Sampling Equation Derivation of LBH-RTM and LBS-RTM

Weiwei Yang
Computer Science
University of Maryland
College Park, MD
wwyang@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber@colorado.edu

Philip Resnik
Linguistics and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

1 Sampling Block Assignments

The joint probability of ALL link weights $\boldsymbol{A}$ and document block assignments $\boldsymbol{y}$ is

$$\Pr(\boldsymbol{A}, \boldsymbol{y} \,|\, a, b, \gamma) = \Pr(\boldsymbol{A} \,|\, \boldsymbol{y}, a, b) \Pr(\boldsymbol{y} \,|\, \gamma). \tag{1}$$

1.1 Undirected Links

We further expand $\Pr(\boldsymbol{A}, \boldsymbol{y} \,|\, a, b, \gamma)$ for undirected graph as

$$\begin{aligned}
& \Pr(\boldsymbol{A}, \boldsymbol{y} \,|\, a, b, \gamma) & (2)\\
= & \int\int \Pr(\boldsymbol{A} \,|\, \boldsymbol{y}, \boldsymbol{\Omega}) \Pr(\boldsymbol{\Omega} \,|\, a, b) \Pr(\boldsymbol{y} \,|\, \boldsymbol{\mu}) \Pr(\boldsymbol{\mu} \,|\, \gamma) \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} & (3)\\
= & \int\int \prod_{l \le l'} \prod_{d \in l, d' \in l'} \frac{\Omega_{l,l'}^{A_{d,d'}}}{A_{d,d'}!} \exp\left(-\Omega_{l,l'}\right) \prod_{l \le l'} \frac{b^a}{\Gamma(a)} \Omega_{l,l'}^{a-1} \exp\left(-b\Omega_{l,l'}\right) \prod_{l=1}^{L} \mu_l^{N_l} \frac{1}{\Delta(\gamma)} \prod_{l=1}^{L} \mu_l^{\gamma-1} \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} & (4)\\
\propto & \int\int \prod_{l \le l'} \Omega_{l,l'}^{S_w(l,l')+a-1} \exp\left(-(S_e(l,l')+b)\Omega_{l,l'}\right) \prod_{l=1}^{L} \mu_l^{N_l+\gamma-1} \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} & (5)\\
\propto & \Delta(\boldsymbol{N_l} + \gamma) \prod_{l \le l'} \frac{\Gamma(S_w(l,l') + a)}{(S_e(l,l') + b)^{S_w(l,l')+a}}, & (6)
\end{aligned}$$

where $S_w(l, l')$ is the weight sum of OBSERVED links between blocks l and l'; $S_e(l, l')$ is the number of ALL POSSIBLE links (i.e. assuming all links are observed) between blocks l and l'. Specifically, $S_e(l, l')$ is defined as

$$S_e(l, l') = \begin{cases} N_l \times N_{l'} & l \ne l' \\ \frac{1}{2} N_l (N_l - 1) & l = l' \end{cases} \tag{7}$$

where N_l denotes the number of documents assigned the block l.

$\Delta(\boldsymbol{N_l} + \gamma)$ is defined as

$$\Delta(\boldsymbol{N_l} + \gamma) = \frac{\prod_{l=1}^{L} \Gamma(N_l + \gamma)}{\Gamma(\sum_{l=1}^{L} N_l + L\gamma)}. \tag{8}$$

