A note on a Variational Bayes derivation of 
full Bayesian Latent Dirichlet Allocation

Daichi Mochihashi
ATR Spoken Language Translation Research Laboratories,
Kyoto, Japan
daichi.mochihashi@atr.jp
2004.10.20

Abstract
This note provides derivations and formulae for a full Bayesian treatment of
Latent Dirichlet Allocation \[1\], which are mentioned but omitted in its full
paper \[2\]. Further, we extend it a little to accommodate a non-uniform lexical
prior. By using this treatment, we can get appropriately smoothed estimates
of class unigrams \(\theta = p(w_n|z_k)\).

First, we assume that a corpus \(w\) consists of:

\[ w = w_1, w_2, \ldots, w_D \tag{1} \]

\[ w_d = w_1, w_2, \ldots, w_{N_d}. \tag{2} \]

Then, we write a log likelihood for \(w\) given \(\alpha, \eta\), and approximate it as shown below.

\[
\log p(w|\alpha, \eta) = \log \int p(w, \beta|\alpha, \eta) d\beta 
\]

\[
= \log \int \frac{p(w, \beta|\alpha, \eta)}{q(\beta)} q(\beta) d\beta \tag{3} \]

\[
\geq \int q(\beta|\lambda) \log \frac{p(w, \beta|\alpha, \eta)}{q(\beta|\lambda)} d\beta \tag{4} \]

\[
= - \int q(\beta|\lambda) \log q(\beta|\lambda) d\beta + \int q(\beta|\lambda) \log p(w|\alpha, \beta) p(\beta|\eta) d\beta \tag{5} \]

\[
= - \int q(\beta|\lambda) \log q(\beta|\lambda) d\beta + \int q(\beta|\lambda) \log p(w|\alpha, \beta) d\beta + \int q(\beta|\lambda) \log p(\beta|\eta) d\beta \tag{6} \]

\[ \equiv L. \tag{7} \]

Here, \(p(w|\alpha, \beta)\) is a standard model of latent Dirichlet allocation, and decomposed as follows.

\[
\log p(w|\alpha, \beta) = \sum_{d=1}^{D} \log \int \sum_z p(w_d, z, \theta|\alpha, \beta) d\theta 
\]

\[
= \sum_{d=1}^{D} \log \int \sum_z \frac{p(w_d, z, \theta|\alpha, \beta)}{q(z, \theta)} q(z, \theta) d\theta \tag{9} \]

\[
\geq \sum_{d=1}^{D} \int \sum_z q(z, \theta) \log \frac{p(w_d, z, \theta|\alpha, \beta)}{q(z, \theta)} d\theta \tag{10} \]

\[
= \sum_{d=1}^{D} \int \sum_z q(\theta) q(z) \left[ \log p(\theta|\alpha) + \sum_n \log p(z_n|\theta) + \sum_n \log p(w_n|z_n, \beta) \right] d\theta 
\]

\[
- \int \sum_z q(\theta) q(z) \log q(\theta) q(z) d\theta. \tag{12} \]
1 Solution for $q(\beta|\lambda)$

Therefore, we can collect terms that contain $q(\beta|\lambda)$ from $L$, and apply Lagrangians:

\[
\begin{align*}
J(\beta) &= -\int \prod_k q(\beta_k|\lambda_k) \sum_k \log q(\beta_k|\lambda_k) d\beta \\
&\quad + \sum_d \int \prod_k q(\beta_k|\lambda_k) \left[ \sum_z q(z) \sum_n \log p(w_n|z_n,\beta) \right] d\beta \\
&\quad + \int \prod_k q(\beta_k|\lambda_k) \sum_k \log p(\beta_k|\eta) d\beta \\
&\quad + \sum_k \mu_k \left( \int q(\beta_k|\lambda_k)d\beta_k - 1 \right)
\end{align*}
\]

