Collapsed Variational Inference for LDA

BT Thomas Yeo

1 LDA

We shall follow the same notation as Blei et al. (2003). In other words, we consider full LDA model with hyperparameters $\alpha$ and $\eta$ on $\theta$ and $\beta$ respectively, where $\theta$ parameterizes $P(\text{topic} \mid \text{document})$ and $\beta$ parameterizes $P(\text{word} \mid \text{topic})$, $w_{dn}$ refers to $n$-th word of the $d$-th document. $z_{dn}$ refers to the topic that generates the $n$-th word of the $d$-th document.

1.1 Variational Inference Setup

Cost function (aim is to maximize likelihood of observed words $w$ with respect to $\alpha, \eta$):

\[
\log p(w \mid \alpha, \eta) = \log \int \int \sum_z p(\theta, \beta, z, w \mid \alpha, \eta) d\beta d\theta
\]

\[
= \log \int \int \sum_z \frac{q(\beta, \theta, z) p(\theta, \beta, z, w \mid \alpha, \eta)}{q(\beta, \theta, z)} d\beta d\theta
\]

\[
= \log \mathbb{E}_q \left( \frac{p(\theta, \beta, z, w \mid \alpha, \eta)}{q(\beta, \theta, z)} \right)
\]

\[
\geq \mathbb{E}_q \left( \log p(\theta, \beta, z, w \mid \alpha, \eta) \right) - \mathbb{E}_q \left( \log q(\beta, \theta, z) \right)
\]

\[
= -F(q(\beta, \theta, z))
\]

$F$ is called the variational free energy

Note that

\[
\mathbb{E}_q \left( \log p(\theta, \beta, z, w \mid \alpha, \eta) \right) - \mathbb{E}_q \left( \log q(\beta, \theta, z) \right) + \text{KL}(q(\beta, \theta, z) \mid \mid p(\beta, \theta, z \mid w, \alpha, \eta))
\]

\[
= \mathbb{E}_q \left( \log p(\theta, \beta, z, w \mid \alpha, \eta) \right) - \mathbb{E}_q \left( \log q(\beta, \theta, z) \right) + \mathbb{E}_q \left( \log \frac{q(\beta, \theta, z)}{p(\beta, \theta, z \mid w, \alpha, \eta)} \right)
\]

\[
= \mathbb{E}_q \left( \log \frac{p(\theta, \beta, z, w \mid \alpha, \eta)}{p(\beta, \theta, z \mid w, \alpha, \eta)} \right)
\]

\[
= \log p(w \mid \alpha, \eta)
\]

Therefore

\[
\log p(w \mid \alpha, \eta) - (-F(q(\beta, \theta, z))) = \text{KL}(q(\beta, \theta, z) \mid \mid p(\beta, \theta, z \mid w, \alpha, \eta))
\]

In other words, for the the negative free energy ($-F$) to be equal to $\log p(w \mid \alpha, \eta)$, $q(\beta, \theta, z)$ should be equal to $p(\beta, \theta, z \mid w, \alpha, \eta)$.

In EM, we compute $q(\beta, \theta, z) = p(\beta, \theta, z \mid w, \alpha, \eta)$ exactly in the $E$-step and then conditioned on this particular $q$, we maximize with respect to $\alpha$ and $\eta$ in the M-step. In LDA, this is not tractable. In variational EM, $q(\beta, \theta, z)$ is approximated with a simpler class of functions. The best $q$ function within this class of functions is optimized during the $E$-step and then conditioned on this particular $q$, we maximize with respect to $\alpha$ and $\eta$ in the M-step.