We then derive the Gibbs sampling equation for document d, given the block assignments of other documents and link weights excluding d, as

$$
\begin{aligned}
& \Pr\left(y_d = l \mid \boldsymbol{A}_{-d}, \boldsymbol{y}_{-d}, a, b, \gamma\right) & (9)\\
= & \frac{\Pr\left(\boldsymbol{A}, \boldsymbol{y} \mid a, b, \gamma\right)}{\Pr\left(\boldsymbol{A}_{-d}, \boldsymbol{y}_{-d} \mid a, b, \gamma\right)} & (10)\\
\propto & \prod_{l'=1}^{L} \frac{\Gamma(S_w(l,l')+a)}{(S_e(l,l')+b)^{S_w(l,l')+a}} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{\Gamma(S_w^{-d}(l,l')+a)} \frac{\Gamma(D-1+L\gamma)}{\Gamma(D+L\gamma)} \frac{\Gamma(N_l+\gamma)}{\Gamma(N_l^{-d}+\gamma)} & (11)\\
\propto & \prod_{l'=1}^{L} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{(S_e(l,l')+b)^{S_w(l,l')+a}} \prod_{i=0}^{S_w(d,l')-1} (S_w^{-d}(l,l')+a+i) \frac{N_l^{-d}+\gamma}{D-1+L\gamma} & (12)\\
\propto & (N_l^{-d}+\gamma) \prod_{l'=1}^{L} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{(S_e^{-d}(l,l')+b+S_e(d,l'))^{S_w^{-d}(l,l')+a+S_w(d,l')}} \prod_{i=0}^{S_w(d,l')-1} (S_w^{-d}(l,l')+a+i), & (13)
\end{aligned}
$$

where $S_w(d,l')$ denotes the weight sum of OBSERVED links between document d and block l'; $S_e(d,l')$ denotes the number of ALL POSSIBLE links between document d and block l'. Namely, $S_e(d,l') = N_{l'}$.

1.2 Directed Links

The expansion of $\Pr\left(\boldsymbol{A}, \boldsymbol{y} \mid a, b, \gamma\right)$ for directed graph is

$$
\begin{aligned}
& \Pr\left(\boldsymbol{A}, \boldsymbol{y} \mid a, b, \gamma\right) & (14)\\
\propto & \int\!\!\int \prod_{l,l'} \prod_{d \in l, d' \in l'} \frac{\Omega_{l,l'}^{A_{d,d'}}}{A_{d,d'}!} \exp\left(-\Omega_{l,l'}\right) \prod_{l,l'} \frac{b^a}{\Gamma(a)} \Omega_{l,l'}^{a-1} \exp\left(-b\Omega_{l,l'}\right) \prod_{l=1}^{L} \mu_l^{N_l} \frac{1}{\Delta(\gamma)} \prod_{l=1}^{L} \mu_l^{\gamma-1} \mathrm{d}\boldsymbol{\Omega}\,\mathrm{d}\boldsymbol{\mu} & (15)\\
\propto & \Delta(\boldsymbol{N_l}+\gamma) \prod_{l,l'} \frac{\Gamma(S_w(l,l')+a)}{(S_e(l,l')+b)^{S_w(l,l')+a}}, & (16)
\end{aligned}
$$

where $S_e(l,l')$ is defined as

$$
S_e(l,l') = \begin{cases} N_l \times N_{l'} & l \neq l' \\ N_l(N_l-1) & l = l' \end{cases} \qquad (17)
$$

