

# Collapsing the Correlated Topic Model

Arun Tejasvi Chaganty, Kirtika Ruchandani, Balaraman Ravindran

Department of Computer Science and Engineering,  
IIT Madras, Chennai, India - 600036

**Abstract.** The Correlated Topic Model (CTM) was proposed by Blei et. al [BL07] to capture topic correlations. It does so by assuming the topics are drawn from a log-normal distribution. The log-normal and multinomial distributions are not conjugate, thus leading to complex inference and learning algorithms. We outline several approximations we explored to ‘collapse’ the CTM to a simpler form that can be easily parallelised using a GPU. Parallelisation is key to scaling the model to large web-scale data.

## 1 Introduction

The latent Dirichlet allocation (LDA) model proposed in [BNJ03] assumes the topic proportions for a document are independent, and can be drawn from a  $K$ -dimensional Dirichlet. Relaxing this assumption, the Correlated Topic Model (CTM) in [BL07] proposes drawing  $K - 1$  ‘topic proportions’ from a log-normal distribution,

$$\frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\log\theta - \mu)^T \Sigma^{-1}(\log\theta - \mu)\right) \frac{\theta_j}{\sum_{j'} \theta_{j'}}.$$

This distribution is neither conjugate to the multinomial, nor marginalisable in the closed form due to the normaliser in the denominator. This makes inferring using the model challenging, and the standard approach for collapsed inference [TNWI06] inapplicable.

At the same time, several methods to parallelise the inference of the simpler LDA model have been proposed. Yan et. al propose an algorithm to split data for inference on graphics processing units (GPUs) [YXQ]. Smola and Narayana-murthy propose a parallel architecture to process thousands of documents per hour [SN10]. Our objective is to achieve the same with this non-conjugate model.

We describe the notation we will use, and provide some background on the correlated topic model in [Section 2](#). Our first approach involved approximations to the integral ([Section 3](#)). Next, we tried a variation of the model, replacing the log-normal distribution with a normal and Dirichlet ([Section 4](#)). Finally, we describe an approach wherein we approximate the marginal using the MAP value of the distribution ([Section 5](#)). Partial experimental results are described within the sections. We briefly describe our observations and future work in [Section ??](#).

## 2 Background

### 2.1 Notation

In light of the variety of notation used across topic modelling literature, we define here a uniform notation we will adhere to:

– **Indices:**

- $i$ : Document index
- $j$ : Topic index
- $k$ : Vocabulary index
- $w$ : Word index (within a document)
- When multiple indices of the same type are required, we will use  $i'$ ,  $j'$ , etc.
- When not obvious, we will use  $\cdot$  to indicate an index we have marginalised over. For example,  $\theta_{\cdot j}$  refers to the  $j$ -th topic marginalised over all the documents.

– **Counts:**

- $D$ : Number of documents
- $D_i$ : Number of words in document  $i$
- $K$ : Number of topics
- $V$ : Number of distinct vocabulary terms
- $n_{i,k}$ : Number of occurrences of the  $k$ -th term in the  $i$ -th document
- $n_{ij}$ : Number of occurrences of the  $j$ -th topic in the  $i$ -th document

– **Model Hyper-parameters:**

- $\alpha$ : Dirichlet parameters for topic proportions in the LDA
- $\mu$ : Mean parameter for the log-normal distribution in the CTM
- $\Sigma$ : Covariance parameter for the log-normal distribution in the CTM
- $\beta$ : Dirichlet parameters for word proportions in the LDA

– **Model Parameters:**

- $\theta_{ij}$ : Probability of drawing topic  $j$  in document  $i$
- $\eta_{ij}$ : Log-likelihood of drawing topic  $j$  in document  $i$ , i.e.  $\theta_{ij} = \frac{\exp(\eta_{ij})}{\sum_{j'=0}^K \exp \eta_{ij'}}$ .
- $\omega_{jk}$ : Probability of drawing term  $k$  from topic  $j$
- $z_i$ : Topic assignment vector in document  $i$
- $w_{iw}$ :  $w$ -th word in document  $i$
- $v_k$ :  $k$ -th vocabulary term

### 2.2 Topic Models

A topic in the context of recent NLP literature is simply a bag-of-words, a distribution over the probability of occurrences of vocabulary, usually a Dirichlet. Thus, in a sports topic, the probability of picking up a word like “coach”, or “strategy” might be higher than in a music topic.

Most topic models can be described by the following generative procedure,

1. Draw the topic-word distribution,  $\omega_j \sim p(\omega_j | \beta_j)$ . This is usually a Dirichlet distribution.

2. For each document  $i$ , sample the topic proportions  $\eta_i$ .
3. Draw the document length  $W$ .
4. For each word,
  - (a) Draw a topic from the multinomial distribution  $P(z_{ij}|\eta_i)$ .
  - (b) Draw a word from the multinomial distribution  $P(v_{ik}|\omega_j)$ .

### 2.3 Latent Dirichlet Allocation

The LDA model [BNJ03] models both the topic proportion distribution, and the topic-word distribution with Dirichlets,  $\alpha$  and  $\beta$  respectively. The conjugacy between the Dirichlet and multinomial distributions allows the model to be inferred using an efficient Gibbs sampler, as well a expectation-maximisation (EM) approach using variational Bayes updates. The variational updates are rather simple to compute.  $\beta$  can be updated analytically in each maximisation step, and  $\alpha$  can be estimated efficiently using a linear time Newton-Raphson algorithm. The latter is possible because the topics are uncorrelated, leading to a diagonal Hessian matrix.

These inference procedures can be further improved using mean-field approximations, i.e. collapsed Gibbs sampling or collapsed variational Bayes updates, which will be described in a later subsection.

### 2.4 Correlated Topic Model

The CTM proposes using a log-normal distribution for the topic proportions, capturing the topic correlations in the covariance matrix. Each topic has a Dirichlet distribution words are drawn from, represented by  $\beta_j$ . Briefly,

$$\begin{aligned}
 p(\eta, z, w, \omega | \mu, \Sigma, \beta) &= p(\omega | \beta) \prod_i^D p(\eta_i | \mu, \Sigma) \prod_w^{D_i} p(z_{iw} = j | \eta_i) p(w_{iw} = v_k | z_{ij}) \\
 &= p(\omega | \beta) \prod_i^D p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \\
 p(\eta_i | \mu, \Sigma) &= \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\eta_i - \mu)^T \Sigma^{-1}(\eta_i - \mu)\right) \\
 p(z_{ij} | \eta_i) &= \frac{\exp(\eta_{ij})}{\sum_{j'=0}^K \exp \eta_{ij'}} \\
 p(v_k | z_{ij}) &= \omega_{jk} \\
 p(\omega_j | \beta_j) &= \frac{\Gamma\left(\sum_k^V \beta_{jk}\right)}{\prod_k^V \Gamma(\beta_{jk})} \prod_k^V \omega_{jk}^{\beta_{jk}-1}.
 \end{aligned}$$

The normal distribution above and the multinomial distribution from which topics and words are drawn from is non-conjugate, thus requiring Metropolis-Hastings in order to perform MCMC inference. The preferred alternative is variational inference, wherein the KL-divergence of a fully factorised model with

respect to the above model is minimised. The variational model is,

$$\begin{aligned}
q\left(\eta, z, \omega | \lambda, \nu, \phi, \tilde{\beta}\right) &= \prod_{ij} q(z_{ij} | \phi_i) \prod_j^K q(\eta_j | \lambda_j, \nu_j) \prod_k q(v_k | \tilde{\beta}_k) \\
q(\eta_j, | \lambda_j, \nu_j) &= \frac{1}{|2\pi\nu_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\eta_j - \lambda_j)^T \nu_j^{-1} (\eta_j - \lambda_j)\right) \\
q(z_{iw} = j | \phi_i) &= \phi_{ijw} \\
q(v_k | \tilde{\beta}_k) &= \omega_k \\
q(\omega_j | \beta_j) &= \frac{\Gamma\left(\sum_k^V \tilde{\beta}_{jk}\right)}{\prod_k^V \Gamma\left(\tilde{\beta}_{jk}\right)} \prod_k^V \omega_{jk}^{\tilde{\beta}_{jk}-1}.
\end{aligned}$$

The variational updates are far more involved here, requiring iterations between each of the variational parameters  $z, \lambda, \nu$ , and an additional parameter  $\zeta$ . Further, the parameters  $\mu$  and  $\Sigma$  require coordinate ascent optimisation to be estimated. Details of inference routine can be found in the appendix of the paper by Blei et. al [BL07].