1.2 Motivation for Collapsed Variational Inference

In the original LDA paper (Blei et al., 2003), the class of proposal distributions was:

\[
q(\theta, \beta, z \mid \gamma, \phi, \lambda) = q(\beta \mid \lambda)q(\theta \mid \gamma)q(z \mid \phi)
\]

\[
= \prod_{k=1}^K q(\beta_k \mid \lambda_k) \prod_{d=1}^D q_d(\theta_d \mid \gamma_d)q(z_d \mid \phi_d)
\]

(12)

(13)
In Teh et al. (2007), the class of proposal distributions was

\[
q(\theta, z, \beta|\phi) = q(\theta, \beta|z)q(z|\phi)
\]

\[
=q(\theta, \beta|z, w, \alpha, \eta)q(z|\phi) \quad \text{notice the first term is exact}
\]

\[
=p(\theta, \beta|z, w, \alpha, \eta) \prod_{d=1}^{D} q(z_d|\phi_d)
\]

Observe that Blei’s class of distribution is a subset of Teh’s: \(\prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} q_0(\theta_d|\gamma_d) \subset q(\theta, \beta|z, w, \alpha, \eta)\) since the former class of distributions assume independence between \(\beta\) and \(\theta\), while the latter makes no assumption about their independence. Therefore intuitively, Teh’s approximation is better. We can also prove this more formally:

\[
\text{KL}(q(\beta, \theta, z)||p(\beta, \theta, z|w, \alpha, \eta)) = \text{KL}(q(z)q(\beta, \theta|z)||p(\beta, \theta, z|w, \alpha, \eta))
\]

\[
= E_q \left( \log \frac{q(z)q(\beta, \theta|z)}{p(\beta, \theta, z|w, \alpha, \eta)} \right)
\]

\[
= E_q \left( \log \frac{q(z)}{p(z|w, \alpha, \eta)} + \log \frac{q(\beta, \theta|z)}{p(\beta, \theta|z, w, \alpha, \eta)} \right)
\]

\[
= E_q \left( \log \frac{q(z)}{p(z|w, \alpha, \eta)} \right) + E_q(\beta, \theta|z) \left( \log \frac{q(\beta, \theta|z)}{p(\beta, \theta|z, w, \alpha, \eta)} \right)
\]

\[
\geq \text{KL}(q(z)||p(z|w, \alpha, \eta)) \quad \text{with equality when } q(\beta, \theta|z) = p(\beta, \theta|z, w, \alpha, \eta)
\]

Therefore KL divergence of Teh’s approximation is at most equal to the KL divergence of Blei’s approximation because the second KL divergence term is zero for Teh’s approximation but non-zero for Blei’s approximation. The question then is whether Teh’s approximation is tractable.

### 1.3 Simplifying the lower bound on \(\log p(w|\alpha, \eta)\) with Teh’s approximation

We can plug in \(q(\beta, \theta, z) = p(\theta, \beta|z, w, \alpha, \eta)q(z|\phi)\) into Eq. 14 and simplify. However, it’s simpler to restart from \(\log p(w|\alpha, \eta)\):

\[
\log p(w|\alpha, \eta) = \log \sum_z p(z, w|\alpha, \eta)
\]

\[
= \log \sum_z q(z)p(z, w|\alpha, \eta)
\]

\[
= \log E_q \left( \frac{p(z, w|\alpha, \eta)}{q(z)} \right)
\]

\[
\geq E_q(\log p(z, w|\alpha, \eta)) - E_q(\log q(z))
\]

The class of proposal distribution we will consider is \(q(z) = \prod_{d=1}^{D} q(z_d|\phi_d) = \prod_{d,n} q(z_{dn}|\phi_{dn})\). Note that \(q(z_{dn}|\phi_{dn})\) is a categorical distribution with parameters \(\phi_{dn}\). In other works \(\phi_{dn}\) is a vector of length \(K\) (corresponding to topics), such that \(\sum_k \phi_{dnk} = 1\).