The Gibbs sampling equation is derived as

$$
\begin{aligned}
& \Pr\left(y_d = l \mid \boldsymbol{A}_{-d}, \boldsymbol{y}_{-d}, a, b, \gamma\right) & (18)\\
= & \frac{\Pr\left(\boldsymbol{A}, \boldsymbol{y} \mid a, b, \gamma\right)}{\Pr\left(\boldsymbol{A}_{-d}, \boldsymbol{y}_{-d} \mid a, b, \gamma\right)} & (19)\\
\propto & \prod_{l'=1, l'\neq l}^{L} \frac{\Gamma(S_w(l,l')+a)}{(S_e(l,l')+b)^{S_w(l,l')+a}} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{\Gamma(S_w^{-d}(l,l')+a)} \\
& \prod_{l'=1, l'\neq l}^{L} \frac{\Gamma(S_w(l',l)+a)}{(S_e(l',l)+b)^{S_w(l',l)+a}} \frac{(S_e^{-d}(l',l)+b)^{S_w^{-d}(l',l)+a}}{\Gamma(S_w^{-d}(l',l)+a)} \\
& \frac{\Gamma(S_w(l,l)+a)}{(S_e(l,l)+b)^{S_w(l,l)+a}} \frac{(S_e^{-d}(l,l)+b)^{S_w^{-d}(l,l)+a}}{\Gamma(S_w^{-d}(l,l)+a)} \frac{\Gamma(D-1+L\gamma)}{\Gamma(D+L\gamma)} \frac{\Gamma(N_l+\gamma)}{\Gamma(N_l^{-d}+\gamma)} & (20)\\
\propto & \prod_{l'=1, l'\neq l}^{L} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{(S_e(l,l')+b)^{S_w(l,l')+a}} \prod_{i=0}^{S_w(d,l')-1} (S_w^{-d}(l,l')+a+i) \\
& \prod_{l'=1, l'\neq l}^{L} \frac{(S_e^{-d}(l',l)+b)^{S_w^{-d}(l',l)+a}}{(S_e(l',l)+b)^{S_w(l',l)+a}} \prod_{i=0}^{S_w(l',d)-1} (S_w^{-d}(l',l)+a+i)
\end{aligned}
$$

$$\frac{(S_e^{-d}(l,l)+b)^{S_w^{-d}(l,l)+a}}{(S_e(l,l)+b)^{S_w(l,l)+a}} \prod_{i=0}^{S_w(d,l)+S_w(l,d)-1} (S_w^{-d}(l,l)+a+i) \frac{N_l^{-d}+\gamma}{D-1+L\gamma} \tag{21}$$

$$\begin{aligned}
\propto \quad & \prod_{l'=1,l'\neq l}^{L} \frac{(S_e^{-d}(l,l')+b)^{S_w^{-d}(l,l')+a}}{(S_e^{-d}(l,l')+b+S_e(d,l'))^{S_w^{-d}(l,l')+a+S_w(d,l')}} \prod_{i=0}^{S_w(d,l')-1} (S_w^{-d}(l,l')+a+i) \\
& \prod_{l'=1,l'\neq l}^{L} \frac{(S_e^{-d}(l',l)+b)^{S_w^{-d}(l',l)+a}}{(S_e^{-d}(l',l)+b+S_e(l',d))^{S_w^{-d}(l',l)+a+S_w(l',d)}} \prod_{i=0}^{S_w(l',d)-1} (S_w^{-d}(l',l)+a+i) \\
& \frac{(S_e^{-d}(l,l)+b)^{S_w^{-d}(l,l)+a}}{(S_e^{-d}(l,l)+b+S_e(l,d)+S_e(d,l))^{S_w^{-d}(l,l)+a+S_w(d,l)+S_w(l,d)}} \\
& (N_l^{-d}+\gamma) \prod_{i=0}^{S_w(d,l)+S_w(l,d)-1} (S_w^{-d}(l,l)+a+i).
\end{aligned} \tag{22}$$

2 Sampling Topic Assignments

The joint probability of topic assignments $\Pr\left(\boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta, \boldsymbol{\pi}, \boldsymbol{y}\right)$ is

$$\Pr\left(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{B} \mid \alpha, \beta, \boldsymbol{\pi}, \boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{23}$$

$$= \int\int \Pr\left(\boldsymbol{z} \mid \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \mid \alpha, \boldsymbol{\pi}, \boldsymbol{y}\right) \Pr\left(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi} \mid \beta\right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi} \cdot \Pr\left(\boldsymbol{B} \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{24}$$

$$\begin{aligned}
= & \int\int \left(\prod_{d=1}^{D}\prod_{k=1}^{K} \theta_{d,k}^{N_{d,k}}\right) \left(\prod_{d=1}^{D} \frac{1}{\Delta(\alpha\boldsymbol{\pi}_{\boldsymbol{y_d}})} \prod_{k=1}^{K} \theta_{d,k}^{\alpha\pi_{y_d,k}-1}\right) \left(\prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{N_{k,v}}\right) \\
& \left(\prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta-1}\right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi} \cdot \prod_{d,d'} \Psi\left(B_{d,d'} \mid \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)
\end{aligned} \tag{25}$$