## 2.5 Collapsed Gibbs Sampling and the Collapsed Variational Bayes

Collapsed Gibbs sampling and collapsed variational bayes inference are based on a mean-field approximation of the distribution [TNWI06]. In this approach, hidden parameters like  $\theta$  and  $\omega$  are marginalised before proceeding with inference; this is straight-forward when  $p(\theta|\cdot)$  and  $p(\omega|\cdot)$  are conjugate with  $p(z|\theta)$  and  $p(v|\omega)$ . This is the case for LDA, leading to the following updates,

$$p(z_{iw} = j | z^{-iww}, w, \alpha, \beta) \propto (\alpha_j + n_{ij}^{-iww}) (\beta_{jk} + n_{.jk}^{-iww}) \left( \sum_{k'} \beta_{jk'} + n_{.jk'}^{-iww} \right)^{-1}$$

and

$$\begin{aligned}
\phi_{ijk} &\propto (\alpha_j + \mathbb{E}[n_{ij}^{-iww}]) (\beta_{jk} + \mathbb{E}[n_{.jk}^{-iww}]) \left( \mathbb{E} \left[ \sum_{k'} \beta_{jk'} + n_{.jk'}^{-iww} \right] \right)^{-1} \\
&\exp \left( -\frac{\text{Var}[n_{ij}^{-iww}]}{2(\alpha_j + \mathbb{E}[n_{ij}^{-iww}])^2} - \frac{\text{Var}[n_{.jk}^{-iww}]}{2(\beta_{jk} + \mathbb{E}[n_{.jk}^{-iww}])^2} + \frac{\sum_{k'} \text{Var}[n_{.jk'}^{-iww}]}{2(\sum_{k'} \beta_{jk'} + \mathbb{E}[n_{.jk'}^{-iww}])^2} \right).
\end{aligned}$$

Computing the mean field approximation of the CTM is difficult because of the normalising factor,  $\sum_{j'} \exp(\eta_{j'})$ . In subsequent sections, we will describe our various attempts at working around this factor.

### 3 Approach 1: Approximating the Integral

The first step in deriving the CVB form for the CTM is marginalising over the hidden parameters  $\eta$  and  $\omega$ ,

$$\begin{aligned}
 p(\eta, z, w, \omega | \mu, \Sigma, \beta) &= p(\omega | \beta) \left( \prod_i^D p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \right) \left( \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \right) \\
 p(z, w | \mu, \Sigma, \beta) &= \int d\eta \, d\beta \, p(\eta, z, w, \omega | \mu, \Sigma, \beta) \\
 &= \left( \prod_i^D \int d\eta_i \, p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \right) \\
 &\quad \int d\omega \, p(\omega | \beta) \left( \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \right)
 \end{aligned}$$

The second part of this integral simply evaluates to  $\text{Dir}(\beta+n)$ , because of the conjugacy of the Dirichlet and multinomial distributions. The first part however, does not reduce to anything so simple,

$$\begin{aligned}
 \int d\eta_i \, p(\eta_i | \mu, \Sigma) \prod_j^K p(z_j | \eta_i)^{n_{ij}} &= \int d\eta_i \, \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\eta_i - \mu)^T \Sigma^{-1}(\eta_i - \mu)\right) \\
 &\quad \prod_j^K \frac{\exp \eta_j^{n_{ij}}}{\sum_{j'} \exp \eta_{j'}^{n_{ij'}}}.
 \end{aligned}$$

This integral can not be further reduced because of the denominator. One approach to proceed is by constraining the denominator to another parameter,  $\gamma$ <sup>1</sup>. Note that because we are going to marginalise out  $\gamma$ , we instead place a constraint on the *expected* value of the denominator.

$$\begin{aligned}
 \gamma &= \sum_j^K \exp \eta_j \\
 \mathbb{E}[\gamma] &= \sum_j^K \exp\left(\mu_j + \frac{1}{2}\Sigma_{jj}\right)
 \end{aligned}$$

This leads to the following variational updates,

<sup>1</sup> Blei et. al. also resort to bounding the denominator, but do so when optimising the Lagrangian

$$\begin{aligned}
\log(\phi_{ijk}) &\leftarrow \left( \frac{1}{2} \sum_{j'} \Sigma_{jj'} \left( \mathbf{E}_j^{-ik} [n_{ij'.}] \right) + \frac{1}{2} \Sigma_{jj} \left( \mathbf{E}_j^{-ik} [n_{ij'.}] + n_{i.k} \right) + \mu_j \right) \\
&+ \frac{\sum_{m=0}^{n_{i.k}-1} -\log \left( \beta_j + \mathbf{E}_j^{-ik} [n_{.j.}] + m \right) + \frac{\text{Var}_j^{-ik} [n_{.j.}]}{2(\beta_j + \mathbf{E}_j^{-ik} [n_{.j.}] + m)^2}}{n_{i.k}} \\
&+ \frac{\sum_{m=0}^{n_{i.k}-1} \log \left( \beta_{jk} + \mathbf{E}_{jk}^{-ik} [n_{.jk}] + m \right) - \frac{\text{Var}_j^{-ik} [n_{.jk}]}{2(\beta_{jk} + \mathbf{E}_{jk}^{-ik} [n_{.jk}] + m)^2}}{n_{i.k}} \\
&- \log \left( \sum_k \phi_{ijk} \right) \\
&\approx \left( \frac{1}{2} \sum_{j'} \Sigma_{jj'} \left( \mathbf{E}_j^{-ik} [n_{ij'.}] \right) + \frac{1}{2} \Sigma_{jj} \left( \mathbf{E}_j^{-ik} [n_{ij'.}] + n_{i.k} \right) + \mu_j \right) \\
&+ -\log \left( \beta_j + \mathbf{E}_j^{-ik} [n_{.j.}] + \frac{n_{i.k} - 1}{2} \right) + \frac{\text{Var}_j^{-ik} [n_{.j.}]}{2 \left( \beta_j + \mathbf{E}_j^{-ik} [n_{.j.}] + \frac{n_{i.k} - 1}{2} \right)^2} \\
&+ \log \left( \beta_{jk} + \mathbf{E}_{jk}^{-ik} [n_{.jk}] + \frac{n_{i.k} - 1}{2} \right) - \frac{\text{Var}_j^{-ik} [n_{.jk}]}{2 \left( \beta_{jk} + \mathbf{E}_{jk}^{-ik} [n_{.jk}] + \frac{n_{i.k} - 1}{2} \right)^2} \\
&- \log \left( \sum_k \phi_{ijk} \right).
\end{aligned}$$

A detailed derivation, along with parameter updates can be found in [Section A](#).

Another approach could be to consider a second order Taylor expansion of the expectation of  $\frac{\prod_j \exp \eta_j^{n_{.j.}}}{(\sum_{j'} \exp n_{ij'.})^n}$ ,

$$\begin{aligned}
p(z, w | \mu, \Sigma, \beta) &= \int d\eta \, d\beta \, p(\eta, z, w, \omega | \mu, \Sigma, \beta) \\
&= \mathbf{E}_\eta [\mathbf{E}_\omega [p(\eta, z, w, \omega | \mu, \Sigma, \beta)]] \\
&= \mathbf{E}_\eta \left[ \frac{\exp \eta_j^{n_{ij.}}}{(\sum_j \exp \eta_j)^N} \right] \mathbf{E}_\omega \left[ \prod_{jk} n_{.jk} \right] \\
&= \mathbf{E}_\eta \left[ \frac{\prod_j \exp \eta_j^{n_{.j.}}}{(\sum_j \exp \eta_j)^N} \right] \prod_j \frac{\text{Dir}(\beta_j)}{\text{Dir}(\beta_j + n_{.j.})} \\
&= \mathbf{E}_\theta \left[ \frac{\prod_j \theta_j^{n_{.j.}}}{(\sum_{j'} \theta_{j'})^N} \right] \prod_j \frac{\text{Dir}(\beta_j)}{\text{Dir}(\beta_j + n_{.j.})}
\end{aligned}$$

This approach leads to the following updates to  $\phi$ ,

$$\begin{aligned}
\log(\phi_{ijk}) &\approx (\mu_j + \frac{1}{2}\Sigma_{jj} + \sum_{j'} \frac{1}{2}(\exp(\Sigma_{jj'}) - 1)[E_j^{-ik}[n_{.j'.}] - \delta_{jj'} - N_i \frac{\exp(\mu_{j'} + \frac{1}{2}\Sigma_{j'j'})}{\sum_{j''} \exp(\mu_{j''} + \frac{1}{2}\Sigma_{j''j''})}]) \\
&+ -\log(\beta_j + E_j^{-ik}[n_{.j.}] + \frac{n_{i.k} - 1}{2}) + \frac{\text{Var}_j^{-ik}[n_{.j.}]}{2(\beta_j + E_j^{-ik}[n_{.j.}] + \frac{n_{i.k} - 1}{2})^2} \\
&+ \log(\beta_{jk} + E_{jk}^{-ik}[n_{.jk}] + \frac{n_{i.k} - 1}{2}) - \frac{\text{Var}_j^{-ik}[n_{.jk}]}{2(\beta_{jk} + E_{jk}^{-ik}[n_{.jk}] + \frac{n_{i.k} - 1}{2})^2} \\
&- \log(\sum_k \phi_{ijk})
\end{aligned}$$

The updates are somewhat intuitive. The first two terms are the expected values of  $\phi_{ijk}$ , given just  $\mu$ ,  $\Sigma$ . The remaining terms project the “difference” between the observed values  $E_j^{-ik}[n_{.j'.}]$  and the expected values  $N_i \frac{\alpha_{j'}}{A}$  (noting that  $\frac{\alpha_{j'}}{A} = \theta_{j'}$ , the proportion of  $j'$  in the document) onto  $j$  using the covariance matrix.