(13)

\[
\begin{align*}
\therefore \frac{\partial J(\beta)}{\partial \beta_k} &= -\int \frac{\partial}{\partial \beta_k} \prod_k q(\beta_k|\lambda_k) \log q(\beta_k|\lambda_k) d\beta \\
&\quad + \mu_k \\
&\quad + \log p(\beta_k|\eta) \\
&\quad + \sum_d \sum_z q(z) \sum_n \log p(w_n|z_n,\beta) \\
&= -\log q(\beta_k|\lambda_k) + \mu_k + \log p(\beta_k|\eta) \\
&\quad + \sum_{d=1}^D \sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v \log \beta_{kv}
\end{align*}
\]

(14)

\[
\begin{align*}
&= 0.
\end{align*}
\]

(15)

(16)

Then,

\[
\begin{align*}
\log q(\beta_k|\lambda_k) &= \mu_k + \log p(\beta_k|\eta) + \sum_{d=1}^D \sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v \log \beta_{kv}
\end{align*}
\]

(17)

\[
\begin{align*}
\therefore q(\beta_k|\lambda_k) \propto p(\beta_k|\eta) \exp(\sum_{d=1}^D \sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v \log \beta_{kv}) \\
&= p(\beta_k|\eta) \cdot \beta_{d1}^\sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v \\
&\propto \text{Dir}(\beta_k|\eta + \sum_{d=1}^D \sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v)
\end{align*}
\]

(18)

(19)

(20)

\[
\begin{align*}
\iff \lambda_k &= \eta + \sum_{d=1}^D \sum_{n=1}^N \sum_{v=1}^V \phi_{dnk} w_{dn}^v.
\end{align*}
\]

(21)
2 Newton-Raphson iteration for $\eta$

Here, we derive a Newton-Raphson iteration for $\eta$, a hyperparameter that works as a smoothing term for $\beta$.

First, we extract a term that contains $\eta$, from $L$:

$$L_{\eta} = \int q(\beta|\lambda) \log p(\beta|\eta) d\beta$$

$$= \int \prod_k q(\beta_k|\lambda_k) \sum_k \log p(\beta_k|\eta) d\beta$$

$$= \sum_k \int q(\beta_k|\lambda_k) \log \left( \frac{\Gamma(V\eta)}{\Gamma(\eta)} \prod_v \beta_{kv}^{\eta-1} \right) d\beta_k$$

$$= \sum_k \left[ \log \Gamma(V\eta) - V \log \Gamma(\eta) + \int \text{Dir}(\beta_k|\lambda_k) \sum_v (\eta - 1) \log \beta_{kv} d\beta_k \right]$$

$$= K \left( \log \Gamma(V\eta) - V \log \Gamma(\eta) + \sum_k \sum_v (\eta - 1) \int \text{Dir}(\beta_k|\lambda_k) \log \beta_{kv} d\beta_k \right)$$

$$= K \left( \log \Gamma(V\eta) - V \log \Gamma(\eta) + (\eta - 1) \sum_k \sum_v \{ \Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv}) \} \right)$$

We denote $\sum_k \sum_v \Psi(\lambda_{kv}) = \Psi(\sum_v \lambda_{kv})$ as $P$. Then,

$$\frac{\partial L_{\eta}}{\partial \eta} = K \left( \log \Gamma(V\eta) - V \log \Gamma(\eta) + \right) + P. \quad (28)$$

Here, we can derive a Newton-Raphson update for scalar hyperparameter $\eta$.

$$\therefore \quad \eta^\text{new} = \eta - \frac{K \left( \log \Gamma(V\eta) - V \log \Gamma(\eta) \right) + (\eta - 1) \cdot P}{K \left( \log \Gamma(V\eta) - V \log \Gamma(\eta) \right) + P} \quad (29)$$

$$= \eta - \frac{\log \Gamma(V\eta) - V \log \Gamma(\eta) + (\eta - 1) \cdot P/K}{\Psi(V\eta) - \Psi(\eta) + P/(KV)} \quad (30)$$

$$= \eta - \frac{\log \Gamma(V\eta)/V - \log \Gamma(\eta) + (\eta - 1) \cdot P/(KV)}{\Psi(V\eta) - \Psi(\eta) + P/(KV)}. \quad \blacksquare \quad (31)$$
3 Newton-Raphson iteration for $\eta$ (extended)

Equation (31) is the update formula for scalar $\eta$, that is mentioned but omitted in [2]. However, this means we are doing a generalized Laplace smoothing (Lidstone’s law) [3]; that is, it gives a \textit{uniform} smoothing term to class unigrams $k_v$, no matter what word $v$ is.