$$\begin{aligned}
= & \int\int \left(\prod_{d=1}^{D} \frac{1}{\Delta(\alpha\boldsymbol{\pi}_{\boldsymbol{y_d}})} \prod_{k=1}^{K} \theta_{d,k}^{N_{d,k}+\alpha\pi_{y_d,k}-1}\right) \left(\prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{N_{k,v}+\beta-1}\right) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi} \\
& \prod_{d,d'} \Psi\left(B_{d,d'} \mid \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)
\end{aligned} \tag{26}$$

$$= \prod_{d=1}^{D} \frac{\Delta(\boldsymbol{N_d}+\alpha\boldsymbol{\pi}_{\boldsymbol{y_d}})}{\Delta(\alpha\boldsymbol{\pi}_{\boldsymbol{y_d}})} \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{N_k}+\beta)}{\Delta(\beta)} \prod_{d,d'} \Psi\left(B_{d,d'} \mid \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right). \tag{27}$$

The Gibbs sampling equation is then derived as

$$\Pr\left(z_{d,n}=k \mid \boldsymbol{z_{-d,n}}, w_{d,n}=v, \boldsymbol{w_{-d,n}}, \boldsymbol{B}, \alpha, \beta, \boldsymbol{\pi}, \boldsymbol{y_{-d}}, y_d=l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{28}$$

$$= \frac{\Pr\left(z_{d,n}=k, \boldsymbol{z_{-d,n}}, w_{d,n}=v, \boldsymbol{w_{-d,n}}, \boldsymbol{B} \mid \alpha, \beta, \boldsymbol{\pi}, \boldsymbol{y_{-d}}, y_d=l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)}{\Pr\left(\boldsymbol{z_{-d,n}}, \boldsymbol{w_{-d,n}}, \boldsymbol{B_{-d,n}} \mid \alpha, \beta, \boldsymbol{\pi}, \boldsymbol{y_{-d}}, y_d=l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)} \tag{29}$$

$$\begin{aligned}
= & \frac{\Delta(\boldsymbol{N_d}+\alpha\boldsymbol{\pi_l})}{\Delta(\boldsymbol{N_d^{-d,n}}+\alpha\boldsymbol{\pi_l})} \frac{\Delta(\boldsymbol{N_k}+\beta)}{\Delta(\boldsymbol{N_k^{-d,n}}+\beta)} \\
& \prod_{d'} \frac{\Psi\left(B_{d,d'} \mid z_{d,n}=k, \boldsymbol{z_{-d,n}}, \boldsymbol{z_{d'}}, w_{d,n}=v, \boldsymbol{w_{-d,n}}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)}{\Psi\left(B_{d,d'} \mid \boldsymbol{z_{-d,n}}, \boldsymbol{z_{d'}}, \boldsymbol{w_{-d,n}}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)}
\end{aligned} \tag{30}$$

$$\begin{aligned}
\propto & \left(N_{d,k}^{-d,n}+\alpha\pi_{l,k}^{-d,n}\right) \frac{N_{k,v}^{-d,n}+\beta}{N_{k,\cdot}^{-d,n}+V\beta} \\
& \prod_{d'} \Psi\left(B_{d,d'} \mid z_{d,n}=k, \boldsymbol{z_{-d,n}}, \boldsymbol{z_{d'}}, w_{d,n}=v, \boldsymbol{w_{-d,n}}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right),
\end{aligned} \tag{31}$$

where $\pi_{l,k}^{-d,n}$ is estimated based on maximal path assumption [1, 3]

$$\pi_{l,k}^{-d,n} = \frac{\sum_{d':y_{d'}=l} N_{d',k}^{-d,n} + \alpha'}{\sum_{d':y_{d'}=l} N_{d',\cdot}^{-d,n} + K\alpha'}. \tag{32}$$

