Sampling Equation Derivation for Lex-MED-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 Topics

The probability that document d and d' are linked is defined as

$$p(y_{d,d'} \,|\, \boldsymbol{\eta}, \boldsymbol{\tau}, \overline{\boldsymbol{z}}_d, \overline{\boldsymbol{z}}_{d'}, \overline{\boldsymbol{w}}_d, \overline{\boldsymbol{w}}_{d'}) = \exp\left(-2c\max(0, \zeta_{d,d'})\right), \tag{1}$$

where $\overline{\boldsymbol{z}}_d = \frac{1}{N_d}\sum_n z_{d,n}$ and $\overline{\boldsymbol{w}}_d = \frac{1}{N_d}\sum_n w_{d,n}$; $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are weight vectors for two documents' element-wise products of topic proportions and word proportions respectively; c is the regularization parameter; $\zeta_{d,d'}$ is defined as

$$\zeta_{d,d'} = 1 - y_{d,d'}(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'})), \tag{2}$$

where $\circ$ denotes element-wise product of two vectors.

Equation 1 can be expressed [1] as

$$p(y_{d,d'} \,|\, \boldsymbol{\eta}, \boldsymbol{\tau}, \overline{\boldsymbol{z}}_d, \overline{\boldsymbol{z}}_{d'}, \overline{\boldsymbol{w}}_d, \overline{\boldsymbol{w}}_{d'}) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right) \mathrm{d}\lambda_{d,d'}, \tag{3}$$

by introducing a latent variable $\lambda_{d,d'}$.

Therefore the joint probability of Lex-MED-RTM is

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) \propto \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{N_k} + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \prod_{d=1}^{D} \frac{\Delta(\boldsymbol{N_d} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \prod_{d,d'} \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \tag{4}$$

where D and K are numbers of documents and topics respectively; d and d' denote the document pairs that are actually linked; $\Delta(\cdot)$ is defined as

$$\Delta(\boldsymbol{x}) = \frac{\prod_{i=1}^{\dim(\boldsymbol{x})} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\boldsymbol{x})} x_i)}, \tag{5}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

Then the Gibbs sampling equation can be derived as

$$\begin{aligned}
p(z_{d,n} = k \,|\, \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \boldsymbol{y}) &\propto \frac{p(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y})}{p(\boldsymbol{z}_{-d,n}, \boldsymbol{w}_{-d,n}, \boldsymbol{y})} && (6)\\
&\propto \frac{\Delta(\boldsymbol{N_k} + \boldsymbol{\beta})}{\Delta(\boldsymbol{N_k^{-d,n}} + \boldsymbol{\beta})} \frac{\Delta(\boldsymbol{N_d} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{N_d^{-d,n}} + \boldsymbol{\alpha})} \prod_{d'} \frac{\exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right)}{\exp\left(-\frac{(c\zeta_{d,d'}^{-d,n} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right)} && (7)\\
&\propto \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} (N_{d,k}^{-d,n} + \alpha) \prod_{d'} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), && (8)
\end{aligned}$$

where $N_{k,v}$ denotes the count of word v assigned to topic k; $N_{d,k}$ is the number of tokens in document d that are assigned to topic k. Marginal counts are denoted by $\cdot$; $^{-d,n}$ denotes that the count excludes token n in document d; d' denotes the indexes of documents which are actually linked to document d.