## 4 Approach 2: Substituting the Log-Normal

In this approach, we modify the original distribution to use the log-normal distribution as a *prior for a Dirichlet over topic distributions*. This is identical to the LDA model, with a log-normal prior, which still provides a mechanism to capture correlations between the topic proportions in a document.

$$\begin{aligned}
p(w|\mu, \Sigma, \beta) &= \int d\alpha \sum_z p(z, w, \alpha|\mu, \Sigma, \beta) \\
&= \int d\alpha \sum_z p(\alpha|\mu, \Sigma) p(z|\alpha) p(w|\beta) \\
p(\alpha|\mu, \Sigma) &= \frac{|2\pi\Sigma|^{-\frac{1}{2}}}{\prod_j \theta_j} \exp\left(-\frac{1}{2}(\log(\theta) - \mu)^T \Sigma^{-1} (\log(\theta) - \mu)\right) \\
p(z|\alpha) &= \frac{\Gamma(\sum_j \alpha_j)}{\Gamma(\sum_j \alpha_j + n_j)} \prod_j \frac{\Gamma(\alpha_j + n_j)}{\Gamma(\alpha_j)} \\
p(w|\beta) &= \prod_j \frac{\Gamma(\sum_k \beta_{jk})}{\Gamma(\sum_k \beta_{jk} + n_{jk})} \prod_j \prod_k \frac{\Gamma(\beta_{jk} + n_{jk})}{\Gamma(\beta_{jk})}
\end{aligned}$$

The variational updates are,

$$\begin{aligned}
\log(\phi_{ijk}) = & -\log\left(\alpha_i + \mathbb{E}^{-ik}[n_i] + \frac{n_{ik} - 1}{2}\right) + \frac{\text{Var}^{-ik}[n_i]}{2\left(\alpha_i + \mathbb{E}^{-ik}[n_i] + \frac{n_{ik} - 1}{2}\right)^2} \\
& + \log\left(\alpha_{ij} + \mathbb{E}^{-ik}[n_{ij}] + \frac{n_{ik} - 1}{2}\right) - \frac{\text{Var}^{-ik}[n_{ij}]}{2\left(\alpha_{ij} + \mathbb{E}^{-ik}[n_{ij}] + \frac{n_{ik} - 1}{2}\right)^2} \\
& + -\log\left(\beta_j + \mathbb{E}^{-ik}[n_j] + \frac{n_{ik} - 1}{2}\right) + \frac{\text{Var}^{-ik}[n_j]}{2\left(\beta_j + \mathbb{E}^{-ik}[n_j] + \frac{n_{ik} - 1}{2}\right)^2} \\
& + \log\left(\beta_{jk} + \mathbb{E}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2}\right) - \frac{\text{Var}^{-ik}[n_{jk}]}{2\left(\beta_{jk} + \mathbb{E}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2}\right)^2} \\
& - \log\left(\sum_k \phi_{ijk}\right)
\end{aligned}$$

We evaluated our model on 180 randomly selected documents from the AP corpus. The corpus was pre-processed by removing stopwords and lemmatising, with a final vocabulary of 9586 words across 2246 documents.

The model was run starting with a  $\mu$  of 0 and  $\Sigma$  with an uncorrelated unit-variance each, and a random initialisation to the topic-vocabulary parameters  $\beta$ <sup>2</sup>. The expectation and maximisation steps were run till convergence of 0.001% and 0.01% respectively. This gave log-likelihood score of  $-5.401326e + 07$ .

In comparison Blei, et. al's implementation of CTM [BL07], with similar parameter settings, gave a log-likelihood score of  $-2.27487e + 05$

### 5 Approach 3: Using MAP in place of the marginal

A typical Gibbs sampler would proceed as follows,

1. Begin with a random initialisation of  $z$ .
2. Choose a word  $i \in [1, \dots, N_d]$  at random. Choose a topic for it according to the distribution,  $P(z^i = j | z^{-i}, w, N, \beta)$ ,

$$P(z^i = j | z^{-i}, w, N, \beta) = \frac{P(z^i = j, z^{-i}, w, N, \beta)}{\sum_{j'} P(z^i = j', z^{-i}, w, N, \beta)}.$$

3. Continue until the Markov chain converges to the stationary distribution.

The key step in this approach is in computing  $P(z^i = j, z^{-i}, w, N, \beta)$ ,

$$P(z^i = j, z^{-i}, N, \beta) = \int d\eta d\omega P(\omega|\beta)P(\eta|N)P(z^i = j|\eta)P(w^i|z^i, \omega) \prod_{i' \neq i}^N P(z^{i'}|\eta)P(w^{i'}|z^{i'}, \omega).$$

<sup>2</sup> Uniform initialisation of the topics results in indistinguishable topics, and is thus not a sensible option

The conjugacy properties of the Dirichlet and the multinomial, let us integrate out that part of the equation,

$$\begin{aligned} \int d\omega P(\omega|\beta) \prod_i^N P(w^i|z^i = j, \omega) &= \int d\omega \text{Dir}(\omega|\beta) \prod_j^K \prod_k^V \omega_{jk}^{n_{jk}} \\ &= \prod_j^K \frac{\Gamma(\sum_k^V \beta_{jk})}{\Gamma(\sum_k^V \beta_{jk} + n_{jk})} \prod_k^V \frac{\Gamma(\beta_{jk})}{\Gamma(\beta_{jk} + n_{jk})}. \end{aligned}$$

On the other hand, not every member of the exponential family is a conjugate of the multinomial distribution. We propose using the MAP estimate of  $\eta$  instead of evaluating under expectation.

$$P(z|N) \approx \max_{\eta} P(\eta|N)P(z|\eta) \quad (1)$$

$$\log(P(\eta|N)P(z|\eta)) = \log Z(N) + N^T \mathbf{T}(\eta) + n^T \eta - N \log \sum_j \exp(\eta_j)$$

$$\frac{\partial}{\partial \eta_i} \log(P(\eta|N)P(z|\eta)) = N^T \frac{\partial}{\partial \eta_i} \mathbf{T}(\eta) + n_i - N \frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)}.$$

Thus, for the MAP estimate of  $\eta$ ,

$$\begin{aligned} \frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)} &= \frac{n_i}{N} + \frac{N^T \frac{\partial}{\partial \eta_i} \mathbf{T}(\eta)}{N} \\ N^T \sum_i \frac{\partial}{\partial \eta_i} \mathbf{T}(\eta) &= 0 \end{aligned}$$

This is condition basically allows  $\frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)}$  to be larger than the empirical topic proportion,  $\frac{n_i}{N}$  for a particular  $i$ , as long as overall, these additions cancel out. As a special case that certainly satisfies these constraints, consider  $\frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)} = \frac{n_i}{N}$  for all  $i$ . Then,  $\eta = \log n + \log K - \log N$ . Finally, we can solve for  $K$ , with the over-constrained system,  $N^T \frac{\partial}{\partial \eta} \mathbf{T}(\eta) = 0$ .

Putting in these values of  $\eta$  into Equation (1),

$$\log P(z|N) = \log Z(N) + N^T \mathbf{T}(\log n + \log \frac{K}{N}) + n^T \log n + N \log \frac{K}{N} - n.$$

Let us consider the special case where  $P_N$  is the multivariate Gaussian distribution, with mean  $\mu$  and covariance  $\Sigma$ . In this case,

$$\begin{aligned} T(\theta) &= \langle \theta_1, \theta_2, \dots, \theta_K, \theta_{11}, \theta_{12}, \dots, \theta_{KK} \rangle \\ N &= \langle -\Sigma_{11}^{-1} \mu_1, -\Sigma_{22}^{-1} \mu_2, \dots, -\Sigma_{KK}^{-1} \mu_K, \Sigma_{11}^{-1}, \Sigma_{12}^{-1}, \dots, \Sigma_{KK}^{-1} \rangle. \end{aligned}$$

We can compute  $\log K$  in this case,

$$\begin{aligned} N^T \frac{\partial}{\partial \eta_i} \mathbb{T}(\eta) &= 0 \\ -\Sigma_{ii}^{-1} \mu_i + \sum_j \Sigma_{ij}^{-1} \eta_j &= 0 \\ -\Sigma_{ii}^{-1} \mu_i + \sum_j \Sigma_{ij}^{-1} (\log n_j + \log \frac{K}{N}) &= 0. \end{aligned}$$

Thus,

$$\log \frac{K}{N} = \frac{\Sigma_{ii}^{-1} \mu_i - \sum_j \Sigma_{ij}^{-1} \log n_j}{\sum_j \Sigma_{ij}^{-1}}.$$

The Gibbs sampler really only requires,  $\frac{P(z^i=j, z^{-i}, w, N, \beta)}{\sum_{j'} P(z^i=j', z^{-i}, w, N, \beta)}$ , thus we can drop constant terms. With the topic-word distribution, we get,

$$\frac{P(w|z^i=j, z^{-i}, \beta)}{\sum_{j'} P(w|z^i=j', z^{-i}, \beta)} = \frac{(\beta_{jk} + n_{jk}^{-i})(\sum_k \beta_{jk} + n_{jk}^{-i})^{-1}}{\sum_{j'} (\beta_{j'k} + n_{j'k}^{-i})(\sum_k \beta_{j'k} + n_{j'k}^{-i})^{-1}}.$$

