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 α and η on β and θ respectively, where θ parameterizes $P(\text{topic}|\text{document})$ and β parameterizes $P(\text{word}|\text{topic})$, 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 α, η):

$$\begin{aligned}
\log p(w|\alpha,\eta) &= \log \int_\theta \int_\beta \sum_z p(\theta,\beta,z,w|\alpha,\eta) d\beta d\theta && (1)\\
&= \log \int_\theta \int_\beta \sum_z \frac{q(\beta,\theta,z) p(\theta,\beta,z,w|\alpha,\eta)}{q(\beta,\theta,z)} d\beta d\theta && (2)\\
&= \log E_q\left(\frac{p(\theta,\beta,z,w|\alpha,\eta)}{q(\beta,\theta,z)}\right) && (3)\\
&\geq E_q(\log p(\theta,\beta,z,w|\alpha,\eta)) - E_q(\log q(\beta,\theta,z)) && (4)\\
&= -F(q(\beta,\theta,z)) \quad F \text{ is called the variational free energy} && (5)
\end{aligned}$$

Note that

$$\begin{aligned}
&E_q(\log p(\theta,\beta,z,w|\alpha,\eta)) - E_q(\log q(\beta,\theta,z)) + \mathrm{KL}(q(\beta,\theta,z)||p(\beta,\theta,z|w,\alpha,\eta)) && (6)\\
&= E_q(\log p(\theta,\beta,z,w|\alpha,\eta)) - E_q(\log q(\beta,\theta,z)) + E_q\left(\log \frac{q(\beta,\theta,z)}{p(\beta,\theta,z|w,\alpha,\eta)}\right) && (7)\\
&= E_q\left(\log \frac{p(\theta,\beta,z,w|\alpha,\eta)}{p(\beta,\theta,z|w,\alpha,\eta)}\right) && (8)\\
&= E_q(\log p(w|\alpha,\eta)) && (9)\\
&= \log p(w|\alpha,\eta) && (10)
\end{aligned}$$

Therefore

$$\log p(w|\alpha,\eta) - (-F(q(\beta,\theta,z))) = \mathrm{KL}(q(\beta,\theta,z)||p(\beta,\theta,z|w,\alpha,\eta)) \quad (11)$$

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

In EM, we compute $q(\beta,\theta,z) = p(\beta,\theta,z|w,\alpha,\eta)$ exactly in the E-step and then conditioned on this particular q, we maximize with respect to α and η 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 α and η 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:

$$\begin{aligned}
q(\theta,z,\beta|\gamma,\phi,\lambda) &= q(\beta|\lambda)q(\theta|\gamma)q(z|\phi) && (12)\\
&= \prod_{k=1}^K q(\beta_k|\lambda_k) \prod_{d=1}^D q_d(\theta_d|\gamma_d) q(z_d|\phi_d) && (13)
\end{aligned}$$

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

$$q(\theta, z, \beta|\phi) = q(\theta, \beta|z)q(z|\phi) \tag{14}$$
$$= p(\theta, \beta|z, w, \alpha, \eta)q(z|\phi) \quad \text{notice the first term is exact} \tag{15}$$
$$= p(\theta, \beta|z, w, \alpha, \eta) \prod_{d=1}^{D} q(z_d|\phi_d) \tag{16}$$

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

$$\mathrm{KL}(q(\beta, \theta, z)||p(\beta, \theta, z|w, \alpha, \eta)) = \mathrm{KL}(q(z)q(\beta, \theta|z)||p(\beta, \theta, z|w, \alpha, \eta)) \tag{17}$$
$$= E_q\left(\log \frac{q(z)q(\beta, \theta|z)}{p(\beta, \theta, z|w, \alpha, \eta)} \right) \tag{18}$$
$$= E_q\left(\log \frac{q(z)}{p(z|w, \alpha, \eta)} + \log \frac{q(\beta, \theta|z)}{p(\beta, \theta|z, w, \alpha, \eta)} \right) \tag{19}$$
$$= E_q\left(\log \frac{q(z)}{p(z|w, \alpha, \eta)} \right) + E_{q(z)}\left(E_{q(\beta,\theta|z)}\left(\log \frac{q(\beta, \theta|z)}{p(\beta, \theta|z, w, \alpha, \eta)} \right) \right) \tag{20}$$
$$= E_{q(z)}\left(\log \frac{q(z)}{p(z|w, \alpha, \eta)} \right) + E_{q(z)}\left(\mathrm{KL}\Big(q(\beta, \theta|z)||p(\beta, \theta|z, w, \alpha, \eta)\Big) \right) \tag{21}$$
$$\geq \mathrm{KL}(q(z)||p(z|w, \alpha, \eta)) \quad \text{with equality when } q(\beta, \theta|z) = p(\beta, \theta|z, w, \alpha, \eta) \tag{22}$$