The next step is to expand the hinge loss term as

$$
\begin{aligned}
\exp\left(-\frac{(c\zeta_{d,d'}+\lambda_{d,d'})^2}{2\lambda_{d,d'}}\right) &\propto \exp\left(-\frac{c^2\zeta_{d,d'}^2+2\lambda_{d,d'}c\zeta_{d,d'}}{2\lambda_{d,d'}}\right) && (9)\\
&\propto \exp\left(-\frac{c^2(-2y_{d,d'}(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d\circ\overline{\boldsymbol{z}}_{d'})+\boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d\circ\overline{\boldsymbol{w}}_{d'}))+(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d\circ\overline{\boldsymbol{z}}_{d'})+\boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d\circ\overline{\boldsymbol{w}}_{d'}))^2)}{2\lambda_{d,d'}}\right) && (10)\\
&\quad\exp\left(\frac{2\lambda_{d,d'}cy_{d,d'}(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d\circ\overline{\boldsymbol{z}}_{d'})+\boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d\circ\overline{\boldsymbol{w}}_{d'}))}{2\lambda_{d,d'}}\right) && (11)\\
&\propto \exp\left(\frac{cy_{d,d'}(c+\lambda_{d,d'})(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d\circ\overline{\boldsymbol{z}}_{d'}))}{\lambda_{d,d'}}-c^2\frac{(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d\circ\overline{\boldsymbol{z}}_{d'})+\boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d\circ\overline{\boldsymbol{w}}_{d'}))^2)}{2\lambda_{d,d'}}\right) && (12)\\
&= \exp\left(\frac{cy_{d,d'}(c+\lambda_{d,d'})(\sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}}+\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}})}{\lambda_{d,d'}}\right) && (13)\\
&\quad\exp\left(-c^2\frac{(\sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}}+\frac{\eta_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}})^2}{2\lambda_{d,d'}}\right) && (14)\\
&\propto \exp\left(\frac{cy_{d,d'}(c+\lambda_{d,d'})\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}}}{\lambda_{d,d'}}\right) && (15)\\
&\quad\exp\left(-c^2\frac{\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}}(\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}})}{2\lambda_{d,d'}}\right) && (16)\\
&\propto \exp\left(\frac{cy_{d,d'}(c+\lambda_{d,d'})\eta_k N_{d',k}}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}\right) && (17)\\
&\quad\exp\left(-c^2\frac{\eta_k^2N_{d',k}^2+2\eta_kN_{d',k}(\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})}{2\lambda_{d,d'}N_{d,\cdot}^2N_{d',\cdot}^2}\right). && (18)
\end{aligned}
$$

In the sampling process, we only consider linked documents, which means that $y_{d,d'}=1$, so $y_{d,d'}$ can be removed in the sampling equation.

2 Optimizing Topic and Lexical Regression Parameters

Assuming that each element of topic regression parameters $\boldsymbol{\eta}$ and lexical regression parameters $\boldsymbol{\tau}$ is given a Gaussian prior $\mathcal{N}(0,\nu^2)$, the likelihood of $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are computed as

$$p(\boldsymbol{\eta},\boldsymbol{\tau}\,|\,\boldsymbol{z},\boldsymbol{w},\boldsymbol{\lambda})\propto\exp\left(-\sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2}-\sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2}-\sum_{d,d'}\frac{(\lambda_{d,d'}+c\zeta_{d,d'})^2}{2\lambda_{d,d'}}\right). \tag{19}$$

Therefore, the log likelihood $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})$ is

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau}) \propto -\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{d,d'} \frac{(\lambda_{d,d'} + c\zeta_{d,d'})^2}{2\lambda_{d,d'}}. \tag{20}$$

It can be further expanded as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau}) &\propto -\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{d,d'} \frac{c^2\zeta_{d,d'}^2 + 2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} && (21)\\
&= -\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - && (22)\\
&\quad \sum_{d,d'} \frac{c^2(1 - (\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'})))^2 + 2c\lambda_{d,d'}(1 - (\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'})))}{2\lambda_{d,d'}} && (23)\\
&\propto -\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} + && (24)\\
&\quad \sum_{d,d'} \frac{2c(c + \lambda_{d,d'})(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'})) - c^2(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'}))^2}{2\lambda_{d,d'}}. && (25)
\end{aligned}$$

Let

$$W_{d,d'} = \boldsymbol{\eta}^{\mathrm{T}}(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \boldsymbol{\tau}^{\mathrm{T}}(\overline{\boldsymbol{w}}_d \circ \overline{\boldsymbol{w}}_{d'}), \tag{26}$$

then $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})$ is

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau}) \propto -\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} + \sum_{d,d'} \frac{2c(c + \lambda_{d,d'})W_{d,d'} - c^2 W_{d,d'}^2}{2\lambda_{d,d'}}. \tag{27}$$