If we wish to have a similar relation with  $P(z|N)$ , we will need to make use of the relation,

$$\log(n+1) = \log n + 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left( \frac{1}{2n+1} \right)^{2k+1} = \log n + \zeta,$$

when  $n$  is an integer. Using this in the log-likelihood equation,

$$\begin{aligned} \log \frac{K}{N} &= \frac{\Sigma_{jj}^{-1} \mu_j - \sum_{j'} \Sigma_{jj'}^{-1} \log n_{j'}}{\sum_{j'} \Sigma_{jj'}^{-1}} \\ &= \frac{\Sigma_{jj}^{-1} \mu_j - \sum_{j'} \Sigma_{jj'}^{-1} \log n_{j'}^{-i} - \Sigma_{jj}^{-1} \zeta_j}{\sum_{j'} \Sigma_{jj'}^{-1}} \\ &= \log \frac{K^{-i}}{N} - \frac{\Sigma_{jj}^{-1}}{\sum_{j'} \Sigma_{jj'}^{-1}} \zeta_j. \end{aligned}$$

Finally, the sufficient statistic function  $T$  can be decomposed as follows,

$$\begin{aligned} N^T T(x+y) &= \sum_j \Sigma_{jj}^{-1} \mu_j (x+y)_j + \sum_j \sum_{j'} \Sigma_{jj'}^{-1} (x+y)_j (x+y)_{j'} \\ &= \sum_j \Sigma_{jj}^{-1} \mu_j x_j + \sum_j \Sigma_{jj}^{-1} \mu_j y_j + \sum_j \sum_{j'} \Sigma_{jj'}^{-1} (x_j x_{j'} + x_j y_{j'} + y_j x_{j'}) + y_j y_{j'} \\ &= N^T T(x) + N^T T(y) + \sum_j \sum_{j'} 2 \Sigma_{jj'}^{-1} x_j y_{j'} \end{aligned}$$

In particular, if  $y$  is non-zero for exactly one index,  $j$ , the last term reduces to  $2y_j \sum_{j'} \Sigma_{jj'}^{-1} x_{j'}$ . Putting these together, we can write  $P(z|N)$  in a reduced form,

$$\begin{aligned} \log P(z|N) &= \log Z(N) + N^T \text{T} \left( \log \frac{n^{-i} K^{-i}}{N} + \left(1_j - \frac{\Sigma_{jj}^{-1}}{\sum_{j'} \Sigma_{jj'}^{-1}}\right) \zeta_j \right) + \sum_{j'} \log \Gamma(n_{j'}^{-i} + \delta_{jj'}) + N \log \frac{K}{N} \\ &\equiv K + N^T \text{T} \left( \left(1_j - \frac{\Sigma_{jj}^{-1}}{\sum_{j'} \Sigma_{jj'}^{-1}}\right) \zeta_j \right) + 2\zeta_j \left( \Sigma_j^{-1} - \frac{\Sigma_{jj}^{-1}}{\sum_{j'} \Sigma_{jj'}^{-1}} \sum_{j''} \Sigma_{j''}^{-1} \right)^T \log \frac{n^{-i} K^{-i}}{N} + \log n_j^{-i} \end{aligned}$$

Combining this with the topic-word distribution, we have a selection criteria for the Gibbs sampler.

## References

- BL07. David M. Blei and John D. Lafferty. A Correlated Topic Model of Science. *Annals of Applied Statistics*, 1(1):17–35, June 2007.
- BNJ03. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, May 2003.
- SN10. Alexander Smola and Shравan Narayanamurthy. An Architecture for Parallel Topic Models. In *VLDB*, pages 703–710, 2010.
- TNWI06. Yee Whye Teh, David Newman, Max Welling, and U C Irvine. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *NIPS*, 2006.
- YXQ. Feng Yan, Ningyi Xu, and Yuan Qi. Parallel Inference for Latent Dirichlet Allocation on Graphics Processing Units. In *NIPS*, pages 1–9.

## A Calculations for Normaliser Approximation of the CTM

### A.1 Marginalising $p$

In this section, we marginalise  $p(\eta, z, w, \omega | \mu, \Sigma, \beta)$  over  $\eta$  and  $\omega$ .

$$\begin{aligned} p(\eta, z, w, \omega | \mu, \Sigma, \beta) &= p(\omega | \beta) \left( \prod_i^D p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \right) \left( \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \right) \\ p(z, w | \mu, \Sigma, \beta) &= \int d\eta \, d\beta \, p(\eta, z, w, \omega | \mu, \Sigma, \beta) \\ &= \int d\eta \left( \prod_i^D p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \right) \int d\omega \, p(\omega | \beta) \left( \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \right) \end{aligned}$$

Note that in (2), the second integral computes the posterior of a Dirichlet, and is thus simply  $\text{Dir}(\beta + n)$ , i.e. another Dirichlet with updated pseudo-counts.

It is the first integral that requires attention. Also, note that the first integral is *document-specific* (whereas the second is across documents). Let us look at the integral for any particular document.

$$\begin{aligned}
\int d\eta p(\eta|\mu, \Sigma) \prod_j^K p(z_j|\eta)^{n_j} &= \int d\eta \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\eta - \mu)^T \Sigma^{-1}(\eta - \mu)\right) \prod_j^K \frac{\exp \eta_j^{n_j}}{\gamma} \\
&= \int d\eta \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \exp\left(-\frac{1}{2}(\eta - \mu)^T \Sigma^{-1}(\eta - \mu) + \sum_j n_j \eta_j\right) \\
&= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \int d\eta \exp\left(-\frac{1}{2}\eta^T \Sigma^{-1} \eta + n^T (\eta + \mu)\right) \\
&= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \int d\eta \exp\left(-\frac{1}{2}\eta^T \Sigma^{-1} \eta + n^T \eta + n^T \mu\right) \\
&= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \exp(n^T \mu) \int d\eta \exp\left(-\frac{1}{2}\eta^T \Sigma^{-1} \eta + n^T \eta\right) \quad (3)
\end{aligned}$$

It turns out that there is a very simple formula for a Gaussian integral of this form,

$$\int d\eta \exp\left(-\frac{1}{2}x^T A x + b^T x\right) = (2\pi)^{\frac{K}{2}} |A|^{-\frac{1}{2}} \exp\left(\frac{1}{2}b^T A^{-1}b\right) \quad (4)$$

Using the result of (4) in (3), we arrive upon the following,

$$\begin{aligned}
\int d\eta p(\eta|\mu, \Sigma) \prod_j^K p(z_j|\eta)^{n_j} &= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \exp(n^T \mu) \int d\eta \exp\left(-\frac{1}{2}\eta^T \Sigma^{-1} \eta + n^T \eta\right) \\
&= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \gamma^{N+K}} \exp(n^T \mu) (2\pi)^{\frac{K}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \exp\left(\frac{1}{2}n^T \Sigma n\right) \\
&= \frac{1}{\gamma^{N+K}} \exp\left(\frac{1}{2}n^T \Sigma n + n^T \mu\right) \quad (5)
\end{aligned}$$

Note that this is not exactly a valid probability distribution, and should be normalised by the following value,

$$\int dz p(z|\mu, \Sigma) = \sum_{z=N} \frac{1}{\gamma^{N+K}} \exp\left(\frac{1}{2}n^T \Sigma n + n^T \mu\right) \quad (6)$$

We will revisit this normalisation factor (the partition function) in the next section where we use a bound on the log-likelihood

Putting (5) into the original posterior calculation, (2), we get,

$$\begin{aligned}
p(z, w | \mu, \Sigma, \beta) &= \int d\eta \left( \prod_i^D p(\eta_i | \mu, \Sigma) \prod_j^K p(z_{ij} | \eta_i)^{n_{ij}} \right) \int d\beta p(\omega | \beta) \left( \prod_j^K \prod_k^V p(v_k | z_{ij})^{n_{jk}} \right) \\
&= \prod_i^D \frac{1}{\gamma^{N_i+K}} \exp\left(\frac{1}{2} n_i^T \Sigma n_i + n_i^T \mu\right) \prod_j^K \frac{\Gamma\left(\sum_k^V \beta_{jk}\right)}{\Gamma\left(\sum_k^V \beta_{jk} + n_{jk}\right)} \prod_k^V \frac{\Gamma(\beta_{jk} + n_{jk})}{\Gamma(\beta_{jk})} \\
&= \prod_i^D \frac{1}{\gamma^{N_i+K}} \exp\left(\frac{1}{2} n_i^T \Sigma n_i + n_i^T \mu\right) \prod_j^K \frac{\Gamma(\beta_j)}{\Gamma(\beta_j + n_j)} \prod_k^V \frac{\Gamma(\beta_{jk} + n_{jk})}{\Gamma(\beta_{jk})} \quad (7)
\end{aligned}$$

