Latent Dirichlet Allocation Posterior Inference

Wang Yiran
amyran_1122@hotmail.com
November 28, 2017

1 Introduction of Latent Dirichlet Allocation (LDA)

1.1 Properties of LDA

LDA probability model posits that documents exhibit multiple topics!

1. Treat data as observations that arise from a generative probability process that includes hidden variables (e.g. thematic structure of the collection)
2. Infer the hidden structure using posterior inference while only observe documents. What are the topics that describe this collection?
3. Situate new data into estimated model.

1.2 Generative Model

(Blei et al., 2003) Each document is a random mixture of corpus-wide topics, while each word is drawn from one of those topics.
Define terms as following:

Terms	Description
$word$, $\mathbf{w}$	The unit-basis vectors defined to be an item from a vocabulary indexed by 1, ..., V. Thus, the vth word in the vocabulary is represented by a Vvector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.
$document$, $\mathbf{D}$	A sequence of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_N)$, where w_n is the nth word in the sequence.
$corpus$	A collection of M documents denoted by $D = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N)$.

Table 1: Terms Definition

LDA generative process:

1. Determine number of words in document. Choose $N \sim Poisson(\xi)$.
2. Choose a topic mixture for the documents over a fixed set of topics.
3. Generate words in document. For each of the N words w_n:
   (a) Choose a topic based on the document multinomial distribution $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n based on the topic's multinomial distribution from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Each piece of structure as shown in Figure is a random variable and the joint distribution explains as follow:

$$p(\mathbf{D}, \mathbf{w}) = \prod_{d=1}^{D} p(\theta_d \mid \alpha)(\prod_{n=1}^{N} p(z_n \mid \theta_d) p(w_{d,n} \mid z_{d,n}, \beta), \text{ where } p(\theta_d \mid \alpha) \sim Dirichlet \tag{1}$$

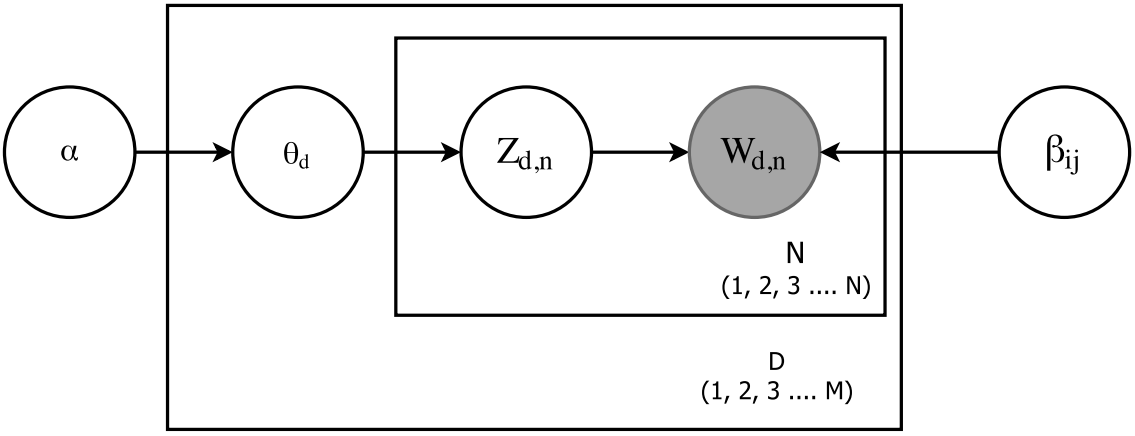

Figure 1: Graphical model representation of LDA

Parameter	Description
α	Dirichlet parameters.
θ_d	Per-doc topic proportions $\theta_d \in R^k$. k dimensional Dir. random variable θ can take values in (k-1)-simplex with $p.d.f\ p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$.
$Z_{d,n}$	Topic assignment drawn from distribution θ_d, which takes values from 1, 2 .... k.
$W_{d,n}$	Observed n^{th} word from d^{th} document.
β_{ij}	$\beta_{ij} = p(w^j = 1 \mid z^i = 1)$ Topic's distribution over words; per-corpus topic distribution.

