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

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

where $\overline{z}_d = \frac{1}{N_d} \sum_n z_{d,n}$ and $\overline{w}_d = \frac{1}{N_d} \sum_n w_{d,n}$; $\eta$ and $\tau$ 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

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

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

Equation 1 can be expressed [1] as

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

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

Therefore the joint probability of Lex-MED-RTM is

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

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

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

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

Then the Gibbs sampling equation can be derived as

$$p(z_{d,n} = k | z_{-d,n}, w_{-d,n}, \mathbf{y}) \propto \frac{p(z, w, y)}{p(z_{-d,n}, w_{-d,n}, y)} \propto \frac{\Delta(N_k + \beta)}{\Delta(N_k - d,n + \beta)} \frac{\Delta(N_d + \alpha)}{\Delta(N_d - d,n + \alpha)} \prod_{d'} \frac{1}{\sqrt{2\pi \lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \tag{6}$$

$$\propto \frac{N_{k,v}}{N_{k,v} - \zeta_{d,n} + \alpha} (N_{d,k} - d,n + \alpha) \prod_{d'} \frac{1}{\sqrt{2\pi \lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \tag{7}$$

$$\propto \frac{N_{k,v}}{N_{k,v} - \zeta_{d,n} + \alpha} (N_{d,k} - d,n + \alpha) \prod_{d'} \frac{1}{\sqrt{2\pi \lambda_{d,d'}}} \exp\left(-\frac{(c\zeta_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}}\right), \tag{8}$$
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 \(-\cdot\cdot\cdot\) \(-\cdot\cdot\cdot\). 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

\[
\exp \left( -\frac{(c_{d,d'} + \lambda_{d,d'})^2}{2\lambda_{d,d'}} \right) \propto \exp \left( -\frac{c^2_{d,d'} + 2\lambda_{d,d'} c_{d,d'}}{2\lambda_{d,d'}} \right) \quad (9)
\]

\[
\exp \left( -\frac{c^2(-2\gamma_{d,d'}(\eta^T z_{d} + \tau^T w_{d} + \tau^T w_{d'}))^2}{2\lambda_{d,d'}} \right) \quad (10)
\]

\[
\exp \left( -\frac{c^2 y_{d,d'}(\eta^T z_{d} + \tau^T w_{d} + \tau^T w_{d'}))^2}{2\lambda_{d,d'}} \right) \quad (11)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\eta^T z_{d} + \tau^T w_{d} + \tau^T w_{d'}))^2 \right) \quad (12)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (13)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (14)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (15)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (16)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (17)
\]

\[
\exp \left( c y_{d,d'}(c + \lambda_{d,d'})(\sum_{k'=1}^{K} \eta_{k'k} N_{d,k'} N_{d',k'} + \sum_{v=1}^{V} \tau_{v} N_{d,v} N_{d',v}))^2 \right) \quad (18)
\]

