Appendix for Hierarchical Dirichlet Scaling Process for Multi-labeled Data

Dongwoo Kim DW.KIM@KAIST.AC.KR

KAIST, Daejeon, Korea

Alice Oh ALICE.OH@KAIST.EDU

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,

$$\ln \int p(\mathcal{D}, \Theta) d\Theta \geq \int Q(\Psi) \ln \frac{P(\mathcal{D}, \Theta)}{Q(\Psi)} d\Theta, \tag{1}$$

where $\mathcal{D}$ is the training set of documents and labels. Ψ denotes the set of variational parameters, Θ 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

$$\begin{aligned}
q(z_{mn}) &= \text{Multinomial}(z_{mn} | \gamma_{mn1}, \gamma_{mn2}, ..., \gamma_{mnT}) \\
q(\pi_{mk}) &= \text{Gamma}(\pi_{mk} | a^{\pi}_{mk}, b^{\pi}_{mk}) \\
q(w_{jk}) &= \text{InvGamma}(w_{jk} | a^{w}_{jk}, b^{w}_{jk}) \\
q(\phi_k) &= \text{Dirichlet}(\phi_k | \eta_{k1}, \eta_{k2}, ..., \eta_{kI}) \\
q(V_k) &= \delta_{V_k}.
\end{aligned} \tag{3}$$

The evidence of lower bound (ELBO) is

$$
\begin{aligned}
L(\mathcal{D},\Psi) &= \mathbb{E}_q[\ln p(\mathcal{D},\Psi)] + \mathbb{H}[Q] \qquad (4)\\
&= \mathbb{E}_q[\sum_{m=1}^{M}\sum_{n=1}^{N_m}\ln p(\mathrm{x}_{mn}|z_{mn},\Phi)] + \mathbb{E}_q[\sum_{m=1}^{M}\sum_{n=1}^{N_m}\ln p(z_{mn}|\pi_m)]\\
&\quad + \mathbb{E}_q[\sum_{m=1}^{M}\sum_{k=1}^{\infty}\ln p(\pi_{mk}|V_k,w_k,\mathbf{r}_m)] + \mathbb{E}_q[\sum_{k=1}^{\infty}\ln p(V_k|\alpha)] + \mathbb{E}_q[\sum_{j=1}^{J}\sum_{k=1}^{\infty}\ln p(w_{jk}|a^w,b^w)]\\
&\quad + \mathbb{E}_q[\sum_{k=1}^{\infty}\ln p(\phi_k|\eta)] - \mathbb{E}_q[\ln Q]\\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T}\gamma_{mnk}\mathbb{E}_q[\ln p(\mathrm{x}_{mn}|\phi_k)] + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{T}\gamma_{mnk}\mathbb{E}_q[\ln p(z_{mn}=k|\pi_m)]\\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{T}\mathbb{E}_q[\ln p(\pi_{mk}|V_k,w_k,\mathbf{r}_m)] + \sum_{k=1}^{T}\mathbb{E}_q[\ln p(V_k|\alpha)] + \sum_{k=1}^{\infty}\sum_{j=1}^{J}\mathbb{E}_q[\ln p(w_{jk}|a^w,b^w)] - \mathbb{E}_q[\ln Q]
\end{aligned}
$$

where the expectations of latent variables under the variational distribution Q are

$$
\begin{aligned}
&\mathbb{E}_q[\pi_{mk}] = a^{\pi}_{mk}/b^{\pi}_{mk}\\
&\mathbb{E}_q[\ln\pi_{mk}] = \psi(a^{\pi}_{mk}) - \ln b^{\pi}_{mk}\\
&\mathbb{E}_q[w_{jk}] = b^{w}_{jk}/(a^{w}_{jk}-1)\\
&\mathbb{E}_q[w^{-1}_{jk}] = a^{w}_{jk}/b^{w}_{jk}\\
&\mathbb{E}_q[\ln w_{jk}] = \ln b^{w}_{jk} - \psi(a^{w}_{jk})\\
&\mathbb{E}_q[\ln\phi_{ki}] = \psi(\eta_{kd}) - \psi(\sum_{i}^{\prime}\psi_{ki'})
\end{aligned}
$$