## A.2 Inference and Parameter Estimation

Because of the inherent difficulty in computing the exact posterior  $p(z|w, \mu, \Sigma, \beta)$ , we approximate the distribution with a fully factorised model (which marginalising out  $\eta$  and  $\omega$  has made particularly simple),

$$q(z|\phi) = \prod_i \prod_n q(z_{in} | \phi_{in})$$

Because the order of words does not matter in this model, the product over words,  $n$ , can be rewritten as a product over the vocabulary,  $k$ .

$$q(z|\phi) = \prod_i \prod_k q(z_{ik} | \phi_{ik})$$

Our objective is to minimise the KL-divergence of these two distributions. This can be shown equivalent to maximising (8) using Jensen's inequality.

$$\begin{aligned}
\mathcal{L} &= \text{KL}(q(z|\phi) || p(z|w, \mu, \Sigma, \beta)) \\
&\equiv \mathbb{E}_q[\log(p(z, w | \mu, \Sigma, \beta))] - \mathbb{E}_q[\log(q(z|\phi))] \quad (8) \\
&= \mathbb{E}_q[\log(p(z|\mu, \Sigma))] + \mathbb{E}_q[\log(p(w|z, \beta))] - \mathbb{E}_q[\log(q(z|\phi))]
\end{aligned}$$

When looking at  $\log(p(z|\mu, \Sigma))$ , we will have to consider the partition function discussed in (6) to get the correct likelihood.

$$\begin{aligned}
\log\left(\int dz p(z|\mu, \Sigma)\right) &\geq \int dz \log(p(z|\mu, \Sigma)) \\
&= -(N+K) \log(\gamma^{N+K}) + \sum_{\Sigma z=N} \frac{1}{2} n^T \Sigma n + n^T \mu \\
&= -(N+K) \log(\gamma^{N+K}) + \sum_{\Sigma z=N} \sum_j \sum_j' \frac{1}{2} n_j \Sigma_{jj'} n_{j'} + \sum_j n_j \mu_j \\
&= -(N+K) \log(\gamma^{N+K}) + \sum_{n=0}^N \binom{N-n+K-2}{K-2} \frac{n(n+1)^2}{6} \sum_j \sum_j' \Sigma_{jj'} \\
&\quad + \sum_{n=0}^N \binom{N-n+K-1}{K-1} \frac{n(n+1)}{2} \sum_j \mu_j
\end{aligned}$$

We can ignore the  $\gamma$  term. This value is formidable to compute. Adding the additional constraints,

$$\begin{aligned}
\mathcal{L} &= \sum_i^D \mathbb{E}_q \left[ \frac{1}{2} n_i^T \Sigma n_i + n_i^T \mu \right] \\
&+ \sum_j^K \left( \mathbb{E}_q \left[ \log \left( \frac{\Gamma(\beta_j)}{\Gamma(\beta_j + n_j)} \right) \right] + \sum_k^V \mathbb{E}_q \left[ \log \left( \frac{\Gamma(\beta_{jk} + n_{jk})}{\Gamma(\beta_{jk})} \right) \right] \right) \\
&- \sum_i^D \sum_k^V \sum_j^K n_{ik} \mathbb{E}_q [\log(\phi_{ijk})] \\
&+ \sum_i^D \sum_k^V \lambda_i \left( \left( \sum_j^K \phi_{ijk} \right) - 1 \right) \\
&+ \tilde{\lambda} \left( \sum_j^K \exp \left( \mu_j + \frac{1}{2} \Sigma_{jj} \right) \right) - \gamma
\end{aligned} \tag{9}$$

Let us simplify each term of (8) one at a time.

$$\mathbb{E}_q[\log(p(z|\mu, \Sigma))]$$

$$\begin{aligned}
\mathbb{E}_q [T_1] &= \sum_i^D \mathbb{E}_q \left[ \frac{1}{2} n_i^T \Sigma n_i + n_i^T \mu \right] \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q [n_i^T \Sigma n_i] + \mathbb{E}_q [n_i]^T \mu \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q [n_{ij} \Sigma_{jj} n_{ij}] + \mathbb{E}_q [n_i]^T \mu \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q \left[ n_{ij} \Sigma_{j-j} n_{i-j} + \frac{1}{2} n_{ij} \Sigma_{jj} n_{ij} \right] + \mathbb{E}_q [n_i]^T \mu \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q [n_{ij}] \Sigma_{j-j} \mathbb{E}_q [n_{i-j}] + \frac{1}{2} \mathbb{E}_q [n_{ij} \Sigma_{jj} n_{ij}] + \mathbb{E}_q [n_i]^T \mu \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q [n_{ij}] \Sigma_{j-j} \mathbb{E}_q \left[ n_{i-j} \frac{1}{2} \right] + \frac{1}{2} \Sigma_{jj} \left( \text{Var}_q [n_{ij}] + \mathbb{E}_q [n_{ij}]^2 \right) + \mathbb{E}_q [n_i]^T \mu \\
&= \sum_i^D \frac{1}{2} \mathbb{E}_q [n_i]^T \Sigma \mathbb{E}_q [n_i] + \frac{1}{2} \text{Var}_q [n_i]^T \text{diag}(\Sigma) + \mathbb{E}_q [n_i]^T \mu
\end{aligned}$$

We abuse notation slightly, using the Einstein convention, and simplifying  $n_{ij}$  to a vector  $n_i$ .

The expectations and variations of  $n_{ij}$ ,  $n_j$  and  $n_{jk}$  are,

$$\begin{aligned} \mathbb{E}_q [n_{ij}] &= \sum_k n_{ik} \phi_{ijk} & \text{Var}_q [n_{ij}] &= \sum_n n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk}) \\ \mathbb{E}_q [n_j] &= \sum_i \sum_k n_{ik} \phi_{ijk} & \text{Var}_q [n_j] &= \sum_i \sum_k n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk}) \\ \mathbb{E}_q [n_{jk}] &= \sum_i n_{ik} \phi_{ijk} & \text{Var}_q [n_{jk}] &= \sum_i n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk}) \end{aligned}$$

As  $n_{ij}$ , etc. are the sums of a large number of Bernoulli variables, we can apply a Gaussian approximation. Because a Gaussian only has first and second moments (mean and variance respectively), this approximation lets us use a second order Taylor's expansion of a function about it's expected value, i.e.,

$$\begin{aligned} \mathbb{E} [f(x)] &= f(\mathbb{E}[x]) + \frac{1}{2} f''(\mathbb{E}[x]) \text{Var}[x] \\ \mathbb{E} [\log(\Gamma(x))] &= \log(\Gamma(\mathbb{E}[x])) + \frac{1}{2} \text{Var}[x] \Psi'(\mathbb{E}[x]) \end{aligned} \quad (10)$$

Because  $\Psi'$  is a rather complex function, we resort to Stirling's approximation for  $\Gamma(x)$ ,

$$\begin{aligned} \Gamma(x) &\approx \sqrt{\frac{2\pi}{x}} x^x e^{-x} \\ \mathbb{E} [\log(\Gamma(x))] &\approx \mathbb{E} \left[ \log(\sqrt{2\pi}) - \frac{1}{2} \log(x) + x(\log(x) - 1) \right] \\ &= \log(\sqrt{2\pi}) - \frac{1}{2} \log(\mathbb{E}[x]) + \mathbb{E}[x] (\log(\mathbb{E}[x]) - 1) \\ &\quad + \frac{1}{2} \text{Var}[x] \left( \frac{1}{2\mathbb{E}[x]^2} + \frac{1}{\mathbb{E}[x]} \right) \end{aligned} \quad (11)$$

$\mathbb{E}_q [\log(p(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta}))]$

$$\begin{aligned} \mathbb{E}_q [T_2] &= \sum_j^K \mathbb{E}_q [\log(\Gamma(\beta_j)) - \log(\Gamma(\beta_j + n_j))] + \sum_j^K \sum_k^V \mathbb{E}_q [\log(\Gamma(\beta_{jk} + n_{jk})) - \log(\Gamma(\beta_{jk}))] \\ &= \sum_j^K \log(\Gamma(\beta_j)) - \log(\Gamma(\beta_j + \mathbb{E}_q [n_j])) - \frac{1}{2} \text{Var}_q [n_j] \Psi'(\beta_j + \mathbb{E}_q [n_j]) \\ &\quad + \sum_j^K \sum_k^V \log(\Gamma(\beta_{jk})) - \log(\Gamma(\beta_{jk} + \mathbb{E}_q [n_{jk}])) + \frac{1}{2} \text{Var}_q [n_{jk}] \Psi'(\beta_{jk} + \mathbb{E}_q [n_{jk}]) \\ &\approx \sum_j^K \frac{1}{2} (\log(\beta_j + \mathbb{E}_q [n_j]) - \log(\beta_j)) - \frac{1}{2} \text{Var}_q [n_j] \left( \frac{1}{\beta_j + \mathbb{E}_q [n_j]} + \frac{1}{2(\beta_j + \mathbb{E}_q [n_j])^2} \right) \\ &\quad + \sum_j^K \sum_k^V \frac{1}{2} (\log(\beta_{jk}) - \log(\beta_{jk} + \mathbb{E}_q [n_{jk}])) + \frac{1}{2} \text{Var}_q [n_{jk}] \left( \frac{1}{\beta_{jk} + \mathbb{E}_q [n_{jk}]} + \frac{1}{2(\beta_{jk} + \mathbb{E}_q [n_{jk}])^2} \right) \end{aligned}$$