Apparently, this is not an adequate approach to smoothing. Instead, when we introduce a vector hyperparameter $\eta = (\eta_1, \eta_2, \ldots, \eta_v)$ and assume a prior distribution $p(\beta|\eta) = \text{Dir}(\beta|\eta_v)$, we get another Bayesian estimate of $\beta$, appropriately smoothed word by word.

Fortunately, inferring $\eta$ can be done by a linear-time Newton-Raphson iteration, as shown below.

$$L_\eta = \int q(\beta|\lambda) \log p(\beta|\eta) d\beta$$

$$= \prod_k q(\beta_k|\lambda_k) \log p(\beta_k|\eta) d\beta_k$$

$$= \sum_k \int q(\beta_k|\lambda_k) \log \left( \frac{\Gamma(\sum_v \eta_v)}{\prod_v \Gamma(\eta_v)} \prod_v \beta_k^{-1} \right) d\beta_k$$

$$= \sum_k \left[ \log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) + \int \text{Dir}(\beta_k|\lambda_k) \sum_v (\eta_v - 1) \log \beta_k d\beta_k \right]$$

$$= K \left( \log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) + \sum_v (\eta_v - 1) \sum_k \left( \psi(\lambda_k) - \psi(\sum_v \lambda_k) \right) \right)$$

$$= K \left( \log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) + \sum_v (\eta_v - 1) \sum_{\lambda_k} \left( \psi(\lambda_k) - \psi(\sum_v \lambda_k) \right) \right)$$

We denote $\sum_k \psi(\lambda_k) - \psi(\sum_v \lambda_k)$ as $P_v$.

Then,

$$\frac{\partial L_\eta}{\partial \eta_i} = K \left( \psi(\sum_i \eta_i) - \psi(\eta_i) \right) + P_i \equiv g(\eta)$$

$$\frac{\partial^2 L_\eta}{\partial \eta_i \partial \eta_j} = \begin{cases} K \psi'((\sum_i \eta_i) - K \psi'(\eta_i) & \text{if } i = j \\ K \psi'(\sum_i \eta_i) & \text{otherwise} \end{cases}$$

Therefore, the Hessian is of the form

$$H = K \cdot (\text{diag}(h) + 1z1^T)$$

where

$$h_i = -\psi'(\eta_i)$$

$$z = \psi'((\sum_i \eta_i))$$

So we can derive a linear-time Newton-Raphson iteration as outlined in [2], as follows.

$$\eta_{\text{new}} = \eta - H(\eta)^{-1}g(\eta)$$

$$(H^{-1}g)_v = \frac{1}{K} \cdot \frac{e - P_v + K \left( \psi'(\eta_v) - \psi'(\sum_v \eta_v) \right)}{\psi'(\eta_v)}$$

$$c = \sum_v \frac{\left( K \psi'(\eta_v) - \psi'(\sum_v \eta_v) \right) - P_v}{\psi'(\sum_v \eta_v)} / \psi'(\eta_v)$$

$$\psi'(\sum_v \eta_v) = \sum_v \psi'(\eta_v)$$

$$\eta_{\text{new}} = \eta - \frac{1}{K} \cdot \frac{e - P_v + K \left( \psi'(\eta_v) - \psi'(\sum_v \eta_v) \right)}{\psi'(\eta_v)} \psi'(\eta_v)$$

$$c = \sum_v \frac{K \psi'(\eta_v) - \psi'(\sum_v \eta_v)}{\psi'(\sum_v \eta_v) - \sum_v \psi'(\eta_v)}$$

4
References