Then, we derive the equations further

$$
\begin{aligned}
L(\mathcal{D},\Psi) = &\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T}\gamma_{mnk}\{\psi(\eta_{k\mathrm{x}_{mn}}) - \psi(\sum_{d}\eta_{kd})\} \qquad (5)\\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T}\gamma_{mnk}\{\mathbb{E}_q[\ln\pi_{mk}] - \mathbb{E}_q[\ln\sum_{k=1}^{T}\pi_{mk}]\}\\
&+ \sum_{m=1}^{M}\sum_{k=1}^{T} -\beta p_k\sum_{j}r_{mj}\{\ln(b^{w}_{jk}) - \psi(a^{w}_{jk})\} + (\beta p_k - 1)\{\psi(a^{\pi}_{mk}) - \ln(b^{\pi}_{mk})\} - \prod_{j}\left(\frac{a^{w}_{jk}}{b^{w}_{jk}}\right)^{r_{mj}}\frac{a^{\pi}_{mk}}{b^{\pi}_{mk}} - \ln\Gamma(\beta p_k)\\
&+ \sum_{k=1}^{T}\ln\Gamma(\alpha+1) - \ln\Gamma(\alpha) + (\alpha-1)\ln(1-V_k)\\
&+ \sum_{k=1}^{T}\sum_{j=1}^{J}a^{w}\ln b^{w} - \ln\Gamma(a^{w}) - (a^{w}+1)\{\ln b^{w} - \psi(a^{w})\} - a^{w} - \mathbb{E}_q[\ln Q].
\end{aligned}
$$

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 π_{mk}, we can derive the optimal form of variational distribution as follows

$$\begin{aligned} q(\pi_{mk}) &\propto \exp\left\{\mathbb{E}_{q_{-\pi_{mk}}}[\ln p(\pi_{mk}|z,\pi_{-mk},\mathbf{r}_m,V)]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-\pi_{mk}}}[\ln p(z|\pi_m)+\ln p(\pi_m)|\mathbf{r}_m,V)]\right\} \\ &\propto Z^{\beta p_k+\sum_{n=1}^{N_m}\gamma_{mnk}} e^{b^{\pi}_{mk}\prod_j \mathbb{E}_q[w_{jk}^{-r_{mj}}]+\frac{N_m}{\xi_m}} \end{aligned} \tag{6}$$

where update for ξ_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 π_{mk} is

$$q(\pi_{mk})\sim \text{Gamma}(\beta p_k+\sum_{n=1}^{N_m}\gamma_{mnk},\quad b^{\pi}_{mk}\prod_j \mathbb{E}_q[w_{jk}^{-r_{mj}}]+\frac{N_m}{\xi_m}). \tag{7}$$

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

$$\begin{aligned} q(w_{jk}) &\propto \exp\left\{\mathbb{E}_{q_{-w_{jk}}}[\ln p(w_{jk}|\pi,w_{-jk},a^w,b^w)]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-w_{jk}}}[\sum_{m=1}^{M}\ln p(\pi_{mk}|\beta p_k,\prod_{j=1}^{J}w_{jk}^{-r_{mj}})+\ln p(w_{jk}|a^w,b^w)]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-w_{jk}}}[-\beta p_k\sum_{m=1}^{M}r_{mj}\ln w_{jk}-\sum_m\prod_j w_{jk}^{-r_{mj}}\pi_{mk}-(a^w+1)\ln w_{jk}-\frac{b^w}{w_{jk}}]\right\} \\ &\propto w_{jk}^{-\mathbb{E}_q[\beta p_k]\sum_{m=1}^{M}r_{mj}-a^w-1} e^{(-\sum_{\{m:r_{mj}=1\}}\prod_{\{j':r_{mj'}=1/j\}}\mathbb{E}_q[w_{j'k}^{-1}]\mathbb{E}_q[\pi_{mk}]-b^w)\frac{1}{w_{jk}}} \end{aligned} \tag{8}$$

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

$$q(w_{jk})\sim \text{InvGamma}(\mathbb{E}_q[\beta p_k]\sum_m r_{mj}+a^w,\quad \sum_{m'}\prod_{j'/j}\mathbb{E}_q[w_{j'k}^{-1}]\mathbb{E}_q[\pi_{m'k}]+b^w) \tag{9}$$

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 η. As the result indicates our model uses 50 to 100 topics. The HDSP tends to use more topics than the HDP.

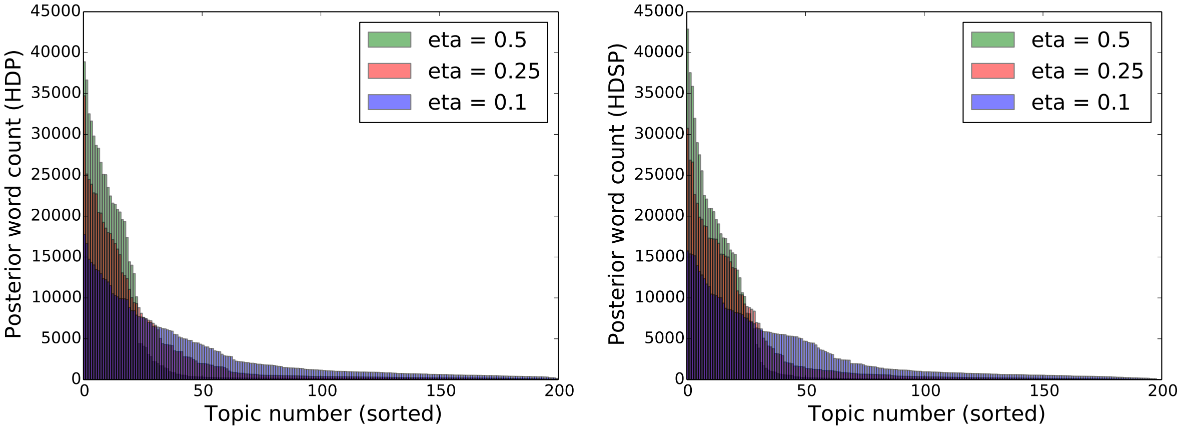
C. More Results

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

Bishop, Christopher M and Nasrabadi, Nasser M. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