**$\phi$  Updates** We are now ready to find updates to  $\phi$ .

$$\begin{aligned}
\mathcal{L} &= E_q [\log (p(z, w))] - \sum_i^D \sum_j^K \sum_k^V n_{ik} E_q [\log (\phi_{ijk})] + \sum_i^D \sum_k^V \lambda_i \left( \left( \sum_j^K \phi_{ijk} \right) - 1 \right) \\
&= \sum_{ik} \sum_j E_{q(z_{-ik})} [\log (p(z_{-ik}, w|z_{ik} = j)) P(z_{ik} = j)] \\
&\quad - \sum_i^D \sum_j^K \sum_k^V n_{ik} E_q [\log (\phi_{ijk})] + \sum_i^D \sum_k^V \lambda_i \left( \left( \sum_j^K \phi_{ijk} \right) - 1 \right) \\
\frac{\partial}{\partial \phi_{ijk}} \mathcal{L} &= \frac{\partial}{\partial \phi_{ijk}} \phi_{ijk} E_{q(z_{-ik})} [\log (p(z_{-ik}, w|z_{ik} = j))] - n_{ik} \phi_{ijk} \log (\phi_{ijk}) + \lambda_i \phi_{ijk} \\
&= E_{q(z_{-ik})} [\log (p(z_{-ik}, w|z_{ik} = j))] - n_{ik} \log (\phi_{ijk}) + \lambda_i - n_{ik}
\end{aligned}$$

Setting to zero, we get,

$$\phi_{ijk} \propto \exp \left( \frac{E_j^{-ik} [\log (p(z_{-ik}, w|z_{ik} = j))]}{n_{ik}} \right) \quad (12)$$

We use the notation  $E_j^{-ik}$  for  $E_{q(z_{-ik}, z_{ik}=j)}$ . Note that  $E_j^{-ik}$  is the same as the  $E_q$  values we looked at earlier, but omitting the counts in the  $k$ -th vocabulary index of the  $i$ -th document. The new counts are then,

$$\begin{aligned}
E_j^{-ik} [n_{ij}] &= E_q [n_{ij}] - n_{ik} \phi_{ijk} & \text{Var}_j^{-ik} [n_{ij}] &= \text{Var}_q [n_{ij}] - n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk}) \\
E_j^{-ik} [n_j] &= E_q [n_j] - n_{ik} \phi_{ijk} & \text{Var}_j^{-ik} [n_j] &= \text{Var}_q [n_j] - n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk}) \\
E_j^{-ik} [n_{jk}] &= E_q [n_{jk}] - n_{ik} \phi_{ijk} & \text{Var}_j^{-ik} [n_{jk}] &= \text{Var}_q [n_{jk}] - n_{ik}^2 \phi_{ijk} (1 - \phi_{ijk})
\end{aligned}$$

Noting that the common terms in the expression for  $\phi_{ijk}$  will cancel when normalised, we can represent only the terms that will differ across  $\phi_{ijk}$ , namely the terms involving  $k$ -th vocabulary -  $E_j^{-ik}$  assumes that the  $k$ -th vocabulary comes from topic  $j$ . Consequently, the expected counts in all these cases are

incremented by  $n_{ik}$

$$\begin{aligned}
\mathbb{E}_j^{-ik} [T_1] &= \sum_{j' \neq j} \left( \mathbb{E}_j^{-ik} [n_{ij}] + \frac{1}{2} n_{ik} \right) \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} \left( \mathbb{E}_j^{-ik} [n_{ij}] + \frac{1}{2} n_{ik} \right) \Sigma_{jj} \left( \mathbb{E}_j^{-ik} [n_{ij}] + n_{ik} \right) + \left( \mathbb{E}_j^{-ik} [n_{ij}] \right) \mu_j \\
&\quad - \sum_{j' \neq j} \frac{1}{2} \left( \mathbb{E}_j^{-ik} [n_{ij}] \right) \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} \left( \mathbb{E}_j^{-ik} [n_{ij}] \right) \Sigma_{jj} \left( \mathbb{E}_j^{-ik} [n_{ij}] \right) + \left( \mathbb{E}_j^{-ik} [n_{ij}] \right) \mu_j \\
&\quad + \text{common terms} \\
&= \sum_{j' \neq j} \frac{1}{2} n_{ik} \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} 2 n_{ik} \Sigma_{jj} \mathbb{E}_j^{-ik} [n_{ij'}] + \frac{1}{2} n_{ik}^2 \Sigma_{jj} + n_{ik} \mu_j \\
&\quad + \text{common terms} \\
&= n_{ik} \left( \frac{1}{2} \sum_{j'} \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} \Sigma_{jj} \left( \mathbb{E}_j^{-ik} [n_{ij'}] + n_{ik} \right) + \mu_j \right) \\
&\quad + \text{common terms} \\
\mathbb{E}_j^{-ik} [T_2] &= -\log \left( \Gamma \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] + n_{ik} \right) \right) - \frac{1}{2} \text{Var}_j^{-ik} [n_j] \Psi' \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] + n_{ik} \right) \\
&\quad + \log \left( \Gamma \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] + n_{ik} \right) \right) + \frac{1}{2} \text{Var}_j^{-ik} [n_{jk}] \Psi' \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] + n_{ik} \right) \\
&\quad + \log \left( \Gamma \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] \right) \right) + \frac{1}{2} \text{Var}_j^{-ik} [n_j] \Psi' \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] \right) \\
&\quad - \log \left( \Gamma \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] \right) \right) - \frac{1}{2} \text{Var}_j^{-ik} [n_{jk}] \Psi' \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] \right) \\
&\quad + \text{common terms} \\
&= \sum_{m=0}^{n_{ik}-1} -\log \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] + m \right) + \frac{\text{Var}_j^{-ik} [n_j]}{2 \left( \beta_j + \mathbb{E}_j^{-ik} [n_j] + m \right)^2} \\
&\quad + \sum_{m=0}^{n_{ik}-1} \log \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] + m \right) - \frac{\text{Var}_j^{-ik} [n_{jk}]}{2 \left( \beta_{jk} + \mathbb{E}_j^{-ik} [n_{jk}] + m \right)^2} \\
&\quad + \text{common terms}
\end{aligned}$$

Plugging these results into (12), and approximating the summation with the midpoint value.

$$\begin{aligned}
\log(\phi_{ijk}) &= \left( \frac{1}{2} \sum_{j'} \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} \Sigma_{jj} \left( \mathbb{E}_j^{-ik} [n_{ij'}] + n_{ik} \right) + \mu_j \right) \\
&\quad + \frac{\sum_{m=0}^{n_{ik}-1} -\log\left(\beta_j + \mathbb{E}_j^{-ik} [n_j] + m\right) + \frac{\text{Var}_j^{-ik} [n_j]}{2(\beta_j + \mathbb{E}_j^{-ik} [n_j] + m)^2}}{n_{ik}} \\
&\quad + \frac{\sum_{m=0}^{n_{ik}-1} \log\left(\beta_{jk} + \mathbb{E}_{jk}^{-ik} [n_{jk}] + m\right) - \frac{\text{Var}_j^{-ik} [n_{jk}]}{2(\beta_{jk} + \mathbb{E}_{jk}^{-ik} [n_{jk}] + m)^2}}{n_{ik}} \\
&\quad - \log\left(\sum_k \phi_{ijk}\right) \\
&\approx \left( \frac{1}{2} \sum_{j'} \Sigma_{jj'} \left( \mathbb{E}_j^{-ik} [n_{ij'}] \right) + \frac{1}{2} \Sigma_{jj} \left( \mathbb{E}_j^{-ik} [n_{ij'}] + n_{ik} \right) + \mu_j \right) \\
&\quad + -\log\left(\beta_j + \mathbb{E}_j^{-ik} [n_j] + \frac{n_{ik}-1}{2}\right) + \frac{\text{Var}_j^{-ik} [n_j]}{2\left(\beta_j + \mathbb{E}_j^{-ik} [n_j] + \frac{n_{ik}-1}{2}\right)^2} \\
&\quad + \log\left(\beta_{jk} + \mathbb{E}_{jk}^{-ik} [n_{jk}] + \frac{n_{ik}-1}{2}\right) - \frac{\text{Var}_j^{-ik} [n_{jk}]}{2\left(\beta_{jk} + \mathbb{E}_{jk}^{-ik} [n_{jk}] + \frac{n_{ik}-1}{2}\right)^2} \\
&\quad - \log\left(\sum_k \phi_{ijk}\right)
\end{aligned}$$