Table 2: Parameters Description

1.3 Posterior Inference

(Blei et al., 2003) Infer the hidden structure using posterior inference while only observe document. Approximate per-doc θ assignment given observations, equivalently, predicting the distribution of hidden variables $Z_{d,n}$)
Posterior Inference:

$$\begin{aligned} p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) &= \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)} \\ &= \frac{p(\theta \mid \alpha) P(z_n \mid \theta) p(w_{d,n} \mid z_{d,n}, \beta)}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n \mid z_n, \beta)) d\theta} \end{aligned} \tag{2}$$

where $p(\theta \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$ and $p(\mathbf{z}_n \mid \theta) = \theta_i$ for unique i such that $\mathbf{z}_{n,i} = 1$.
Likelihood term, the probability of $w_{d,n}$ conditioning on latent topic assignment $z_{d,n}$ and distribution β (marginal distribution of a document):

$$\begin{aligned} p(\mathbf{w} \mid \alpha, \beta) &= \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int_\theta (\sum_{i=1}^{k} \theta_i^{\alpha_i - 1}) (\prod_{i=1}^{k} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j} d\theta \end{aligned} \tag{3}$$

Though the posterior distribution is intractable for exact inference, there are various approaches to approximate inference for LDA generative models, this document is going to detailedly illustrated Variational approximation and Gibbs Sampling methods.

2 Variational Approximation Posterior Inference

2.1 Variational Inference

Utilize Jensen's inequality to obtain an adjustable lower bound on the log likelihood.

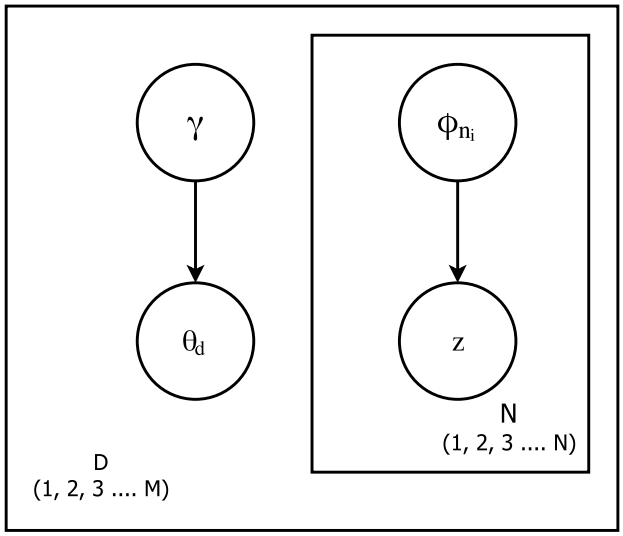

Figure 2: Graphical model representation of variational distribution

Variational Parameters	Description
γ	Dirichlet parameter for θ.
ϕ_{ni}	Multinomial parameters $(\phi_1 \cdots \phi_{ni})$. ϕ_{ni} denotes the probability of nth word is generated by latent topic i.

Table 3: Variational parameters

Approximate the posterior inference (Equation 2) with introduced variational parameters, and the variational distribution is shown as following:

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n) \tag{4}$$

Jensen's Inequality (lower bound of the marginal log-likelihood of documents)

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int_\theta \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta \\
&= \log \int_\theta \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\
&\geq \int_\theta \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta - \int_\theta \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}) d\theta \\
&= E_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})] = \mathcal{L}(\gamma, \phi; \alpha, \beta)
\end{aligned} \tag{5}$$

The difference between log likelihood and lower bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ is the KL divergence between the variational posterior probability and the true variational posterior probability. Hence, we obtain equation

as follow:

$$\log p(\mathbf{w} \mid \alpha, \beta) = \mathcal{L}(\gamma, \phi; \alpha, \beta) + D(q(\theta, \mathbf{z}) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)) \tag{6}$$

Kullback-Leiber (KL) Divergence: $D_{kl}(p \parallel q) = \int p(x) \ln \frac{p(x)}{q(x)dx}$

From equation 6, we can infer that maximizing the lower bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ with respect to variational parameters γ and ϕ is equivalent to minimizing the KL divergence. Apply the mean-field assumption to the lower bound and then expand it by using the factorization of p and q:

$$\begin{aligned} \mathcal{L}(\gamma, \phi; \alpha, \beta) = E_q[\log p(\theta \mid \alpha)] + E_q[\log p(\mathbf{z} \mid \theta)] + E_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] \\ - E_q[\log q(\theta)] - E_q[\log q(\mathbf{z})] \end{aligned}_f \tag{7}$$

Expand Equation 7 in terms of the model parameters (α, β) and the variational parameters (γ, ϕ) for

further optimization.