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. 4 and simplify. However, it's simpler to restart from $\log p(w|\alpha, \eta)$:

$$\log p(w|\alpha, \eta) = \log \sum_z p(z, w|\alpha, \eta) \tag{23}$$
$$= \log \sum_z \frac{q(z)p(z, w|\alpha, \eta)}{q(z)} \tag{24}$$
$$= \log E_q\left(\frac{p(z, w|\alpha, \eta)}{q(z)} \right) \tag{25}$$
$$\geq E_q(\log p(z, w|\alpha, \eta)) - E_q(\log q(z)) \tag{26}$$

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 ϕ_{dn}. In other works ϕ_{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

$$E_q(\log q(z)) = \sum_{z_{11}} \cdots \sum_{z_{DN}} q(z_{11}) \cdots q(z_{DN}) \Big[\log q(z_{11}) + \cdots + \log q(z_{DN}) \Big] \tag{27}$$
$$= \sum_d \sum_n \sum_k q(z_{dn} = k) \log q(z_{dn} = k) \tag{28}$$
$$= \sum_d \sum_n \sum_k \phi_{dnk} \log \phi_{dnk} \tag{29}$$

Therefore, Eq. 26 becomes

$$E_q(\log p(z,w|\alpha,\eta)) - E_q(\log q(z)) + \sum_{dn}\mu_{dn}\Big(\sum_k \phi_{dnk} - 1\Big) \tag{30}$$

$$= E_q(\log p(z,w|\alpha,\eta)) - \sum_d\sum_n\sum_k \phi_{dnk}\log\phi_{dnk} + \sum_{dn}\mu_{dn}\Big(\sum_k \phi_{dnk} - 1\Big) \tag{31}$$

$$= E_{q(z_{ij})}\bigg(E_{q(z_{-ij})}\Big(\log p(z,w|\alpha,\eta)\Big)\bigg) - \sum_d\sum_n\sum_k \phi_{dnk}\log\phi_{dnk} + \sum_{dn}\mu_{dn}\Big(\sum_k \phi_{dnk} - 1\Big) \tag{32}$$

$$= \sum_k \phi_{ijk}\bigg(E_{q(z_{-ij})}\Big(\log p(z_{-ij}, z_{ij}=k, w|\alpha,\eta)\Big)\bigg) - \sum_d\sum_n\sum_k \phi_{dnk}\log\phi_{dnk} + \sum_{dn}\mu_{dn}\Big(\sum_k \phi_{dnk} - 1\Big) \tag{33}$$

where i indexes a particular document, j indexes the j-th word of the document. Differentiating with respect to ϕ_{ijl}, we get

$$E_{q(z_{-ij})}\Big(\log p(z_{-ij}, z_{ij}=l, w|\alpha,\eta)\Big) - \log\phi_{ijl} - 1 + \mu_{ij} \tag{34}$$

Equating the above to 0, we get

$$\phi_{ijl} \propto \exp\bigg(E_{q(z_{-ij})}\Big(\log p(z_{-ij}, z_{ij}=l, w|\alpha,\eta)\Big)\bigg) \tag{35}$$

$$\phi_{ijl} = \frac{\exp\bigg(E_{q(z_{-ij})}\Big(\log p(z_{-ij}, z_{ij}=l, w|\alpha,\eta)\Big)\bigg)}{\sum_{k'}\exp\bigg(E_{q(z_{-ij})}\Big(\log p(z_{-ij}, z_{ij}=k', w|\alpha,\eta)\Big)\bigg)} \tag{36}$$

1.4.1 Plugging in the counts

Using the fact about Dirichlet-compound-multinomial distribution (wikipedia link), we get

$$p(z_d|\alpha) = \int p(z_d|\theta_d)p(\theta_d|\alpha)d\theta_d = \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d\cdot\cdot})}\prod_k \frac{\Gamma(\alpha + N_{dk\cdot})}{\Gamma(\alpha)}, \tag{37}$$

where N_{dkv} 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, $N_{\cdot kv} = \sum_d N_{dkv}$ is the number of words in the corpus belonging to topic k and dictionary word v, as well as

$$p(w|z,\eta) = \prod_k \bigg(\frac{\Gamma(V\eta)}{\Gamma(V\eta + N_{\cdot k\cdot})}\prod_v \frac{\Gamma(\eta + N_{\cdot kv})}{\Gamma(\eta)}\bigg). \tag{38}$$