2.1 Sigmoid Loss

We split d' into two subsets: d^+ and d^-. d^+ denotes the documents that have positive links (observed links, with weight 1) with d. d^- denotes the documents that have negative links (sampled from unobserved links, with weight 0). When using sigmoid loss, the probability of a positive link between documents d and d^+ is

$$\Pr\left(B_{d,d^+} = 1 \,|\, \boldsymbol{z_d}, \boldsymbol{z_{d^+}}, \boldsymbol{w_d}, \boldsymbol{w_{d^+}}, y_d, y_{d^+}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{33}$$

$$= \sigma\left(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_{\boldsymbol{d}} \circ \overline{\boldsymbol{z}}_{\boldsymbol{d^+}}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_{\boldsymbol{d}} \circ \overline{\boldsymbol{w}}_{\boldsymbol{d^+}}) + \rho_{y_d,y_{d^+}}\Omega_{y_d,y_{d^+}}\right) \tag{34}$$

$$= \sigma\left(\sum_{k=1}^{K}\eta_k \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d^+,k}}{N_{d^+,\cdot}} + \sum_{v=1}^{V}\tau_v \frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^+,v}}{N_{d^+,\cdot}} + \rho_{y_d,y_{d^+}}\Omega_{y_d,y_{d^+}}\right), \tag{35}$$

where $\sigma(x) = 1/(1+\exp(-x))$.

Contrarily, the probability of a negative link between documents d and d^- is

$$\Pr\left(B_{d,d^-} = 0 \,|\, \boldsymbol{z_d}, \boldsymbol{z_{d^-}}, \boldsymbol{w_d}, \boldsymbol{w_{d^-}}, y_d, y_{d^-}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{36}$$

$$= 1 - \sigma\left(\sum_{k=1}^{K}\eta_k \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d^-,k}}{N_{d^-,\cdot}} + \sum_{v=1}^{V}\tau_v \frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^-,v}}{N_{d^-,\cdot}} + \rho_{y_d,y_{d^-}}\Omega_{y_d,y_{d^-}}\right). \tag{37}$$

Therefore, the Gibbs sampling equation is

$$\Pr\left(z_{d,n} = k \,|\, \mathrm{rest}\right) \tag{38}$$

$$\propto \left(N_{d,k}^{-d,n} + \alpha\pi_{l,k}^{-d,n}\right)\frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta}$$

$$\prod_{d^+}\sigma\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d^+,k}}{N_{d^+,\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d^+,k'}}{N_{d^+,\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^+,v}}{N_{d^+,\cdot}} + \rho_{y_d,y_{d^+}}\Omega_{y_d,y_{d^+}}\right) \tag{39}$$

$$\prod_{d^-}\left(1 - \sigma\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d^-,k}}{N_{d^-,\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d^-,k'}}{N_{d^-,\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^-,v}}{N_{d^-,\cdot}} + \rho_{y_d,y_{d^-}}\Omega_{y_d,y_{d^-}}\right)\right).$$

2.2 Hinge Loss

When using hinge loss, the probability of a link (either positive or negative, but the weight of a negative link is -1) between documents d and d' is

$$\Pr\left(B_{d,d'} \,|\, \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) = \exp\left(-2c\max(0, \zeta_{d,d'})\right), \tag{40}$$

where c is the regularization parameter (it's set to 1 in our experiments, so it does not appear in the paper); $\zeta_{d,d'}$ is defined as

$$\zeta_{d,d'} = 1 - B_{d,d'}R_{d,d'}, \tag{41}$$

$R_{d,d'}$ is defined in Equation 52.