**Parameter Estimation** For parameter estimation, we have to maximise with respect to  $\mu$ ,  $\Sigma$ ,  $\beta$ .

Note that  $\mu$  is *log-convex*, writing  $\mu = \log(\tilde{\mu})$ , the following becomes a convex optimisation problem,

$$\begin{aligned}
\mathcal{L}_\mu &= \sum_i^D \mathbb{E}_q [n_i]^T \mu + \tilde{\lambda} \exp\left(\mu_j + \frac{1}{2} \Sigma_{jj}\right) \\
&= \sum_i^D \mathbb{E}_q [n_i]^T \log(\tilde{\mu}) + \tilde{\lambda} \tilde{\mu}_j \exp\left(\frac{1}{2} \Sigma_{jj}\right) \\
\frac{\partial}{\partial \tilde{\mu}_j} \mathcal{L}_\mu &= \frac{\left(\sum_i^D \mathbb{E}_q [n_{ij}]\right)}{\tilde{\mu}_j} + \tilde{\lambda} \exp\left(\frac{1}{2} \Sigma_{jj}\right)
\end{aligned}$$

Setting to zero, and using the fact that  $\sum_i \sum_j \mathbb{E}_q[n_{ij}] = N$ , the total number of words, we find,

$$\begin{aligned}\tilde{\mu}_j &\propto \frac{\left(\sum_i^D \mathbb{E}_q[n_{ij}]\right)}{\exp\left(\frac{1}{2}\Sigma_{jj}\right)} \\ \tilde{\mu}_j &= \gamma \frac{\left(\sum_i^D \mathbb{E}_q[n_{ij}]\right)}{N \exp\left(\frac{1}{2}\Sigma_{jj}\right)} \\ \mu_j &= \log(\gamma) - \log(N) + \log\left(\sum_i^D \mathbb{E}_q[n_{ij}]\right) - \frac{1}{2}\Sigma_{jj}\end{aligned}$$

When trying to optimise  $\Sigma$ , we have the additional condition that  $\Sigma$  must be positive semi-definite (to be a covariance matrix), thus, we must consider the following optimisation problem,

$$\begin{aligned}\mathcal{L}_\Sigma &= \sum_i^D \frac{1}{2} \mathbb{E}_q[n_i] \Sigma \mathbb{E}_q[n_i] + \frac{1}{2} \text{Var}_q[n_i]^T \text{diag}(\Sigma) \\ \text{subject to} \quad &\sum_j^K \exp\left(\mu_j + \frac{1}{2}\Sigma_{jj}\right) - \gamma = 0 \\ &\frac{1}{2}(\Sigma_{jj} + \Sigma_{j'j'}) - \Sigma_{jj'} \geq 0\end{aligned}$$

To convert this into a standard convex optimisation problem, we consider  $\tilde{\Sigma} = 2\log(\Sigma)$ ,

$$\begin{aligned}\mathcal{L}_{\tilde{\Sigma}} &= \sum_i^D \frac{1}{2} \mathbb{E}_q[n_i] 2\log(\tilde{\Sigma}) \mathbb{E}_q[n_i] + \frac{1}{2} \text{Var}_q[n_i]^T \text{diag}\left(2\log(\tilde{\Sigma})\right) \\ \text{subject to} \quad &\sum_j^K \exp(\mu_j) \tilde{\Sigma}_{jj} - \gamma = 0 \\ &\sqrt{\tilde{\Sigma}_{jj}\tilde{\Sigma}_{j'j'}} - \tilde{\Sigma}_{jj'} \geq 0\end{aligned}$$

This problem can be efficiently solved using Newton's method. We can further show the Hessian for this problem is diagonal, and hence invertible in linear time,

$$\begin{aligned}\frac{\partial}{\partial \Sigma_{jj'}} \mathcal{L}_{\tilde{\Sigma}} &= 2 \frac{\left(\sum_i^D \frac{1}{2} \mathbb{E}_q[n_{ij}] \mathbb{E}_q[n_{ij'}] + \frac{1}{2} \text{Var}_q[n_{ij}] \delta_{jj'}\right)}{\tilde{\Sigma}_{jj'}} \\ \frac{\partial}{\partial \Sigma_{\tilde{j}\tilde{j}'}} \mathcal{L}_{\tilde{\Sigma}} &= -2\delta_{\tilde{j}'\tilde{j}} \delta_{\tilde{j}\tilde{j}} \frac{\left(\sum_i^D \frac{1}{2} \mathbb{E}_q[n_{ij}] \mathbb{E}_q[n_{ij'}] + \delta_{jj'} \frac{1}{2} \text{Var}_q[n_{ij}]\right)}{\tilde{\Sigma}_{jj'}^2}\end{aligned}$$

It is also possible to find a ML estimate for  $\mu$  and  $\Sigma$ ,

$$\begin{aligned}\mu_j &= \frac{1}{D} \sum_i \log(\gamma) + \log(\mathbb{E}_q[n_{ij}]) - \log(N_i) \\ \Sigma_{jj'} &= \frac{1}{D} \sum_i (\log(\gamma) + \log(\mathbb{E}_q[n_{ij}]) - \log(N_i) - \mu_j) (\log(\gamma) + \log(\mathbb{E}_q[n_{ij'}]) - \log(N_i) - \mu_{j'})\end{aligned}$$

Using the results presented in the original LDA paper,  $\beta$  can be updated using the expected counts,

$$\begin{aligned}\beta_{jk} &= \sum_i^D n_{ik} \phi_{ijk} \\ &= \mathbb{E}_q[n_{jk}]\end{aligned}$$

## B Calculations for Expectation Approximation of the CTM

We note that the integral computed in the earlier section really came from finding the expected value, i.e.,

$$\begin{aligned}p(z, w | \mu, \Sigma, \beta) &= \int d\eta \, d\beta \, p(\eta, z, w, \omega | \mu, \Sigma, \beta) \\ &= \mathbb{E}_\eta [\mathbb{E}_\omega [p(\eta, z, w, \omega | \mu, \Sigma, \beta)]] \\ &= \mathbb{E}_\eta \left[ \frac{\exp \eta_j^{n_j}}{(\sum_j \exp \eta_j)^N} \right] \mathbb{E}_\omega \left[ \prod_{jk} n_{jk} \right] \\ &= \mathbb{E}_\eta \left[ \frac{\prod_j \exp \eta_j^{n_j}}{(\sum_j \exp \eta_j)^N} \right] \prod_j \frac{\text{Dir}(\beta_j)}{\text{Dir}(\beta_j + n_j)} \\ &= \mathbb{E}_\theta \left[ \frac{\prod_j \theta_j^{n_j}}{(\sum_{j'} \theta_{j'})^N} \right] \prod_j \frac{\text{Dir}(\beta_j)}{\text{Dir}(\beta_j + n_j)}\end{aligned}$$

Earlier we used an approximation to find the first term. Now, we will use a second-order Taylor expansion, noting that  $\theta$  is a log-normal distribution, and replacing  $\mathbb{E}[\theta_j]$  with  $a_j$ , and  $\sum_j a_j$  with  $A$  for brevity,

$$\begin{aligned}
\mathbb{E}_\theta [\theta_j] &= \exp\left(\mu_j + \frac{1}{2}\Sigma_{jj}\right) \\
\mathbb{E}_\theta \left[ \frac{\prod_j \theta_j^{n_j}}{(\sum_j \theta_j)^N} \right] &= \frac{\prod_j \mathbb{E}[\theta_j]^{n_j}}{(\sum_j \mathbb{E}[\theta_j])^N} + \sum_{jj'} \frac{1}{2} \text{Var}(\theta_j, \theta_{j'}) \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_{j'}} \left[ \frac{\prod_j \theta_j^{n_j}}{(\sum_j \theta_j)^N} \right] \\
&= \frac{\prod_j a_j^{n_j}}{A^N} \\
&\quad + \sum_{jj'} \frac{1}{2} a_j a_{j'} (\exp(\Sigma_{jj'}) - 1) \left[ \frac{1}{a_j a_{j'}} \left( n_j n_{j'} - \delta_{jj'} n_j - \frac{N}{A} (n_j a_{j'} + n_{j'} a_j) + a_j a_{j'} \frac{N + N^2}{A^2} \right) \right] \\
&= \frac{\prod_j a_j^{n_j}}{A^N} \\
&\quad + \frac{\prod_j a_j^{n_j}}{A^N} \sum_{jj'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ n_j n_{j'} - \delta_{jj'} n_j - \frac{N}{A} (n_j a_{j'} + n_{j'} a_j) + a_j a_{j'} \frac{N + N^2}{A^2} \right] \\
&= \frac{\prod_j a_j^{n_j}}{A^N} \left( 1 + \sum_{jj'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ n_j n_{j'} - \delta_{jj'} n_j - \frac{N}{A} (n_j a_{j'} + n_{j'} a_j) + a_j a_{j'} \frac{N + N^2}{A^2} \right] \right)
\end{aligned}$$

When computing the likelihood score, we must take the log of this value. We will again use a first order Taylor expansion of  $\log(1+x)$  as  $x$  is expected to be small (**VERIFY**).

$$\begin{aligned}
\log \left( \mathbb{E}_\theta \left[ \frac{\prod_j \theta_j^{n_j}}{(\sum_j \theta_j)^N} \right] \right) &= \sum_j n_j \log(a_j) - N \log(A) \\
&\quad + \sum_{jj'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ n_j n_{j'} - \delta_{jj'} n_j - \frac{N}{A} (n_j a_{j'} + n_{j'} a_j) \right] \\
&\quad + \frac{N + N^2}{2A^2} \sum_{jj'} (\exp(\Sigma_{jj'}) - 1) a_j a_{j'}
\end{aligned}$$