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 η. Using both two equations, we get

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

$$= \log p(z|\alpha)p(w|z,\eta) \tag{39}$$

$$= \log\bigg(\prod_d \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d\cdot\cdot})}\prod_k \frac{\Gamma(\alpha + N_{dk\cdot})}{\Gamma(\alpha)}\bigg)\bigg(\prod_k \frac{\Gamma(V\eta)}{\Gamma(V\eta + N_{\cdot k\cdot})}\prod_v \frac{\Gamma(\eta + N_{\cdot kv})}{\Gamma(\eta)}\bigg) \tag{40}$$

$$= \sum_d \log\frac{\Gamma(K\alpha)}{\Gamma(K\alpha + N_{d\cdot\cdot})} + \sum_d\sum_k \log\frac{\Gamma(\alpha + N_{dk\cdot})}{\Gamma(\alpha)} + \sum_k \log\frac{\Gamma(V\eta)}{\Gamma(V\eta + N_{\cdot k\cdot})} + \sum_k\sum_v \log\frac{\Gamma(\eta + N_{\cdot kv})}{\Gamma(\eta)} \tag{41}$$

$$= -\sum_d\sum_{m=0}^{N_{d\cdot\cdot}-1}\log(K\alpha + m) + \sum_d\sum_k\sum_{m=0}^{N_{dk\cdot}-1}\log(\alpha + m) - \sum_k\sum_{m=0}^{N_{\cdot k\cdot}-1}\log(V\eta + m) + \sum_k\sum_v\sum_{m=0}^{N_{\cdot kv}-1}\log(\eta + m) \tag{42}$$

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:

$$\sum_d \sum_k \sum_{m=0}^{N_{dk\cdot}-1} \log(\alpha + m) \tag{43}$$

$$= \Big(\sum_{d \neq i} \sum_k \sum_{m=0}^{N_{dk\cdot}-1} \log(\alpha + m) \Big) + \Big(\sum_k \sum_{m=0}^{N_{ik\cdot}^{-ij}-1} \log(\alpha + m) \Big) + \log(\alpha + N_{il\cdot}^{-ij}), \tag{44}$$

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:

$$-\sum_k \sum_{m=0}^{N_{\cdot k\cdot}-1} \log(V\eta + m) = \Big(-\sum_k \sum_{m=0}^{N_{\cdot k\cdot}^{-ij}-1} \log(V\eta + m) \Big) - \log(V\eta + N_{\cdot l\cdot}^{-ij}), \tag{45}$$

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:

$$\sum_k \sum_v \sum_{m=0}^{N_{\cdot kv}-1} \log(\eta + m) = \Big(\sum_k \sum_v \sum_{m=0}^{N_{\cdot kv}^{-ij}-1} \log(\eta + m) \Big) + \log(\eta + N_{\cdot lw_{ij}}^{-ij}), \tag{46}$$

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

Therefore Eq. (36) becomes

$$\phi_{ijl} = \frac{\exp(E_{q(z^{-ij})}\big(\log(\alpha + N_{il\cdot}^{-ij}) - \log(V\eta + N_{\cdot l\cdot}^{-ij}) + \log(\eta + N_{\cdot lw_{ij}}^{-ij})\big))}{\sum_{k'} \exp(E_{q(z^{-ij})}\big(\log(\alpha + N_{ik'\cdot}^{-ij}) - \log(V\eta + N_{\cdot k'\cdot}^{-ij}) + \log(\eta + N_{\cdot k'w_{ij}}^{-ij})\big))} \tag{47}$$

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

$$\phi_{dnk} = \frac{\exp(E_{q(z^{-dn})}\big(\log(\alpha + N_{dk\cdot}^{-dn}) - \log(V\eta + N_{\cdot k\cdot}^{-dn}) + \log(\eta + N_{\cdot kw_{dn}}^{-dn})\big))}{\sum_{k'} \exp(E_{q(z^{-dn})}\big(\log(\alpha + N_{dk'\cdot}^{-dn}) - \log(V\eta + N_{\cdot k'\cdot}^{-dn}) + \log(\eta + N_{\cdot k'w_{dn}}^{-dn})\big))} \tag{48}$$

1.4.2 Approximating the counts N

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

$$E_q(N_{dk\cdot}^{-dn}) = \sum_{n' \neq n} \phi_{dn'k} \qquad \text{Var}_q(N_{dk\cdot}^{-dn}) \sum_{n' \neq n} \phi_{dn'k}(1 - \phi_{dn'k}) \tag{49}$$