### 1.4 Variational E-step

The goal is to maximize Eq. 26 with respect to \(\{\phi_{dn}\}\). We have to add the lagrange multipliers corresponding to \(\sum_k \phi_{dnk} = 1\). Furthermore, note that

\[
E_q(\log q(z)) = \sum_{z_{11}} \cdots \sum_{z_{DN}} q(z_{11}) \cdots q(z_{DN}) \left[ \log q(z_{11}) + \cdots + \log q(z_{DN}) \right]
\]

\[
= \sum_{d} \sum_{n} \sum_{k} q(z_{dn} = k) \log q(z_{dn} = k)
\]

\[
= \sum_{d} \sum_{n} \sum_{k} \phi_{dnk} \log \phi_{dnk}
\]
Therefore, Eq. \[26\] becomes
\[
E_q(\log p(z, w|\alpha, \eta)) - E_q(\log q(z)) + \sum_{dn} \mu_{dn} \left( \sum_k \phi_{dnk} - 1 \right)
\] (30)
\[= E_q(\log p(z, w|\alpha, \eta)) - \sum_d \sum_n \sum_k \phi_{dnk} \log \phi_{dnk} + \sum_{dn} \mu_{dn} \left( \sum_k \phi_{dnk} - 1 \right)
\] (31)
\[= E_q(z_{ij}) \left( E_q(z_{ij}) \left( \log p(z, w|\alpha, \eta) \right) \right) - \sum_d \sum_n \sum_k \phi_{dnk} \log \phi_{dnk} + \sum_{dn} \mu_{dn} \left( \sum_k \phi_{dnk} - 1 \right)
\] (32)
\[= \sum_k \phi_{ijk} \left( E_q(z_{ij}) \left( \log p(z_{ij}, z_{ij} = k, w|\alpha, \eta) \right) \right) - \sum_d \sum_n \sum_k \phi_{dnk} \log \phi_{dnk} + \sum_{dn} \mu_{dn} \left( \sum_k \phi_{dnk} - 1 \right)
\] (33)
where \(i\) indexes a particular document, \(j\) indexes the \(j\)-th word of the document. Differentiating with respect to \(\phi_{ijl}\), we get
\[
E_q(z_{ij}) \left( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \right) - \log \phi_{ijl} - 1 + \mu_{ij}
\] (34)
Equating the above to 0, we get
\[
\phi_{ijl} \propto \exp \left( E_q(z_{ij}) \left( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \right) \right)
\] (35)
\[
\phi_{ijl} = \frac{\exp \left( E_q(z_{ij}) \left( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \right) \right)}{\sum_{k'} \exp \left( E_q(z_{ij}) \left( \log p(z_{ij}, z_{ij} = k', w|\alpha, \eta) \right) \right)}
\] (36)