Next, this expression expected over  $q$  gives us,

$$\begin{aligned}
\mathbb{E}_q[\dots] &= \sum_i \sum_j \mathbb{E}_q[n_j] \log(a_j) - N_i \log(A) \\
&+ \sum_{jj'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ \mathbb{E}_q[n_j n_{j'}] - \delta_{jj'} \mathbb{E}_q[n_j] - \frac{N_i}{A} (\mathbb{E}_q[n_j] a_{j'} + \mathbb{E}_q[n_{j'}] a_j) \right] \\
&+ \frac{N_i + N_i^2}{2A^2} \sum_{jj'} (\exp(\Sigma_{jj'}) - 1) a_j a_{j'} \\
&= \sum_i \sum_j \mathbb{E}_q[n_j] \log(a_j) - N_i \log(A) \\
&+ \sum_{jj'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ \text{Var}_q(n_j, n_{j'}) + \mathbb{E}_q[n_j] \mathbb{E}_q[n_{j'}] - \delta_{jj'} \mathbb{E}_q[n_j] - \frac{N_i}{A} (\mathbb{E}_q[n_j] a_{j'} + \mathbb{E}_q[n_{j'}] a_j) \right] \\
&+ \frac{N_i + N_i^2}{2A^2} \sum_{jj'} (\exp(\Sigma_{jj'}) - 1) a_j a_{j'}
\end{aligned}$$

Finally we find  $\mathbb{E}_j^{-ik}$ ,

$$\begin{aligned}
\mathbb{E}_j^{-ik}[\dots] &= n_{ik} \log(a_j) \\
&+ \sum_{j'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ n_{ik} \mathbb{E}_j^{-ik}[n_{j'}] - \delta_{jj'} n_{ik} - \frac{N_i}{A} (n_{ik} a_{j'}) \right] \\
&= n_{ik} \left( \log(a_j) + \sum_{j'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ \mathbb{E}_j^{-ik}[n_{j'}] - \delta_{jj'} - \frac{N_i}{A} a_{j'} \right] \right)
\end{aligned}$$

The updates to  $\phi_{ijk}$  will then be,

$$\begin{aligned}
\log(\phi_{ijk}) &\approx \left( \mu_j + \frac{1}{2} \Sigma_{jj} + \sum_{j'} \frac{1}{2} (\exp(\Sigma_{jj'}) - 1) \left[ \mathbb{E}_j^{-ik}[n_{j'}] - \delta_{jj'} - N_i \frac{\exp(\mu_{j'} + \frac{1}{2} \Sigma_{j'j'})}{\sum_{j''} \exp(\mu_{j''} + \frac{1}{2} \Sigma_{j''j''})} \right] \right) \\
&+ -\log \left( \beta_j + \mathbb{E}_j^{-ik}[n_j] + \frac{n_{ik} - 1}{2} \right) + \frac{\text{Var}_j^{-ik}[n_j]}{2 \left( \beta_j + \mathbb{E}_j^{-ik}[n_j] + \frac{n_{ik} - 1}{2} \right)^2} \\
&+ \log \left( \beta_{jk} + \mathbb{E}_{jk}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2} \right) - \frac{\text{Var}_j^{-ik}[n_{jk}]}{2 \left( \beta_{jk} + \mathbb{E}_{jk}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2} \right)^2} \\
&- \log \left( \sum_k \phi_{ijk} \right)
\end{aligned}$$

## C Calculations for Dirichlet Approximation of the CTM

We modify the original distribution to use the log-normal distribution as a *prior* for a *Dirichlet over topic distributions*. This is identical to the LDA model, with

a log-normal prior, which still provides a mechanism to capture correlations between the topic proportions in a document.

$$\begin{aligned}
p(w|\mu, \Sigma, \beta) &= \int d\alpha \sum_z p(z, w, \alpha|\mu, \Sigma, \beta) \\
&= \int d\alpha \sum_z p(\alpha|\mu, \Sigma) p(z|\alpha) p(w|\beta) \\
p(\alpha|\mu, \Sigma) &= \frac{|2\pi\Sigma|^{-\frac{1}{2}}}{\prod_j \theta_j} \exp\left(-\frac{1}{2}(\log(\theta) - \mu)^T \Sigma^{-1} (\log(\theta) - \mu)\right) \\
p(z|\alpha) &= \frac{\Gamma\left(\sum_j \alpha_j\right)}{\Gamma\left(\sum_j \alpha_j + n_j\right)} \prod_j \frac{\Gamma(\alpha_j + n_j)}{\Gamma(\alpha_j)} \\
p(w|\beta) &= \prod_j \frac{\Gamma\left(\sum_k \beta_{jk}\right)}{\Gamma\left(\sum_k \beta_{jk} + n_{jk}\right)} \prod_j \prod_k \frac{\Gamma(\beta_{jk} + n_{jk})}{\Gamma(\beta_{jk})} \\
\log(p(w|\mu, \Sigma, \beta)) &= \log \int d\alpha \sum_z p(z, w, \alpha|\mu, \Sigma, \beta) \\
&= \log \int d\alpha \sum_z \frac{p(z, w, \alpha|\mu, \Sigma, \beta) q(z, \alpha|\lambda, \nu)}{q(z, \alpha|\lambda, \nu)} \\
&\geq \log \int d\alpha \sum_z q(z, \alpha|\lambda, \nu) \log p(z, w, \alpha|\mu, \Sigma, \beta) - \log \int d\alpha \sum_z q(z, \alpha|\lambda, \nu) \log q(z, \alpha|\lambda, \nu) \\
&= \mathbb{E}_q[\log p(z, w, \alpha|\mu, \Sigma, \beta)] - \mathbb{E}_q[\log q(z, \alpha|\lambda, \nu)]
\end{aligned}$$

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_q[\log p(\alpha|\mu, \Sigma)] + \mathbb{E}_q[\log(z|\alpha)] + \mathbb{E}_q[\log(w|\beta)] - \mathbb{E}_q[\log q(\alpha|\lambda, \nu)] - \mathbb{E}_q[\log q(z)] \\
&= -\frac{1}{2} \log(|2\pi\Sigma|) - \frac{1}{2} \left( \sum_j \nu_j^2 \Sigma_{jj}^{-1} + \sum_{jj'} \Sigma_{jj'}^{-1} (\log(\alpha_j) - \mu_j) (\log(\alpha_{j'}) - \mu_{j'}) \right) - \sum_j \log(\alpha_j) \\
&\quad + \log \left( \Gamma \left( \overline{\sum_j \alpha_j} \right) \right) - \log \left( \Gamma \left( \overline{\sum_j \alpha_j + n_j} \right) \right) + \sum_j \log(\Gamma(\overline{\alpha_j + n_j})) - \log(\Gamma(\overline{\alpha_j})) \\
&\quad + \sum_j \log \left( \Gamma \left( \overline{\sum_k \beta_{jk}} \right) \right) - \log \left( \Gamma \left( \overline{\sum_k \beta_{jk} + n_{jk}} \right) \right) + \sum_j \sum_k \log(\Gamma(\overline{\beta_{jk} + n_{jk}})) - \log(\Gamma(\overline{\beta_{jk}})) \\
&\quad - \sum_j -\frac{1}{2} \log(2\pi) - \log(\nu) - \frac{1}{2} \left( \left( \frac{\alpha_j - \lambda_j}{\nu_j^2} \right)^2 \right) - \log(\alpha_j) \\
&\quad - \sum_j \sum_k \overline{n_{ik} \log(\phi_{jk})}
\end{aligned}$$

Finally,

$$\begin{aligned}
\log(\phi_{ijk}) &= -\log\left(\alpha_i + \mathbb{E}^{-ik}[n_i] + \frac{n_{ik} - 1}{2}\right) + \frac{\text{Var}^{-ik}[n_i]}{2\left(\alpha_i + \mathbb{E}^{-ik}[n_i] + \frac{n_{ik} - 1}{2}\right)^2} \\
&\quad + \log\left(\alpha_{ij} + \mathbb{E}^{-ik}[n_{ij}] + \frac{n_{ik} - 1}{2}\right) - \frac{\text{Var}^{-ik}[n_{ij}]}{2\left(\alpha_{ij} + \mathbb{E}^{-ik}[n_{ij}] + \frac{n_{ik} - 1}{2}\right)^2} \\
&\quad + -\log\left(\beta_j + \mathbb{E}^{-ik}[n_j] + \frac{n_{ik} - 1}{2}\right) + \frac{\text{Var}^{-ik}[n_j]}{2\left(\beta_j + \mathbb{E}^{-ik}[n_j] + \frac{n_{ik} - 1}{2}\right)^2} \\
&\quad + \log\left(\beta_{jk} + \mathbb{E}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2}\right) - \frac{\text{Var}^{-ik}[n_{jk}]}{2\left(\beta_{jk} + \mathbb{E}^{-ik}[n_{jk}] + \frac{n_{ik} - 1}{2}\right)^2} \\
&\quad - \log\left(\sum_k \phi_{ijk}\right)
\end{aligned}$$