$$
\begin{aligned}
\mathcal{L}(\gamma, \phi; \alpha, \beta) &= E_q(\log \Gamma(\sum_{j=1}^{k} \alpha_j)) - E_q(\sum_{i=1}^{k} \log \Gamma(\alpha_i)) + \sum_{i=1}^{k} (\alpha_i - 1) E_q(\log \theta_i) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \int_{\theta_i} q(\theta_i \mid \gamma_i) \log(\theta_i) d\theta_i \\
&\quad + \sum_{n=1}^{N} p(w_n \mid z_n, \beta) q(z_n \mid \phi) \\
&\quad - \int_{\theta} \sum_{\mathbf{z}} q(\theta \mid \gamma) \log q(\theta) d\theta \\
&\quad - \sum_{i=1}^{k} q(z_i \mid \phi) \log q(z_i) \\
&= \log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1) E_q(\log \theta_i \mid \gamma_i) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} E(\log \theta_i \mid \gamma_i) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} (\log(\beta_{ij}) w_n^j) \phi_{ni} \\
&\quad - \int_{\theta} q(\theta \mid \gamma) \log(\frac{\Gamma(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \prod_{i=1}^{k} \theta_i^{\gamma_i - 1}) d\theta \\
&\quad - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log(\phi_{ni}) \\
&= \log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)(\Psi \gamma_i - \Psi \sum_{j=1}^{k} \gamma_i) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} (\Psi \gamma_i - \Psi \sum_{j=1}^{k} \gamma_i) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} (\log(\beta_{iv}) \phi_{ni} \\
&\quad - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)(\Psi \gamma_i - \Psi \sum_{j=1}^{k} \gamma_i) \\
&\quad - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log(\phi_{ni})
\end{aligned}
\tag{8}
$$

*where β_{iv} denotes $p(w_n^v = 1 \mid z^i = 1)$

Minimizing the divergence between the true posterior probability and the variational posterior probability $D(q(\theta, \mathbf{z}) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta))$ is equivalent to maximize the lower bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ with respect to variational parameters.

2.2 Smoothing —— Fuller Bayesian Approach to LDA

In order to cope with the problems that the zero probability is assigned to new documents, the standard approach is to "smooth" the multinomial parameters by assigning positive probability to all vocabulary items no matter whether they are observed or not. Hence, treat β as a $k * V$ random matrix where each row denotes for each mixture component and each row is independently drawn from an exchangeable Dirichlet distribution with single scalar parameter η. The extended LDA graphical model is shown in Figure 3.

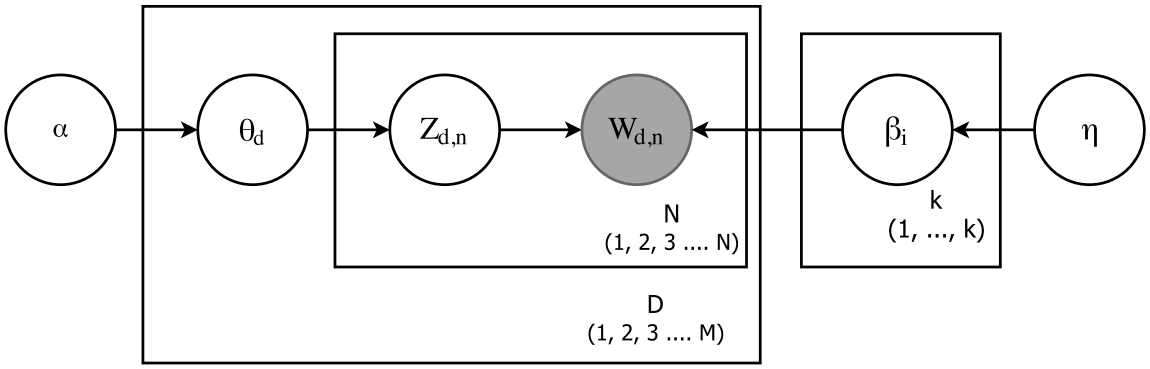

Figure 3: Graphical model representation of smoothed LDA

Parameter	Description
$\beta_i, i \in \{1, ..., k\}$	Random variables that are endowed with a posteior distribution $\sim Dir(\eta)$.
η	Topic hyper-parameter.