### 1.4.1 Plugging in the counts

Using the fact about Dirichlet-compound-multinomial distribution [wikipedia link], we get
\[
p(z_d|\alpha) = \int p(z_d|\theta_d)p(\theta_d|\alpha)d\theta_d = \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d.})} \prod_k \frac{\Gamma(\alpha + N_{dk.})}{\Gamma(\alpha)}
\] (37)
where \(N_{dv}\) is the number of words in the \(d\)-th document belonging to topic \(k\) and corresponding to dictionary word \(v\). Dot implies corresponding indices are summed out. For example, \(N_{kv} = \sum_d N_{dv}\) is the number of words in the corpus belonging to topic \(k\) and dictionary word word \(v\), as well as
\[
p(w|z, \eta) = \prod_k \left( \frac{\Gamma(V \eta)}{\Gamma(V \eta + N_{vk})} \prod_v \frac{\Gamma(\eta + N_{kv})}{\Gamma(\eta)} \right)
\] (38)
The way to think about the above equation is that conditioned on the topics \(z\), we can divide all the words in the corpus \(w\) into words from topic \(1\) to topic \(K\). Words generated from a particular topic are independent of words generated from another topics (hence the \(\prod_k\)). For a given topic \(k\), the words generated by the topic \(k\) follow a Dirichlet-compound-multinomial distribution with hyperparameter \(\eta\). Using both two equations, we get
\[
\log p(z, w|\alpha, \eta)
\]
\[= \log p(z|\alpha)p(w|z, \eta)
\] (39)
\[= \log \left( \prod_d \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d.})} \prod_k \frac{\Gamma(\alpha + N_{dk.})}{\Gamma(\alpha)} \right) \left( \prod_k \frac{\Gamma(V \eta)}{\Gamma(V \eta + N_{vk})} \prod_v \frac{\Gamma(\eta + N_{kv})}{\Gamma(\eta)} \right)
\] (40)
\[= \sum_d \log \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d.})} + \sum_d \sum_k \log \frac{\Gamma(\alpha + N_{dk.})}{\Gamma(\alpha)} + \sum_k \log \frac{\Gamma(V \eta)}{\Gamma(V \eta + N_{vk})} + \sum_v \sum_k \log \frac{\Gamma(\eta + N_{kv})}{\Gamma(\eta)}
\] (41)
\[= - \sum_d \sum_{m=0}^{N_d-1} \log(K\alpha + m) + \sum_d \sum_k \sum_{m=0}^{N_{dk}-1} \log(\alpha + m) - \sum_k \sum_{m=0}^{N_{k-1}} \log(V \eta + m) + \sum_k \sum_v \sum_{m=0}^{N_{kv}-1} \log(\eta + m)
\] (42)

We now come to a tricky part of the derivations! Consider Eq. (36). The numerator and denominator will have four terms corresponding to those from Eq. (42).
• The first term will be the same for both numerator and denominator and will therefore cancel out.

• The second term of \( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \) (i.e., numerator of Eq. (36) excluding the exponential and expectation) corresponds to:

\[
\sum_{d} \sum_{k} \sum_{m=0}^{N_{dk}-1} \log(\alpha + m) = \left( \sum_{d \neq \hat{d}} \sum_{k} \sum_{m=0}^{N_{dk}-1} \log(\alpha + m) \right) + \left( \sum_{k} \sum_{m=0}^{N_{kij}-1} \log(\alpha + m) \right) + \log(\alpha + N_{dk}^{-ij}),
\]

where the first two terms will cancel with the denominator because they are independent of \( l \).

• Similarly, the third term of \( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \) (i.e., numerator of Eq. (36) excluding the exponential and expectation) corresponds to:

\[
- \sum_{k} \sum_{m=0}^{N_{kij}-1} \log(V\eta + m) = \left( - \sum_{k} \sum_{m=0}^{N_{kij}-1} \log(V\eta + m) \right) - \log(\eta + N_{kij}^{-ij}),
\]

where the first term will cancel with the denominator because they are independent of \( l \).

• Finally, the fourth term of \( \log p(z_{ij}, z_{ij} = l, w|\alpha, \eta) \) (i.e., numerator of Eq. (36) excluding the exponential and expectation) corresponds to:

\[
\sum_{k} \sum_{v} \sum_{m=0}^{N_{kwij}-1} \log(\eta + m) = \left( \sum_{k} \sum_{v} \sum_{m=0}^{N_{kij}-1} \log(\eta + m) \right) + \log(\eta + N_{kwij}^{-ij}),
\]

where the first term will cancel with the denominator because they are independent of \( l \).

Therefore Eq. (36) becomes

\[
\phi_{ijl} = \frac{\exp(E_q(z_{ij}))(\log(\alpha + N_{dk}^{-ij}) - \log(V\eta + N_{kij}^{-ij}) + \log(\eta + N_{kwij}^{-ij}))}{\sum_{k'} \exp(E_q(z_{ij}))(\log(\alpha + N_{dk}^{-ij}) - \log(V\eta + N_{kij}^{-ij}) + \log(\eta + N_{kwij}^{-ij}))}
\]