Next step is to compute the derivatives. We first compute $W_{d,d'}$'s derivatives as

$$\begin{aligned}
\frac{\partial W_{d,d'}}{\partial \eta_k} &= \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} && (28)\\
\frac{\partial W_{d,d'}}{\partial \tau_v} &= \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} && (29)\\
\frac{\partial W_{d,d'}^2}{\partial \eta_k} &= 2W_{d,d'} \frac{\partial W_{d,d'}}{\partial \eta_k} = 2W_{d,d'} \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} && (30)\\
\frac{\partial W_{d,d'}^2}{\partial \tau_v} &= 2W_{d,d'} \frac{\partial W_{d,d'}}{\partial \tau_v} = 2W_{d,d'} \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}}. && (31)
\end{aligned}$$

Therefore, the derivatives are

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})}{\partial \eta_k} &\propto -\frac{\eta_k}{\nu^2} + \sum_{d,d'} \frac{cN_{d,k}N_{d',k}(c + \lambda_{d,d'} - cW_{d,d'})}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}} && (32)\\
\frac{\partial \mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau})}{\partial \tau_v} &\propto -\frac{\tau_v}{\nu^2} + \sum_{d,d'} \frac{cN_{d,v}N_{d',v}(c + \lambda_{d,d'} - cW_{d,d'})}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}. && (33)
\end{aligned}$$

3 Sampling Latent Variables

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

$$\begin{aligned} p(\lambda_{d,d'} \,|\, \boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\tau}) &\propto \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{(\lambda_{d,d'} + c\zeta_{d,d'})^2}{2\lambda_{d,d'}}\right) &(34)\\ &\propto \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{c^2\zeta_{d,d'}^2}{2\lambda_{d,d'}} - \frac{\lambda_{d,d'}}{2}\right) &(35)\\ &= \mathrm{GIG}\left(\lambda_{d,d'}; \frac{1}{2}, 1, c^2\zeta_{d,d'}^2\right), &(36)\end{aligned}$$

where GIG is generalized inverse Gaussian distribution which is defined as

$$\mathrm{GIG}(x; p, a, b) = C(p, a, b) x^{p-1} \exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right). \quad (37)$$

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

$$p(\lambda_{d,d'}^{-1} | \boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\tau}) = \mathrm{IG}\left(\lambda_{d,d'}^{-1}; \frac{1}{c|\zeta_{d,d'}|}, 1\right), \quad (38)$$

where

$$\mathrm{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\left(-\frac{b(x-a)^2}{2a^2 x}\right), \quad (39)$$

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

4 Sampling Process

The general sampling process of Lex-MED-RTM is given in Algorithm 1, which is similar to MED-LDA [2].

Algorithm 1 Sampling Process

set $\boldsymbol{\lambda} = 1$ and draw $z_{d,n}$ from a uniform distribution
for $m = 1$ to M **do**
  optimize $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ using L-BFGS (Eqaution 27, 32 and 33)
  for $d = 1$ to D **do**
    for each word n in document d **do**
      draw a topic $z_{d,n}$ from the multinomial distribution (Equation 8, 17 and 18)
    end for
    for each document d' which document d links **do**
      draw $\lambda_{d,d'}^{-1}$ (and then $\lambda_{d,d'}$) from the inverse Gaussian distribution (Equation 38)
    end for
  end for
end for

The sampling process starts from initialization of $\boldsymbol{\lambda}$ and topic assignments. In each iteration, $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are optimized by feeding their likelihood and derivatives to L-BFGS (MALLET provides a nice implementation).[1] When sampling for documents, we first sample each word's topic assignment. Then for each $\lambda_{d,d'}$, we sample its reciprocal from the inverse Gaussian distribution.

[1] MALLET: `http://mallet.cs.umass.edu/`

References

[1] Nicholas G. Polson and Steven L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

[2] Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *The Journal of Machine Learning Research*, 15(1):1073–1110, 2014.