Table 4: Smoothed LDA model parameters description

2.2.1 Variational Bayesian Method

In addition to the Variational Inference method (2.1), a new variational parameter λ is assigned for random variables β. The variational distribution with full Bayesian inference is given as:

$$q(\beta_{1:k}, \mathbf{z}_{1:M}, \theta_{1:M}) = \prod_{i=1}^{k} Dir(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta, \mathbf{z}|\phi, \gamma) \tag{9}$$

The new lower bound is given as:

$$\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta) = E_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z}, \beta)] \tag{10}$$

Similarly, additional parameter λ is updated in E-step (3.1) as following equation:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn} \tag{11}$$

Finally, the hyper-parameters η on the exchangeable Dirichlet in the same approach as the updates on parameter α (3.2.2).

3 EM Algorithm for Topics Estimation

We can approximate empirical Bayes estimates for the LDA model via an alternating variational EM procedure that maximize the lower bound (minimize divergence) with respect to the variational parameters γ and ϕ, and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters α and β. Repeat the following iterative algorithm until the lower bound on the log likelihood converges:

1. (E-step) For each document d, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_d^* : d \in \mathcal{D}\}$.

2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. This is corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step.

3.1 E-step: Variational Multinomial

3.1.1 Maximize Equation 8 with respect to ϕ_{ni}

There exists a constraint on ϕ_{ni} that $\sum_{i=1}^{k} \phi_{ni} = 1$. Utilize the Lagrangian by only considering the terms contain ϕ_{ni} with respective Lagrangian multiplier λ_n for n^{th} document:

$$\mathcal{L}_{[\phi_{ni}]} = \phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j)) + \phi_{ni} \log \beta_{iv} - \phi_{ni} \log \phi_{ni} + \lambda_n(\sum_{i=1}^{k} \phi_{ni} - 1) \tag{12}$$

Then, take the first-order derivative of $\mathcal{L}_{[\phi_{ni}]}$ with respect to ϕ_{ni}:

$$\frac{\partial \mathcal{L}_{[\phi_{ni}]}}{\partial \phi_{ni}} = \psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j)) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda_n \tag{13}$$

Maximize by setting the derivative to zero:

$$\log \phi_{ni} = (\psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j)) + \log \beta_{iv} - 1 + \lambda_n \tag{14}$$

It yields the optimal value of ϕ_{ni}:

$$\phi_{ni} \propto \beta_{iv} \exp(\psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j)) \tag{15}$$

3.1.2 Maximize Equation 8 with respect to γ_i, the ith component of the posterior Dirichlet parameter for θ

Isolate the terms containing γ_i:

$$\begin{aligned}
\mathcal{L}_{[\gamma_i]} &= \sum_{i=1}^{k}(\alpha_i - 1)(\Psi\gamma_i - \Psi\sum_{j=1}^{k}\gamma_j) + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\Psi\gamma_i - \Psi(\sum_{j=1}^{k}\gamma_i)) \\
&\quad - \log\Gamma(\sum_{j=1}^{k}\gamma_j) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)(\Psi\gamma_i - \Psi\sum_{j=1}^{k}\gamma_j) \\
&= \sum_{i=1}^{k}(\psi(\gamma_i) - \psi(\sum_{i=1}^{k}))(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i) - \log\Gamma(\sum_{j=1}^{k}\gamma_i) + \sum_{i=1}^{k}\log\Gamma(\gamma_i)
\end{aligned} \tag{16}$$

Then, take the first-order derivative of $\mathcal{L}_{[\gamma_i]}$ with respect to γ_i:

$$\frac{\partial \mathcal{L}_{[\gamma_i]}}{\partial \gamma_i} = \psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \psi'(\sum_{j=1}^{k} \gamma_j) \sum_{j=1}^{k} (\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j) \tag{17}$$

Setting Eq.(17) to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{18}$$

Since Eq.18 depends on the variational multinomial ϕ, full variational inference requires alternating between Eqs. (15) and (18) until the bound $\mathcal{L}$ converges. The procedure is summarized as following:

Algorithm 1 Variational Parameter Estimation Algorithm

Input: Observed words $\mathbf{w}$ in each document
Output: Variational parameters γ and ϕ

initialize $\phi_{ni}^0 := 1/k$ for all i and n
initialize $\gamma_i := \phi_i + N/k$ for all i
repeat
  for $n = 1$ to N **do**
    for $i = 1$ to k **do**
      $\phi_{ni}^{t+1} := \beta_{iw_n} exp(\psi(\gamma_i^t))$
    end for
    normalize ϕ_n^{t+1} to sum to 1.
  end for
  $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
until convergence

In the language of text, the optimizing parameters $(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$ are document-specific. In particular, we view the Dirichlet parameters γ^* as providing a representation of a document in the topic simplex. From the above procedures, each iteration of variational inference for LDA requires $O((N+1)k)$ operations.