Renaming \( i \) to \( d \), \( j \) to \( n \), and \( l \) to \( k \), we get

\[
\phi_{dsk} = \frac{\exp(E_q(z_{dn})(\log(\alpha + N_{dk}^{-dn}) - \log(V\eta + N_{kn}^{-dn}) + \log(\eta + N_{kwdn}^{-dn}))}{\sum_{k'} \exp(E_q(z_{dn}))(\log(\alpha + N_{dk}^{-dn}) - \log(V\eta + N_{kn}^{-dn}) + \log(\eta + N_{kwdn}^{-dn}))}
\]

### 1.4.2 Approximating the counts \( N \)

- \( N_{dk}^{-dn} = \sum_{n' \neq n} \mathbb{I}(z_{dn'} = k) \), which is a sum of independent (by the mean field approximation) bernoulli variables with probability of “heads” equal to \( \phi_{dsk} \). Therefore the mean and variance of \( N_{dk}^{-dn} \) is given by

\[
E_q(N_{dk}^{-dn}) = \sum_{n' \neq n} \phi_{dsk} \quad \text{Var}_q(N_{dk}^{-dn}) = \sum_{n' \neq n} \phi_{dsk}(1 - \phi_{dsk})
\]

- \( N_{kn}^{-dn} = \sum_{(d', n') = (d, n)} \mathbb{I}(z_{d'n'} = k) \), with mean and variance given by

\[
E_q(N_{kn}^{-dn}) = \sum_{(d', n') \neq (d, n)} \phi_{d'n'k} \quad \text{Var}_q(N_{kn}^{-dn}) = \sum_{(d', n') \neq (d, n)} \phi_{d'n'k}(1 - \phi_{d'n'k})
\]

- \( N_{kwdn}^{-dn} = \sum_{(d', n') \neq (d, n)} \mathbb{I}(z_{d'n'} = k)\mathbb{I}(w_{d'n'} = w_{dn}) \) with mean and variance given by

\[
E_q(N_{kwdn}^{-dn}) = \sum_{(d', n') \neq (d, n)} \phi_{d'n'k}(w_{d'n'} = w_{dn}) \quad \text{Var}_q(N_{kwdn}^{-dn}) = \sum_{(d', n') \neq (d, n)} \phi_{d'n'k}(1 - \phi_{d'n'k})(w_{d'n'} = w_{dn})
\]
As suggested by Teh et al. (2007), we will approximate the random variables $N_{dk}^{-dn}$, $N_{k}^{-dn}$, and $N_{kw,da}^{-dn}$ as Gaussian random variables with mean and variance as discussed above. First, note that

$$f(x) = \log(b + x)$$

$$= f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \cdots$$

$$= f(a) + \frac{1}{b + a}(x - a) - \frac{1}{2(b + a)^2}(x - a)^2 + \cdots$$

Therefore the terms in the numerator of Eq. (48) becomes

- Let $b = \alpha$, $x = N_{dk}^{-dn}$ and $a = E_q(N_{dk}^{-dn})$

$$E_q(z^{-dn})(\log(\alpha + N_{dk}^{-dn}))$$

$$= f\left(E_q(N_{dk}^{-dn})\right) + \frac{1}{\alpha + E_q(N_{dk}^{-dn})}E_q\left(N_{dk}^{-dn} - E_q(N_{dk}^{-dn})\right) - \frac{1}{2(\alpha + E_q(N_{dk}^{-dn}))^2}E_q\left(N_{dk}^{-dn} - E_q(N_{dk}^{-dn})\right)^2$$

$$= \log(\alpha + E_q(N_{dk}^{-dn})) - \frac{\text{Var}_q(N_{dk}^{-dn})}{2(\alpha + E_q(N_{dk}^{-dn}))^2}$$