Equation 40 can be rewritten by introducing a latent variable $\lambda_{d,d'}$ [2] as

$$\Pr\left(B_{d,d'} \,|\, \boldsymbol{z_d}, \boldsymbol{z_{d'}}, \boldsymbol{w_d}, \boldsymbol{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right) \mathrm{d}\lambda_{d,d'}. \tag{42}$$

Thus the Gibbs sampling equation is

$$\Pr\left(z_{d,n} = k \,|\, \text{rest}\right) \propto \left(N_{d,k}^{-d,n} + \alpha\pi_{l,k}\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \prod_{d'} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right). \tag{43}$$

The exponent of final term of the equation above can be expanded as

$$
\begin{aligned}
& -\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}} && (44)\\
\propto\ & -\frac{c^2\zeta_{d,d'}^2 + 2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} && (45)\\
\propto\ & -\frac{c^2(1 - B_{d,d'}R_{d,d'})^2 + 2c\lambda_{d,d'}(1 - B_{d,d'}R_{d,d'})}{2\lambda_{d,d'}} && (46)\\
\propto\ & -\frac{c^2(-2B_{d,d'}R_{d,d'} + R_{d,d'}^2) - 2c\lambda_{d,d'}B_{d,d'}R_{d,d'}}{2\lambda_{d,d'}} && (47)\\
\propto\ & -\frac{c^2R_{d,d'}^2}{2\lambda_{d,d'}} + \frac{cB_{d,d'}(c + \lambda_{d,d'})R_{d,d'}}{\lambda_{d,d'}} && (48)\\
\propto\ & -\frac{c^2\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)^2}{2\lambda_{d,d'}}\\
& + \frac{cB_{d,d'}(c + \lambda_{d,d'})\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)}{\lambda_{d,d'}} && (49)\\
\propto\ & -\frac{c^2\left(\frac{\eta_k^2}{N_{d,\cdot}^2}\frac{N_{d',k}^2}{N_{d',\cdot}^2} + 2\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}}\left(\sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)\right)}{2\lambda_{d,d'}}\\
& + \frac{cB_{d,d'}(c + \lambda_{d,d'})\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}}}{\lambda_{d,d'}} && (50)\\
\propto\ & -\frac{c^2\left(\eta_k^2N_{d',k}^2 + 2\eta_kN_{d',k}\left(\sum_{k'=1}^{K}\eta_{k'}N_{d,k'}^{-d,n}N_{d',k'} + \sum_{v=1}^{V}\tau_vN_{d,v}N_{d',v} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}N_{d,\cdot}N_{d',\cdot}\right)\right)}{2\lambda_{d,d'}N_{d,\cdot}^2N_{d',\cdot}^2}\\
& + \frac{cB_{d,d'}(c + \lambda_{d,d'})\eta_kN_{d',k}}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}. && (51)
\end{aligned}
$$

3 Optimizing Parameters

Let the regression value of documents d and d' be

$$R_{d,d'} = \boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_{\boldsymbol{d}} \circ \overline{\boldsymbol{z}}_{\boldsymbol{d'}}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_{\boldsymbol{d}} \circ \overline{\boldsymbol{w}}_{\boldsymbol{d'}}) + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}. \tag{52}$$

Its partial derivatives are

$$\begin{aligned}
\frac{\partial R_{d,d'}}{\partial \eta_k} &= \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} && (53)\\
\frac{\partial R_{d,d'}}{\partial \tau_v} &= \frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} && (54)\\
\frac{\partial R_{d,d'}}{\partial \rho_{y_d,y_{d'}}} &= \Omega_{y_d,y_{d'}}. && (55)
\end{aligned}$$

