Appendix for Hierarchical Dirichlet Scaling Process for Multi-labeled Data

Dongwoo Kim
KAIST, Daejeon, Korea

Alice Oh
KAIST, Daejeon, Korea

This appendix has been provided to give readers additional information about the HDSP. In this appendix, we provide the detailed derivation of variational inference, the posterior word count analysis, and more examples from the Wikipedia and OHSUMED corpora.

A. Variational inference for HDSP

In this section, we provide the detailed derivation for mean-field variational inference for HDSP. First, the evidence of lower bound for HDSP is obtained by taking a Jensen’s inequality on the marginal log likelihood of the observed data,

\[
\ln \int p(D, \Theta) d\Theta \geq \int Q(\Psi) \ln \frac{p(D, \Theta)}{Q(\Psi)} d\Theta, \tag{1}
\]

where \(D\) is the training set of documents and labels. \(\Psi\) denotes the set of variational parameters, \(\Theta\) denotes the set of model parameters.

We define a fully factorized variational distribution \(Q\) as follows:

\[
Q := \prod_{k=1}^{T} q(\phi_k) q(V_k) \prod_{j=1}^{J} q(w_{jk}) \prod_{m=1}^{M} q(\pi_{mk}) \prod_{n=1}^{N_m} q(z_{mn}), \tag{2}
\]

where

\[
q(z_{mn}) = \text{Multinomial}(z_{mn} | \gamma_{mn1}, \gamma_{mn2}, ..., \gamma_{mNT})
\]
\[
q(\pi_{mk}) = \text{Gamma}(\pi_{mk} | a_{\pi_{mk}}, b_{\pi_{mk}})
\]
\[
q(w_{jk}) = \text{InvGamma}(w_{jk} | a_{w_{jk}}, b_{w_{jk}})
\]
\[
q(\phi_k) = \text{Dirichlet}(\phi_k | \eta_{k1}, \eta_{k2}, ..., \eta_{kI})
\]
\[
q(V_k) = \delta_{V_k}.
\]
The evidence of lower bound (ELBO) is

\[
L(D, \Psi) = \mathbb{E}_Q[\ln p(D, \Psi)] + \mathbb{H}[Q]
\]

\[
= \mathbb{E}_Q[\sum_{m=1}^{M} \sum_{n=1}^{N_m} \ln p(x_{mn}, z_{mn}, \Phi)] + \mathbb{E}_Q[\sum_{m=1}^{M} \sum_{n=1}^{N_m} \ln p(z_{mn} | \pi_m)]
\]

\[
+ \mathbb{E}_Q[\sum_{m=1}^{M} \sum_{k=1}^{\infty} \ln \psi(\pi_{mk} | V_k, w_k, r_m)] + \sum_{m=1}^{M} \sum_{k=1}^{\infty} \mathbb{E}_Q[\ln \psi(V_k | \alpha)] + \sum_{j=1}^{J} \sum_{k=1}^{\infty} \mathbb{E}_Q[\ln p(w_{jk} | a^w, b^w)]
\]

\[
+ \mathbb{E}_Q[\ln \psi(\phi_k | \eta)] - \mathbb{E}_Q[\ln Q]
\]

where the expectations of latent variables under the variational distribution \( Q \) are

\[
\mathbb{E}_Q[\pi_{mk}] = \frac{a^\pi_{mk}}{b^\pi_{mk}}
\]

\[
\mathbb{E}_Q[\ln \pi_{mk}] = \psi(\frac{a^\pi_{mk}}{b^\pi_{mk}}) - \ln b^\pi_{mk}
\]

\[
\mathbb{E}_Q[w_{jk}] = \frac{b^w_{jk}}{a^w_{jk} - 1}
\]

\[
\mathbb{E}_Q[w_{jk}^{-1}] = \frac{a^w_{jk}}{b^w_{jk}}
\]

\[
\mathbb{E}_Q[\ln w_{jk}] = \ln b^w_{jk} - \psi(\frac{a^w_{jk}}{b^w_{jk}})
\]

\[
\mathbb{E}_Q[\ln \psi(\phi_k | \eta)] = \psi(\eta_{kd}) - \psi\left(\sum_i \psi(\eta_{ki})\right)
\]

Then, we derive the equations further

\[
L(D, \Psi) = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{T} \gamma_{mnk} \{\psi(\eta_{kd}) - \psi(\sum_i \eta_{ki})\}
\]

\[
+ \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{T} \gamma_{mnk} \{\mathbb{E}_Q[\ln \pi_{mk}] - \mathbb{E}_Q[\ln \sum_{k=1}^{T} \pi_{mk}]\}
\]

\[
+ \sum_{m=1}^{M} \sum_{j=1}^{T} \alpha_{mk} \{\ln(b^w_{jk}) - \psi(\frac{a^w_{jk}}{b^w_{jk}})\} + (\beta_{mk} - 1) \{\psi(\frac{a^\pi_{mk}}{b^\pi_{mk}}) - \ln(b^\pi_{mk})\} - \prod_j \left(\frac{a^w_{jk}}{b^w_{jk}}\right)^{r_{mj}} \frac{a^\pi_{mk}}{b^\pi_{mk}} - \ln \Gamma(\beta_{mk})
\]

\[
+ \sum_{k=1}^{T} \ln \Gamma(\alpha + 1) - \ln \Gamma(\alpha) + (\alpha - 1) \ln(1 - V_k)
\]

\[
+ \sum_{k=1}^{T} a^w \ln b^w - \ln \Gamma(a^w) - (a^w + 1) \{\ln b^w - \psi(a^w)\} - a^w - \mathbb{E}_Q[\ln Q].
\]

Taking the derivatives of this lower bound with respect to each variational parameter, we can obtain the coordinate ascent updates.