- Let $b = V\eta$, $x = N_{k}^{-dn}$ and $a = E_q(N_{k}^{-dn})$

$$E_q(z^{-dn})(\log(V\eta + N_{k}^{-dn})) = \log(V\eta + E_q(N_{k}^{-dn})) - \frac{\text{Var}_q(N_{k}^{-dn})}{2(V\eta + E_q(N_{k}^{-dn}))^2}$$

- Let $b = \eta$, $x = N_{kw,da}^{-dn}$ and $a = E_q(N_{kw,da}^{-dn})$

$$E_q(z^{-dn})(\log(\eta + N_{kw,da}^{-dn})) = \log(\eta + E_q(N_{kw,da}^{-dn})) - \frac{\text{Var}_q(N_{kw,da}^{-dn})}{2(\eta + E_q(N_{kw,da}^{-dn}))^2}$$

Plugging the above into Eq. (48), we get

$$\phi_{dk} \propto \left(\alpha + E_q(N_{dk}^{-dn})\right)^{-1}\left(V\eta + E_q(N_{k}^{-dn})\right)^{-1}\left(\eta + E_q(N_{kw,da}^{-dn})\right)\times$$

$$\times \exp\left(-\frac{\text{Var}_q(N_{dk}^{-dn})}{2(\alpha + E_q(N_{dk}^{-dn}))^2} + \frac{\text{Var}_q(N_{k}^{-dn})}{2(V\eta + E_q(N_{k}^{-dn}))^2} - \frac{\text{Var}_q(N_{kw,da}^{-dn})}{2(\eta + E_q(N_{kw,da}^{-dn}))^2}\right)$$

### 1.5 M-step

Teh et al. (2007) stopped at the variational E-step. Here, we will consider how to optimize $\alpha$ and $\eta$. In the M-step, we seek to maximize Eq. (26) with respect to $\alpha$ and $\eta$. Keeping only the first term, which contains all the terms that do not contain $\alpha$, we get

$$E_q(\log p(z, w|\alpha, \eta))$$

$$= E_q\left(-\sum_d \sum_{m=0}^{N_{d}^{-1}} \log(K\alpha + m) + \sum_d \sum_k \sum_{m=0}^{N_{dk}^{-1}} \log(\alpha + m) - \sum_k \sum_{m=0}^{N_{k}^{-1}} \log(V\eta + m) + \sum_k \sum_v \sum_{m=0}^{N_{kv}^{-1}} \log(\eta + m)\right)$$
1.5.1 Optimize $\alpha$

Let's consider the terms with $\alpha$:

$$L_\alpha = E_q \left( - \sum_d \sum_{m=0}^{N_d-1} \log(K\alpha + m) + \sum_d \sum_k \sum_{m=0}^{N_d-1} \log(\alpha + m) \right)$$

(63)

$$= - \sum_d \sum_{m=0}^{N_d-1} \log(K\alpha + m) + E_q \left( \sum_d \sum_k \sum_{m=0}^{N_d-1} \log(\alpha + m) \right)$$

(64)

$$= - \sum_d \sum_{m=0}^{N_d-1} \log(K\alpha + m) + \sum_d \sum_k \sum_{n=1}^{N_d} q(N_{dk} = n) \sum_{m=0}^{n-1} \log(\alpha + m)$$

(65)

$$= - \sum_d \sum_{m=0}^{N_d-1} \log(K\alpha + m) + \sum_d \sum_k \sum_{n=1}^{N_d} q(N_{dk} \geq n) \log(\alpha + n - 1)$$

(66)

Differentiating with respect to $\alpha$, we get

$$\frac{\partial L_\alpha}{\partial \alpha} = - \sum_d \sum_{m=0}^{N_d-1} \frac{K}{K\alpha + m} + \sum_d \sum_k \sum_{n=1}^{N_d} q(N_{dk} \geq n) \frac{1}{\alpha + n - 1}$$

(67)

Differentiating one more time, we get

$$\frac{\partial^2 L_\alpha}{\partial \alpha^2} = \sum_d \sum_{m=0}^{N_d-1} \frac{K^2}{(K\alpha + m)^2} - \sum_d \sum_k \sum_{n=1}^{N_d} q(N_{dk} \geq n) \frac{1}{(\alpha + n - 1)^2}$$

(68)

We can use the Hessian and Gradient to compute the Newton-Raphson update.