3.1 Sigmoid Loss

To optimize regression parameters, we first compute the log likelihood of $\boldsymbol{B}$ as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{B}) &= \log \Pr(\boldsymbol{B} \,|\, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}) + \log \Pr(\boldsymbol{\eta} \,|\, \nu) + \log \Pr(\boldsymbol{\tau} \,|\, \nu) + \log \Pr(\boldsymbol{\rho} \,|\, \nu) && (56)\\
&\propto -\sum_{d,d^+} \log\left(1+\exp\left(-R_{d,d^+}\right)\right) + \sum_{d,d^-}\left(\log\left(\exp\left(-R_{d,d^-}\right)\right) - \log\left(1+\exp\left(-R_{d,d^-}\right)\right)\right)\\
&\quad -\sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2} && (57)\\
&\propto -\sum_{d,d'} \log\left(1+\exp\left(-R_{d,d'}\right)\right) - \sum_{d,d^-} R_{d,d^-} - \sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2}. && (58)
\end{aligned}$$

Its derivatives are

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \eta_k} &\propto -\frac{\eta_k}{\nu^2} + \sum_{d,d'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} - \sum_{d,d^-}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d^-,k}}{N_{d^-,\cdot}} && (59)\\
\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \tau_v} &\propto -\frac{\tau_v}{\nu^2} + \sum_{d,d'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)}\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} - \sum_{d,d^-}\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^-,v}}{N_{d^-,\cdot}} && (60)\\
\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \rho_{l,l'}} &\propto -\frac{\rho_{l,l'}}{\nu^2} + \sum_{d\in l, d'\in l'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)}\Omega_{l,l'} - \sum_{d\in l, d^-\in l'}\Omega_{l,l'}. && (61)
\end{aligned}$$

3.2 Hinge Loss

The log likelihood of $\boldsymbol{B}$ is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{B}) &= \log \Pr(\boldsymbol{B} \,|\, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}) + \log \Pr(\boldsymbol{\eta} \,|\, \nu) + \log \Pr(\boldsymbol{\tau} \,|\, \nu) + \log \Pr(\boldsymbol{\rho} \,|\, \nu) && (62)\\
&\propto -\sum_{d,d'}\frac{(c\zeta_{d,d'}+\lambda_{d,d'})^2}{2\lambda_{d,d'}} - \sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2} && (63)\\
&\propto -\sum_{d,d'}\frac{c^2\zeta_{d,d'}^2+2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} - \sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2} && (64)\\
&\propto -\sum_{d,d'}\frac{c^2(1-B_{d,d'}R_{d,d'})^2+2c\lambda_{d,d'}(1-B_{d,d'}R_{d,d'})}{2\lambda_{d,d'}}
\end{aligned}$$

$$-\sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2}-\sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2}-\sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2} \tag{65}$$

$$\propto \quad -\sum_{d,d'}\frac{c^2R_{d,d'}^2-2c(c+\lambda_{d,d'})B_{d,d'}R_{d,d'}}{2\lambda_{d,d'}}-\sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2}-\sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2}-\sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2}. \tag{66}$$

The partial derivatives of $R_{d,d'}^2$ are

$$\frac{\partial R_{d,d'}^2}{\partial \eta_k} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \eta_k} = 2R_{d,d'}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} \tag{67}$$

$$\frac{\partial R_{d,d'}^2}{\partial \tau_v} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \tau_v} = 2R_{d,d'}\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} \tag{68}$$

$$\frac{\partial R_{d,d'}^2}{\partial \rho_{y_d,y_{d'}}} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \rho_{y_d,y_{d'}}} = 2R_{d,d'}\Omega_{y_d,y_{d'}}. \tag{69}$$

So the partial derivatives of $\mathcal{L}(\boldsymbol{B})$ are

$$\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \eta_k} \propto -\sum_{d,d'}\frac{cN_{d,k}N_{d',k}(cR_{d,d'}-(c+\lambda_{d,d'})B_{d,d'})}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}-\frac{\eta_k}{\nu^2} \tag{70}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \tau_v} \propto -\sum_{d,d'}\frac{cN_{d,v}N_{d',v}(cR_{d,d'}-(c+\lambda_{d,d'})B_{d,d'})}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}-\frac{\tau_v}{\nu^2} \tag{71}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{B})}{\partial \rho_{l,l'}} \propto -\sum_{d\in l,d'\in l'}\frac{c\Omega_{l,l'}(cR_{d,d'}-(c+\lambda_{d,d'})B_{d,d'})}{\lambda_{d,d'}}-\frac{\rho_{l,l'}}{\nu^2}. \tag{72}$$

