right) \cdot p\left(\vec{\varphi}_k \mid \vec{\beta}\right) = \mathrm{Dir}\left(\vec{\varphi}_k \mid \vec{n}_k + \vec{\beta}\right) \tag{80}$$

- where $\vec{n}_m$ is the vector of topic observation counts for document m and $\vec{n}_k$ that of term observation counts for topic k. Using the expectation of the Dirichlet distribution, $\langle \mathrm{Dir}(\vec{a}) \rangle = a_i / \sum_i a_i$, on these results yields:

$$\hat{\varphi_{k,t}} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} (n_k^{(t)} + \beta_t)} = E\left(\phi_{k,t} \mid \vec{z}, \vec{w}, \vec{\beta}\right) \tag{81}$$

$$\hat{\vartheta_{m,k}} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)} = E\left(\theta_{m,k} \mid \vec{z}_m, \vec{\alpha}\right) \tag{82}$$

Joint distribution

- In LDA, this joint distribution can be factored: $p(\vec{w}, \vec{z} \mid \vec{\alpha}, \vec{\beta}) = p(\vec{w} \mid \vec{z}, \vec{\beta}) p(\vec{z} \mid \vec{\alpha})$
- The target distribution $p(\vec{w} \mid \vec{z}, \vec{\beta})$ is obtained by integrating over $\underline{\Phi}$, which can be done componentwise using Dirichlet integrals within the product over z:

$$p(\vec{w} \mid \vec{z}, \vec{\beta}) = \int p(\vec{w} \mid \vec{z}, \underline{\Phi}) p(\underline{\Phi} \mid \vec{\beta}) \mathrm{d}\underline{\Phi} \qquad (66)$$

$$= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1} \, \mathrm{d}\vec{\varphi}_z \qquad (67)$$

$$= \prod_{z=1}^{K} \frac{\Delta\left(\vec{n}_z + \vec{\beta}\right)}{\Delta(\vec{\beta})}, \quad \vec{n}_z = \left\{n_z^{(t)}\right\}_{t=1}^{V} \qquad (68)$$

This can be interpreted as a product of K Dirichlet-multinomial models (cf. Eq. 52), representing the corpus by K separate "topic texts".

Joint distribution

Integrating out $\underline{\Theta}$, we obtain:

$$p(\vec{z} \mid \vec{\alpha}) = \int p(\vec{z} \mid \underline{\Theta}) p(\underline{\Theta} \mid \vec{\alpha}) \mathrm{d}\underline{\Theta} \tag{70}$$

$$= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1} \, \mathrm{d}\vec{\vartheta}_m \tag{71}$$

$$= \prod_{m=1}^{M} \frac{\Delta\left(\vec{n}_m + \vec{\alpha}\right)}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \left\{n_m^{(k)}\right\}_{k=1}^{K} \tag{72}$$

The joint distribution therefore becomes:

$$p(\vec{z}, \vec{w} \mid \vec{\alpha}, \vec{\beta}) = \prod_{z=1}^{K} \frac{\Delta\left(\vec{n}_z + \vec{\beta}\right)}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta\left(\vec{n}_m + \vec{\alpha}\right)}{\Delta(\vec{\alpha})} \tag{73}$$

Full conditional

- Using the chain rule and noting that $\vec{w} = \{w_i = t, \vec{w}_{\neg i}\}$ and $\vec{z} = \{z_i = k, \vec{z}_{\neg i}\}$ yields:

$$p\left(z_i = k \mid \vec{z}_{\neg i}, \vec{w}\right) = \frac{p(\vec{w}, \vec{z})}{p\left(\vec{w}, \vec{z}_{\neg i}\right)} = \frac{p(\vec{w} \mid \vec{z})}{p\left(\vec{w}_{\neg i} \mid \vec{z}_{\neg i}\right) p\left(w_i\right)} \cdot \frac{p(\vec{z})}{p\left(\vec{z}_{\neg i}\right)} \tag{74}$$

$$\propto \frac{\Delta\left(\vec{n}_z + \vec{\beta}\right)}{\Delta\left(\vec{n}_{z,\neg i} + \vec{\beta}\right)} \cdot \frac{\Delta\left(\vec{n}_m + \vec{\alpha}\right)}{\Delta\left(\vec{n}_{m,\neg i} + \vec{\alpha}\right)} \tag{75}$$

$$= \frac{\Gamma\left(n_k^{(t)} + \beta_t\right) \Gamma\left(\sum_{t=1}^V n_{k,\neg i}^{(t)} + \beta_t\right)}{\Gamma\left(n_{k,\neg i}^{(t)} + \beta_t\right) \Gamma\left(\sum_{t=1}^V n_k^{(t)} + \beta_t\right)} \cdot \frac{\Gamma\left(n_m^{(k)} + \alpha_k\right) \Gamma\left(\sum_{k=1}^K n_{m,\neg i}^{(k)} + \alpha_k\right)}{\Gamma\left(n_{m,\neg i}^{(k)} + \alpha_k\right) \Gamma\left(\sum_{k=1}^K n_m^{(k)} + \alpha_k\right)} \tag{76}$$

$$= \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^V \left(n_{k,\neg i}^{(t)} + \beta_t\right)} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\left[\sum_{k=1}^K \left(n_m^{(k)} + \alpha_k\right)\right] - 1} \tag{77}$$

where the counts $n_{\cdot,\neg i}^{(\cdot)}$ indicate that the token i is excluded from the corresponding document or topic and the hyperparameters are omitted.

Multinomial parameters

- Multinomial parameters. Finally, we need to obtain the multinomial parameter sets Θ and Φ that correspond to the state of the Markov chain, $\vec{z}$. According to their definitions as multinomial distributions with Dirichlet prior, applying Bayes' rule on the component $z = k$ in Eq. 65 and m in Eq. 69 yields:

$$p\left(\vec{\vartheta}_m \mid \vec{z}_m, \vec{\alpha}\right) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p\left(z_{m,n} \mid \vec{\vartheta}_m\right) \cdot p\left(\vec{\vartheta}_m \mid \vec{\alpha}\right) = \mathrm{Dir}\left(\vec{\vartheta}_m \mid \vec{n}_m + \vec{\alpha}\right) \quad (79)$$

$$p\left(\vec{\varphi}_k \mid \vec{z}, \vec{w}, \vec{\beta}\right) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p\left(w_i \mid \vec{\varphi}_k\right) \cdot p\left(\vec{\varphi}_k \mid \vec{\beta}\right) = \mathrm{Dir}\left(\vec{\varphi}_k \mid \vec{n}_k + \vec{\beta}\right) \quad (80)$$

- where $\vec{n}_m$ is the vector of topic observation counts for document m and $\vec{n}_k$ that of term observation counts for topic k. Using the expectation of the Dirichlet distribution, $\langle \mathrm{Dir}(\vec{a}) \rangle = a_i / \sum_i a_i$, on these results yields:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} (n_k^{(t)} + \beta_t)} = E\left(\phi_{k,t} \mid \vec{z}, \vec{w}, \vec{\beta}\right) \quad (81)$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)} = E\left(\theta_{m,k} \mid \vec{z}_m, \vec{\alpha}\right) \quad (82)$$

Reference

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

Heinrich, Gregor. Parameter estimation for text analysis. Technical report, 2005, https://www.arbylon.net/publications/text-est2.pdf.