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

$$E_q(N_{\cdot k\cdot}^{-dn}) = \sum_{(d',n') \neq (d,n)} \phi_{d'n'k} \qquad \text{Var}_q(N_{\cdot k\cdot}^{-dn}) = \sum_{(d',n') \neq (d,n)} \phi_{d'n'k}(1 - \phi_{d'n'k}) \tag{50}$$

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

$$E_q(N_{\cdot kw_{dn}}^{-dn}) = \sum_{(d',n') \neq (d,n)} \phi_{d'n'k}\mathbb{I}(w_{d'n'} = w_{dn}) \qquad \text{Var}_q(N_{\cdot kw_{dn}}^{-dn}) = \sum_{(d',n') \neq (d,n)} \phi_{d'n'k}(1 - \phi_{d'n'k})\mathbb{I}(w_{d'n'} = w_{dn}) \tag{51}$$

As suggested by Teh et al. (2007), we will approximate the random variables $N_{dk\cdot}^{-dn}$, $N_{\cdot k\cdot}^{-dn}$ and $N_{\cdot kw_{dn}}^{-dn}$ as Gaussian random variables with mean and variance as discussed above. First, note that

$$\begin{aligned} f(x) &= \log(b+x) && (52)\\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots && (53)\\ &= f(a) + \frac{1}{b+a}(x-a) - \frac{1}{2(b+a)^2}(x-a)^2 + \cdots && (54)\end{aligned}$$

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

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

$$E_{q(z^{-dn})}\big(\log(\alpha + N_{dk\cdot}^{-dn})\big) \tag{55}$$

$$= f\Big(E_q(N_{dk\cdot}^{-dn})\Big) + \frac{1}{\alpha + E_q(N_{dk\cdot}^{-dn})} E_q\Big(N_{dk\cdot}^{-dn} - E_q(N_{dk\cdot}^{-dn})\Big) - \frac{1}{2(\alpha + E_q(N_{dk\cdot}^{-dn}))^2} E_q\Big(N_{dk\cdot}^{-dn} - E_q(N_{dk\cdot}^{-dn})\Big)^2 \tag{56}$$

$$= \log(\alpha + E_q(N_{dk\cdot}^{-dn})) - \frac{\text{Var}_q(N_{dk\cdot}^{-dn})}{2(\alpha + E_q(N_{dk\cdot}^{-dn}))^2} \tag{57}$$

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

$$E_{q(z^{-dn})}\big(\log(V\eta + N_{\cdot k\cdot}^{-dn})\big) = \log(V\eta + E_q(N_{\cdot k\cdot}^{-dn})) - \frac{\text{Var}_q(N_{\cdot k\cdot}^{-dn})}{2(V\eta + E_q(N_{\cdot k\cdot}^{-dn}))^2} \tag{58}$$

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

$$E_{q(z^{-dn})}\big(\log(\eta + N_{\cdot kw_{dn}}^{-dn})\big) = \log(\eta + E_q(N_{\cdot kw_{dn}}^{-dn})) - \frac{\text{Var}_q(N_{\cdot kw_{dn}}^{-dn})}{2(\eta + E_q(N_{\cdot kw_{dn}}^{-dn}))^2} \tag{59}$$

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

$$\begin{aligned}\phi_{dnk} \propto &\Big(\alpha + E_q(N_{dk\cdot}^{-dn})\Big)\Big(V\eta + E_q(N_{\cdot k\cdot}^{-dn})\Big)^{-1}\Big(\eta + E_q(N_{\cdot kw_{dn}}^{-dn})\Big)\times \\ &\times \exp\left(-\frac{\text{Var}_q(N_{dk\cdot}^{-dn})}{2(\alpha + E_q(N_{dk\cdot}^{-dn}))^2} + \frac{\text{Var}_q(N_{\cdot k\cdot}^{-dn})}{2(V\eta + E_q(N_{\cdot k\cdot}^{-dn}))^2} - \frac{\text{Var}_q(N_{\cdot kw_{dn}}^{-dn})}{2(\eta + E_q(N_{\cdot kw_{dn}}^{-dn}))^2}\right)\end{aligned} \tag{60}$$

1.5 M-step

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

$$E_q(\log p(z, w|\alpha, \eta)) \tag{61}$$

$$= E_q\Bigg(-\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \log(K\alpha + m) + \sum_d \sum_k \sum_{m=0}^{N_{dk\cdot}-1} \log(\alpha + m) - \sum_k \sum_{m=0}^{N_{\cdot k\cdot}-1} \log(V\eta + m) + \sum_k \sum_v \sum_{m=0}^{N_{\cdot kv}-1} \log(\eta + m)\Bigg) \tag{62}$$