The likelihood of latent variable $\lambda_{d,d'}$ is

$$\Pr\left(\lambda_{d,d'} \mid \boldsymbol{z},\boldsymbol{w},\boldsymbol{y},\boldsymbol{\Omega},\boldsymbol{B},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right) \propto \frac{1}{\sqrt{2\pi\lambda_{d,d'}}}\exp\left(-\frac{(\lambda_{d,d'}+c\zeta_{d,d'})^2}{2\lambda_{d,d'}}\right) \tag{73}$$

$$\propto \frac{1}{\sqrt{2\pi\lambda_{d,d'}}}\exp\left(-\frac{c^2\zeta_{d,d'}^2}{2\lambda_{d,d'}}-\frac{\lambda_{d,d'}}{2}\right) \tag{74}$$

$$\propto \mathrm{GIG}\left(\lambda_{d,d'}\,;\,\frac{1}{2},1,c^2\zeta_{d,d'}^2\right), \tag{75}$$

where GIG denotes inverse Gaussian distribution which is defined as

$$\mathrm{GIG}(x\,;\,p,a,b)=C(p,a,b)x^{p-1}\exp\left(-\frac{1}{2}\left(\frac{b}{x}+ax\right)\right). \tag{76}$$

We can sample $\lambda_{d,d'}^{-1}$ (then $\lambda_{d,d'}$) from an inverse Gaussian distribution

$$\Pr\left(\lambda_{d,d'}^{-1} \mid \boldsymbol{z},\boldsymbol{w},\boldsymbol{y},\boldsymbol{\Omega},\boldsymbol{B},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right) \propto \mathrm{IG}\left(\lambda_{d,d'}^{-1}\,;\,\frac{1}{c|\zeta_{d,d'}|},1\right), \tag{77}$$

where

$$\mathrm{IG}(x\,;\,a,b)=\sqrt{\frac{b}{2\pi x^3}}\exp\left(-\frac{b(x-a)^2}{2a^2x}\right), \tag{78}$$

for $a>0$ and $b>0$.

4 Sampling Process

The sampling process in the paper is very brief due to space limit, so we give detailed ones here.

Algorithm 1 Sampling Process of LBS-RTM

Initialize every topic assignment $z_{d,n}$ from a uniform distribution
for $m = 1$ to M **do**
    Optimize $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ using L-BFGS (Equations 58, 59, 60, and 61)
    for each document $d = 1$ to D **do**
        Draw block assignment y_d from the multinomial distribution (Equation 13)
        for each token n in document d **do**
            Draw a topic assignment $z_{d,n}$ from the multinomial distribution (Equations 39)
        end for
    end for
end for

Algorithm 2 Sampling Process of LBH-RTM

Set every $\lambda_{d,d'} = 1$ and initialize every topic assignment $z_{d,n}$ from a uniform distribution
for $m = 1$ to M **do**
    Optimize $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ using L-BFGS (Equations 66, 70, 71, and 72)
    for each document $d = 1$ to D **do**
        Draw block assignment y_d from the multinomial distribution (Equation 13)
        for each token n in document d **do**
            Draw a topic assignment $z_{d,n}$ from the multinomial distribution (Equations 43 and 51)
        end for
        for each document d' which document d explicitly links **do**
            Draw $\lambda_{d,d'}^{-1}$ (and then $\lambda_{d,d'}$) from the inverse Gaussian distribution (Equation 77)
        end for
    end for
end for

References

[1] Philip J. Cowans. *Probabilistic Document Modelling*. PhD thesis, University of Cambridge, 2006.

[2] Nicholas G. Polson and Steven L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 2011.

[3] Hanna M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.