3.2 M-step: Maximize Log Likelihood

3.2.1 Conditional Multinomial: maximize with respect to β

There exists a constraint on β that $\sum_{i=1}^{V} \beta_{ij} = 1$. Isolates terms contain β only in $\mathcal{L}$ and add Lagarange multipliers (β is optimized over all documents):

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i (\sum_{j=1}^{V} \beta_{ij} - 1) \tag{19}$$

Then, take the first-order derivative of $\mathcal{L}_{[\beta]}$ with respect to β_{ij}:

$$\frac{\partial \mathcal{L}_{[\beta]}}{\partial \beta_{ij}} = (\sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j) \frac{1}{\beta_{ij}} + \lambda_i \tag{20}$$

Maximizing by setting the derivative yields the optimal value of β_{ij}:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j \tag{21}$$

3.2.2 Estimate the Dirichlet parameter α with all observed documents

Isolate the terms that contain α only:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{M}(\log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}((\alpha_i - 1)(\psi(\gamma_{di}) - \psi(\sum_{j=1}^{k}\gamma_{dj})))) \tag{22}$$

Taking the first-order derivative with respect to α_i:

$$\frac{\partial\mathcal{L}_{[\alpha]}}{\partial\alpha_i} = M(\psi(\sum_{j=1}^{k}\alpha_j) - \psi(\alpha_i)) + \sum_{d=1}^{M}(\psi(\gamma_{di}) - \psi(\sum_{j=1}^{k}\gamma_{dj})) \tag{23}$$

The derivative indicates that optimizing of α_i depends on α_j, where $j \neq i$, thus we must use an iterative method to find the maximal α as elaborated in Section 3.2.3.

3.2.3 Newton-Raphson methods for a Hessian with special structure

This section describes a linear algorithm for the usually cubic Newton-Raphson optimization method. This method is used for estimating the Dirichlet parameter α in M-step.

A fixed-point iteration for maximizing the likelihood can be derived as follows. Given an initial guess for α we construct a simple lower bound on the likelihood which is tight at α. The maximum of this bound is computed in closed-form and it becomes the new guess. Such an iteration is guaranteed to converge to a stationary point of the likelihoodin fact it is the same principle behind the EM algorithm (Minka, 1998). For the Dirichlet, the maximum is the only stationary point. Newton-Raphson finds a stationary point of a function by iterating:

$$\alpha_{new} = \alpha_{old} - \mathbf{H}(\alpha_{old})^{-1}\mathbf{g}(\alpha_{old})$$

where $H(\alpha)$ and $g(\alpha)$ are the Hessian matrix and gradient respectively at the point α.

The second-derivatives, i.e. Hessian matrix, obtained from Eqn.(23) are given by:

$$\frac{\partial\mathcal{L}_{[\alpha]}}{\partial\alpha_i^2} = M(\psi^{'}(\sum_{j=1}^{k}\alpha_j) - \psi^{'}(\alpha_i)) \tag{24}$$

$$\frac{\partial\mathcal{L}_{[\alpha]}}{\partial\alpha_i\partial\alpha_j} = M(\psi^{'}(\sum_{j=1}^{k}\alpha_j) \tag{25}$$

Hence, the Hessian matrix can be written in the form as:

$$\mathbf{H} = \mathbf{Q} + \mathbf{1}\mathbf{1}^T z \tag{26}$$

$$z = M(\psi^{'}(\sum_{j=1}^{k}\alpha_j) \tag{27}$$

$$q_{ij} = -M\psi'(\alpha_i)\delta(i,j), \text{ where } \mathbf{Q} \text{ is a diagonal matrix} \tag{28}$$

Apply the matrix inversion lemma and obtain:

$$\mathbf{H} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}\mathbf{1}\mathbf{1}^T\mathbf{Q}^{-1}}{z^{-1} + \sum_{j=1}^{k} q_{jj}^{-1}} \tag{29}$$

$$(\mathbf{H}^{-1}\mathbf{g})_i = \frac{g_i - c}{h_i} \tag{30}$$

$$\text{where } c = \frac{\sum_{j=1}^{k}\frac{g_j}{h_j}}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}} \tag{31}$$

References

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bouguila et al.2004] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. 2004. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543.

[Minka2000] Thomas Minka. 2000. Estimating a dirichlet distribution.