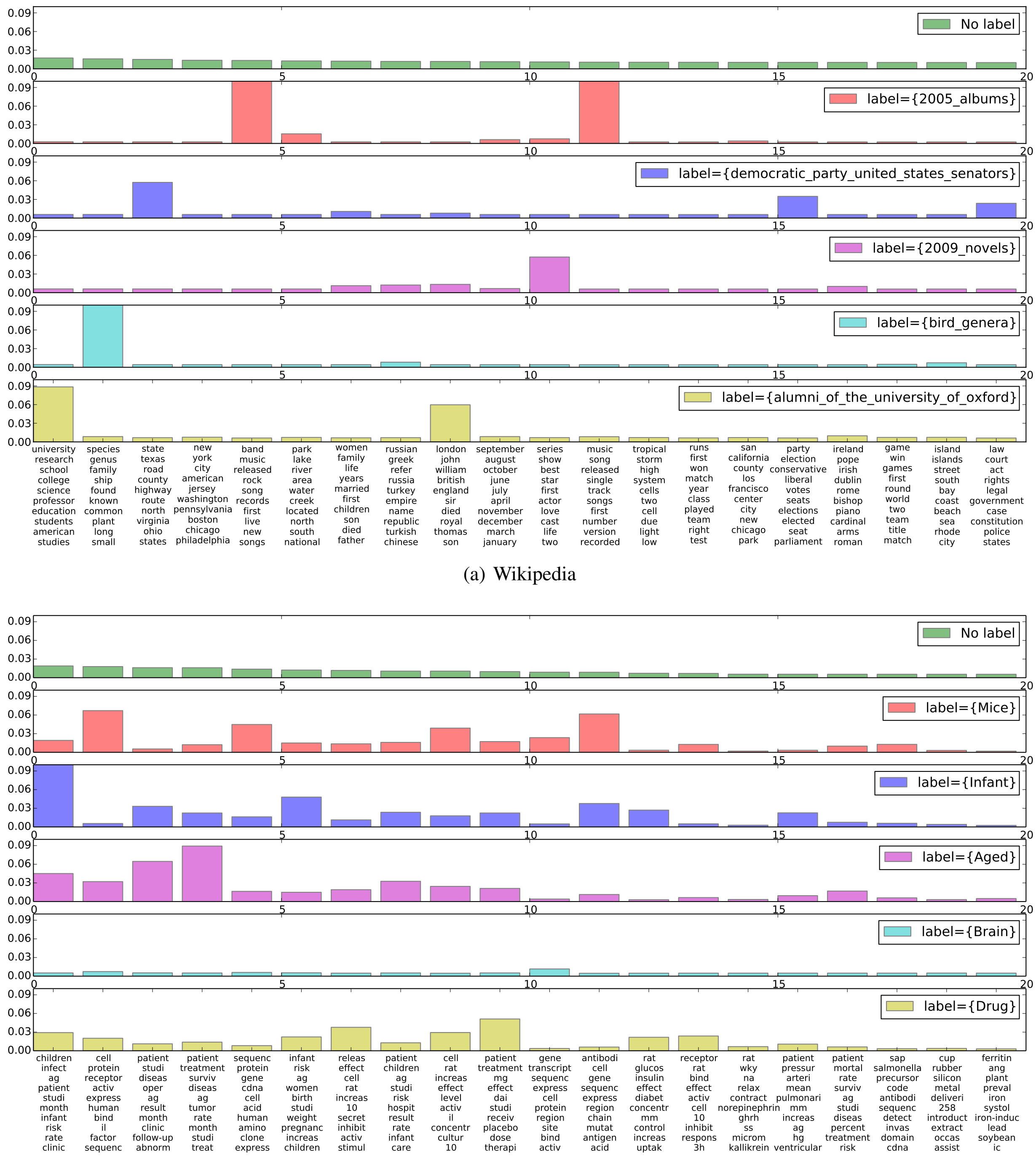
Paisley, John, Wang, Chong, and Blei, David M. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7 (4):997–1034, 2012.



(a) HDP

(b) HDSP

Figure 1. Posterior number of words. We set the truncation level to 200, but only a few topics are used during inference.



(a) Wikipedia

(b) OHSUMED

Figure 2. Expected topic distributions from Wikipedia and OHSUMED. From top to bottom, we compute the expected topic distribution given a label.