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

\[
p(\eta, \tau | z, w, \lambda) \propto \exp \left( -\frac{\sum_{k=1}^{K} \eta_{k}^2}{2\nu^2} - \frac{V}{2\nu^2} - \sum_{v=1}^{V} \tau_{v}^2 - \sum_{d,d'} \frac{(\lambda_{d,d'} + c_{d,d'})^2}{2\lambda_{d,d'}} \right) \quad (19)
\]
Therefore, the log likelihood $\mathcal{L}(\eta, \tau)$ is
\[
\mathcal{L}(\eta, \tau) \propto -\sum_{k=1}^{K} \frac{\eta_k^2}{\nu k^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{\nu v^2} - \sum_{d,d'} \frac{(\lambda_{d,d'} + c\zeta_{d,d'})^2}{2\lambda_{d,d'}}. \tag{20}
\]
It can be further expanded as
\[
\mathcal{L}(\eta, \tau) \propto -\sum_{k=1}^{K} \frac{\eta_k^2}{\nu k^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{\nu v^2} - \sum_{d,d'} \frac{c^2\zeta_{d,d'} + 2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} \tag{21}
\]
\[
= -\sum_{k=1}^{K} \frac{\eta_k^2}{\nu k^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{\nu v^2} - \sum_{d,d'} \frac{c^2(1 - (\eta^T(z_d \circ z_{d'})) + \rho_d^T(\mathbf{w}_d \circ \mathbf{w}_{d'})))^2 + 2c\lambda_{d,d'} (1 - (\eta^T(z_d \circ z_{d'})) + \rho_d^T(\mathbf{w}_d \circ \mathbf{w}_{d'})))}{2\lambda_{d,d'}} \tag{22}
\]
\[
\approx -\sum_{k=1}^{K} \frac{\eta_k^2}{\nu k^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{\nu v^2} + \sum_{d,d'} \frac{2c(c + \lambda_{d,d'})((\eta^T(z_d \circ z_{d'})) + \rho_d^T(\mathbf{w}_d \circ \mathbf{w}_{d'})) - c^2(\eta^T(z_d \circ z_{d'})) + \rho_d^T(\mathbf{w}_d \circ \mathbf{w}_{d'}))^2}{2\lambda_{d,d'}}. \tag{24}
\]
Let
\[
W_{d,d'} = \eta^T(z_d \circ z_{d'}) + \rho_d^T(\mathbf{w}_d \circ \mathbf{w}_{d'}), \tag{26}
\]
then $\mathcal{L}(\eta, \tau)$ is
\[
\mathcal{L}(\eta, \tau) \propto -\sum_{k=1}^{K} \frac{\eta_k^2}{\nu k^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{\nu v^2} + \sum_{d,d'} \frac{2c(c + \lambda_{d,d'})W_{d,d'} - c^2W_{d,d'}^2}{2\lambda_{d,d'}}. \tag{27}
\]
Next step is to compute the derivatives. We first compute $W_{d,d'}$’s derivatives as
\[
\frac{\partial W_{d,d'}}{\partial \eta_k} = \frac{N_{d,k} N_{d',k}}{N_d \cdot N_{d'}}, \tag{28}
\]
\[
\frac{\partial W_{d,d'}}{\partial \tau_v} = \frac{N_{d,v} N_{d',v}}{N_d \cdot N_{d'}}, \tag{29}
\]
\[
\frac{\partial W_{d,d'}^2}{\partial \eta_k} = 2W_{d,d'} \frac{\partial W_{d,d'}}{\partial \eta_k} = 2W_{d,d'} \frac{N_{d,k} N_{d',k}}{N_d \cdot N_{d'}}, \tag{30}
\]
\[
\frac{\partial W_{d,d'}^2}{\partial \tau_v} = 2W_{d,d'} \frac{\partial W_{d,d'}}{\partial \tau_v} = 2W_{d,d'} \frac{N_{d,v} N_{d',v}}{N_d \cdot N_{d'}},. \tag{31}
\]
Therefore, the derivatives are
\[
\frac{\partial \mathcal{L}(\eta, \tau)}{\partial \eta_k} \propto -\frac{\eta_k}{\nu k^2} + \sum_{d,d'} \frac{cN_{d,k} N_{d',k}(c + \lambda_{d,d'}) - cW_{d,d'}}{\lambda_{d,d'} N_d \cdot N_{d'}} \tag{32}
\]
\[
\frac{\partial \mathcal{L}(\eta, \tau)}{\partial \tau_v} \propto -\frac{\tau_v}{\nu v^2} + \sum_{d,d'} \frac{cN_{d,v} N_{d',v}(c + \lambda_{d,d'}) - cW_{d,d'}}{\lambda_{d,d'} N_d \cdot N_{d'}}. \tag{33}
\]
3 Sampling Latent Variables

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

$$p(\lambda_{d,d'} | z, \eta, \tau) \propto \frac{1}{\sqrt{2\pi \lambda_{d,d'}}} \exp\left(-\frac{(\lambda_{d,d'} + c\zeta_{d,d'})^2}{2\lambda_{d,d'}}\right)$$

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

$$= \text{GIG}\left(\lambda_{d,d'}; \frac{1}{2}, 1, c^2\zeta_{d,d'}^2\right),$$

where GIG is generalized inverse Gaussian distribution which is defined as

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

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

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

where

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

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

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

The sampling process starts from initialization of $\lambda$ and topic assignments. In each iteration, $\eta$ and $\tau$ are optimized by feeding their likelihood and derivatives to L-BFGS (MALLET provides a nice implementation). 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.

\[\text{MALLET: http://mallet.cs.umass.edu/}\]
References