To compute $q(N_{dk} \geq n)$, note that $N_{dk} = \sum_n I(z_{dn} = k)$, which we can approximate by a Gaussian with mean $\sum_n \phi_{dnk}$ and variance $\sum_n \phi_{dnk}(1 - \phi_{dnk})$.

1.5.2 Optimize $\eta$

Let's consider the terms with $\eta$ and denoting $N.. = N$ (i.e., total number of words in all the documents), we get

$$L_\eta = E_q \left( - \sum_k \sum_{m=0}^{N_k-1} \log(V\eta + m) + \sum_k \sum_v \sum_{m=0}^{N_v-1} \log(\eta + m) \right)$$

(69)

$$= - \sum_k \sum_{n=1}^{N_k} q(N_k = n) \sum_{m=0}^{n-1} \log(V\eta + m) + \sum_k \sum_v \sum_{n=1}^{N_v} q(N_{kv} = n) \sum_{m=0}^{n-1} \log(\eta + m)$$

(70)

$$= - \sum_k \sum_{n=1}^{N_k} q(N_k \geq n) \log(V\eta + n - 1) + \sum_k \sum_v \sum_{n=1}^{N_v} q(N_{kv} \geq n) \log(\eta + n - 1)$$

(71)

Differentiating with respect to $\alpha$, we get

$$\frac{\partial L_\eta}{\partial \eta} = - \sum_k \sum_{n=1}^{N_k} q(N_k \geq n) \frac{V}{V\eta + n - 1} + \sum_k \sum_v \sum_{n=1}^{N_v} q(N_{kv} \geq n) \frac{1}{\eta + n - 1}$$

(72)

Differentiating one more time, we get

$$\frac{\partial^2 L_\eta}{\partial \eta^2} = \sum_k \sum_{n=1}^{N_k} q(N_k \geq n) \frac{V^2}{(V\eta + n - 1)^2} - \sum_k \sum_v \sum_{n=1}^{N_v} q(N_{kv} \geq n) \frac{1}{(\eta + n - 1)^2}$$

(73)

We can use the Hessian and Gradient to compute the Newton-Raphson update.

To compute $q(N_k \geq n)$, note that $N_k = \sum_n I(z_{dn} = k)$, which we can approximate by a Gaussian with mean $\sum_n \phi_{dnk}$ and variance $\sum_n \phi_{dnk}(1 - \phi_{dnk})$.

To compute $q(N_{kv} \geq n)$, note that $N_{kv} = \sum_n I(z_{dn} = k)I(w_{dn} = v)$, which we can approximate by a Gaussian with mean $\sum_n \phi_{dnk}I(w_{dn} = v)$ and variance $\sum_n \phi_{dnk}(1 - \phi_{dnk})I(w_{dn} = v)$.
1.6 Alternative M-step

As an alternative approach, we can use $\hat{\phi}_{dnk}$ to compute

$$\gamma_{dk} = \alpha_k + \sum_{n=1}^{N} \phi_{dnk}$$  \hspace{2cm} (74)

$$\lambda_{kv} = \eta_{kv} + \sum_{d} \sum_{n=1}^{N} \phi_{dnk} w_{dn}^{v},$$  \hspace{2cm} (75)

which we can then use to update $\alpha$ and $\eta$ exactly the same way as the original LDA (Blei et al., 2003). This is theoretically not as good as the previous section because we are using point estimates of $\theta$ and $\beta$ instead of integrating them out. But the estimates of $\phi$ should be better than the original LDA (Blei et al., 2003), so maybe it will perform better? However, there is no theoretical guarantees unlike the variational E-step.