1.5.1 Optimize α

Let's consider the terms with α:

$$L_\alpha = E_q\Big(-\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \log(K\alpha+m) + \sum_d \sum_k \sum_{m=0}^{N_{dk\cdot}-1} \log(\alpha+m)\Big) \tag{63}$$

$$= -\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \log(K\alpha+m) + E_q\Big(\sum_d \sum_k \sum_{m=0}^{N_{dk\cdot}-1} \log(\alpha+m)\Big) \tag{64}$$

$$= -\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \log(K\alpha+m) + \sum_d \sum_k \sum_{n=1}^{N_{d\cdot\cdot}} q(N_{dk\cdot}=n) \sum_{m=0}^{n-1} \log(\alpha+m) \tag{65}$$

$$= -\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \log(K\alpha+m) + \sum_d \sum_k \sum_{n=1}^{N_{d\cdot\cdot}} q(N_{dk\cdot}\geq n) \log(\alpha+n-1) \tag{66}$$

Differentiating with respect to α, we get

$$\frac{\partial L_\alpha}{\partial \alpha} = -\sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \frac{K}{K\alpha+m} + \sum_d \sum_k \sum_{n=1}^{N_{d\cdot\cdot}} q(N_{dk\cdot}\geq n)\frac{1}{\alpha+n-1} \tag{67}$$

Differentiating one more time, we get

$$\frac{\partial^2 L_\alpha}{\partial \alpha^2} = \sum_d \sum_{m=0}^{N_{d\cdot\cdot}-1} \frac{K^2}{(K\alpha+m)^2} - \sum_d \sum_k \sum_{n=1}^{N_{d\cdot\cdot}} q(N_{dk\cdot}\geq n)\frac{1}{(\alpha+n-1)^2} \tag{68}$$

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

To compute $q(N_{dk\cdot}\geq n)$, note that $N_{dk\cdot} = \sum_n \mathbb{I}(z_{dn}=k)$, 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 η

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

$$L_\eta = E_q\Big(-\sum_k \sum_{m=0}^{N_{\cdot k\cdot}-1} \log(V\eta+m) + \sum_k \sum_v \sum_{m=0}^{N_{\cdot kv}-1} \log(\eta+m)\Big) \tag{69}$$

$$= -\sum_k \sum_{n=1}^{N} q(N_{\cdot k\cdot}=n) \sum_{m=0}^{n-1} \log(V\eta+m) + \sum_k \sum_v \sum_{n=1}^{N_{\cdot\cdot v}} q(N_{\cdot kv}=n) \sum_{m=0}^{n-1} \log(\eta+m) \tag{70}$$

$$= -\sum_k \sum_{n=1}^{N} q(N_{\cdot k\cdot}\geq n) \log(V\eta+n-1) + \sum_k \sum_v \sum_{n=1}^{N_{\cdot\cdot v}} q(N_{\cdot kv}\geq n) \log(\eta+n-1) \tag{71}$$

Differentiating with respect to α, we get

$$\frac{\partial L_\eta}{\partial \eta} = -\sum_k \sum_{n=1}^{N} q(N_{\cdot k\cdot}\geq n)\frac{V}{V\eta+n-1} + \sum_k \sum_v \sum_{n=1}^{N_{\cdot\cdot v}} q(N_{\cdot kv}\geq n)\frac{1}{\eta+n-1} \tag{72}$$

Differentiating one more time, we get

$$\frac{\partial^2 L_\eta}{\partial \eta^2} = \sum_k \sum_{n=1}^{N} q(N_{\cdot k\cdot}\geq n)\frac{V^2}{(V\eta+n-1)^2} - \sum_k \sum_v \sum_{n=1}^{N_{\cdot\cdot v}} q(N_{\cdot kv}\geq n)\frac{1}{(\eta+n-1)^2} \tag{73}$$

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

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

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

1.6 Alternative M-step

As an alternative approach, we can use ϕ_{dnk} to compute

$$\gamma_{dk} = \alpha_k + \sum_{n=1}^{N} \phi_{dnk} \tag{74}$$

$$\lambda_{kv} = \eta_{kv} + \sum_{d} \sum_{n=1}^{N} \phi_{dnk} w_{dn}^{v}, \tag{75}$$

which we can then use to update α and η 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 θ and β instead of integrating them out. But the estimates of ϕ 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.