The optimal form of the variational distribution can be obtained by exponentiating the variational lower bound with all expectations except the parameter of interest (Bishop & Nasrabadi, 2006). For \( \pi_{mk} \), we can derive the optimal form of
Hierarchical Dirichlet Scaling Process

variational distribution as follows

\[
q(\pi_{mk}) \propto \exp \left\{ \mathbb{E}_{q-\pi_{mk}} [\ln p(\pi_{mk}|z, \pi_{-mk}, r_m, V)] \right\} \\
\propto \exp \left\{ \mathbb{E}_{q-\pi_{mk}} [\ln p(z|\pi_m) + \ln p(\pi_m|r_m, V)] \right\} \\
\propto Z^{\beta_p} + \sum_{n=1}^{N_m} \gamma_{mk} b_{mk}^{\pi_m} \prod_j \mathbb{E}_q[w_{jk}^{-r_{mj}}] + \frac{N_m}{\xi_m}
\]

where update for $\xi_m$ is $- \ln \xi_m - (\sum_{k=1}^{T} \mathbb{E}_q[\pi_{mk}] - \xi_m)/\xi_m$. Therefore, the optimal form of variational distribution for $\pi_{mk}$ is

\[
q(\pi_{mk}) \sim \text{Gamma}(\beta_p + \sum_{n=1}^{N_m} \gamma_{mk}, b_{mk}^{\pi_m} \prod_j \mathbb{E}_q[w_{jk}^{-r_{mj}}] + \frac{N_m}{\xi_m}).
\]

We take the same approach described in (Paisley et al., 2012), and the only difference comes from the product of the inverse distance term.

For $w_{jk}$, we can derive the optimal form of the variational distribution as follows

\[
q(w_{jk}) \propto \exp \left\{ \mathbb{E}_{q-w_{jk}} [\ln p(w_{jk}|\pi, w_{-jk}, a^w, b^w)] \right\} \\
\propto \exp \left\{ \mathbb{E}_{q-w_{jk}} \left[ \sum_{m=1}^{M} \ln p(\pi_{mk}|\beta_p, \prod_{j=1}^{J} w_{jk}^{-r_{mj}}) + \ln p(w_{jk}|a^w, b^w) \right] \right\} \\
\propto \exp \left\{ \mathbb{E}_{q-w_{jk}} \left[ -\beta_p \sum_{m=1}^{M} r_{mj} \ln w_{jk} - \sum_{m} \prod_{j} w_{jk}^{-r_{mj}} \pi_{mk} - (a^w + 1) \ln w_{jk} - \frac{b^w}{w_{jk}} \right] \right\} \\
\propto w_{jk}^{-\mathbb{E}_q[\beta_p] \sum_{m=1}^{M} r_{mj} - a^w - 1} e^{- (\sum_{m:r_{mj}=1} \prod_{j':r_{mj'}=1} \mathbb{E}_q[w_{jk}^{-r_{mj}}] \mathbb{E}_q[\pi_{mk}]-b^w)} w_{jk}^{r_{mj}}
\]

Therefore, the optimal form of variational distribution for $w_{jk}$ is

\[
q(w_{jk}) \sim \text{InvGamma}(\mathbb{E}_q[\beta_p] \sum_{m} r_{mj} + a^w, \sum_{m' j'/j} \mathbb{E}_q[w_{jk}^{-r_{mj}}] \mathbb{E}_q[\pi_{mk}'] + b^w)
\]

where $m' = \{ m : r_{mj} = 1 \}$ and $j'/j = \{ j' : r_{mj'} = 1, j' \neq j \}$.

B. Posterior Word Count

Like the HDP and other nonparametric topic models, our model also uses only a few topics even though we set the truncation level to 200. Figure 1 shows the posterior word count for the different values of the Dirichlet topic parameter $\eta$. As the result indicates our model uses 50 to 100 topics. The HDSP tends to use more topics than the HDP.

C. More Results

Figure 2 shows more examples of the expected topic distribution given label from the Wikipedia and OHSUMED corpora. We show the top 20 topics each with the top 10 words.

References


Figure 1. Posterior number of words. We set the truncation level to 200, but only a few topics are used during inference.
Figure 2. Expected topic distributions from Wikipedia and OHSUMED. From top to bottom, we compute the expected topic distribution given a label.