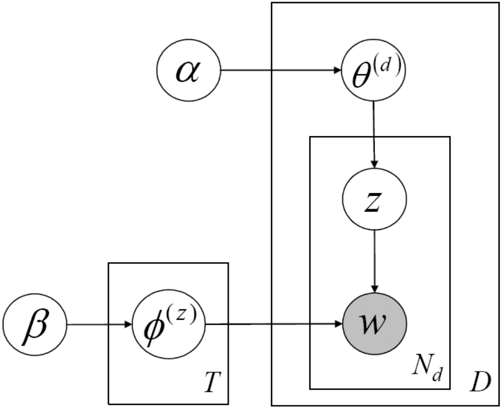
Gibbs Sampling for LDA

Lei Tang

Department of CSE
Arizona State University

January 7, 2008



α, β are fixed hyper-parameters. We need to estimate parameters θ for each document and ϕ for each topic. Z are latent variables. This is different from original LDA work.

Property of Dirichlet

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

The expectation of Dirichlet is

$$E(\mu_k) = \frac{\alpha_k}{\alpha_0}$$

where $\alpha_0 = \sum \alpha_k$.

Gibbs Variants

1. Gibbs Sampling
   - Draw a conditioned on b, c
   - Draw b conditioned on a, c
   - Draw c conditioned on a, b
2. Block Gibbs Sampling
   - Draw a, b conditioned on c
   - Draw c conditioned on a,b
3. Collapsed Gibbs Sampling
   - Draw a conditioned on c
   - Draw c conditioned on a

b is collopsed out during the sampling process.

Collapsed Sampling for LDA

In the original paper "Finding Scientific Topics", the authors are more interested in text modelling, (find out Z), hence, the Gibbs sampling procedure boils down to estimate

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

Here, θ, ϕ are intergrated out. Actually, if we know the exact Z for each document, it's trivial to estimate θ and ϕ.

$$\begin{aligned} P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) &\propto P(z_i = j, \mathbf{z}_{-i}, \mathbf{w}) \\ &= P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\ &= P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) P(z_i = j | \mathbf{z}_{-i}) \end{aligned}$$

The first term is the likelihood and the 2nd term like a prior.

$$
\begin{aligned}
& P(w_i|z_i=j,\mathbf{z}_{-i},\mathbf{w}_{-i}) \\
= & \int P(w_i|z_i=j,\phi^{(j)})P(\phi^{(j)}|\mathbf{z}_{-i},\mathbf{w}_{-i})d\phi^{(j)} \\
= & \int \phi^{(j)}_{w_i} P(\phi^{(j)}|\mathbf{z}_{-i},\mathbf{w}_{-i})d\phi^{(j)}
\end{aligned}
$$

$$
\begin{aligned}
P(\phi^{(j)}|\mathbf{z}_{-i},\mathbf{w}_{-i}) & \propto P(\mathbf{w}_{-i}|\phi^{(j)},z_{-i})P(\phi^{j}) \\
& \sim Dirichlet(\beta + n^{(w)}_{-i,j})
\end{aligned}
$$

Here, $n^{(w)}_{-i,j}$ is the number of instances of word w assigned to topic j. Using the property of expectation of Dirichlet distribution, we have

$$
P(w_i|z_i=j,\mathbf{z}_{-i},\mathbf{w}_{-i}) = \frac{n^{(w_i)}_{-i,j}+\beta}{n^{(\cdot)}_{-i,j}+W\beta}
$$

where $n_{-i,j}$ total number of words assigned to topic j.

$$
\begin{aligned}
& P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\
= & \int P(w_i | z_i = j, \phi^{(j)}) P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)} \\
= & \int \phi^{(j)}_{w_i} P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)}
\end{aligned}
$$

$$
\begin{aligned}
P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) & \propto P(\mathbf{w}_{-i} | \phi^{(j)}, z_{-i}) P(\phi^{j}) \\
& \sim Dirichlet(\beta + n^{(w)}_{-i,j})
\end{aligned}
$$

Here, $n^{(w)}_{-i,j}$ is the number of instances of word w assigned to topic j.
Using the property of expectation of Dirichlet distribution, we have

$$
P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta}
$$

where $n_{-i,j}$ total number of words assigned to topic j.

Similarly, for the 2nd term, we have

$$\begin{aligned} P(z_i = j|\mathbf{z}_{-i}) &= \int P(z_i = j|\theta^{(d)})P(\theta^{(d)}|\mathbf{z}_{-i})d\theta^{(d)} \\ P(\theta^{(d)}|\mathbf{z}_{-i}) &\propto P(\mathbf{z}_{-i}|\theta^{(d)})P(\theta^{(d)}) \\ &\sim Dirichlet(n^{(d)}_{-i,j} + \alpha) \end{aligned}$$

where $n^{(d)}_{-i,j}$ is the number of words assigned to topic j excluding current one.

$$P(z_i = j|z_{-i}) = \frac{n^{(d)}_{-i,j} + \alpha}{n^{(d)}_{-i,\cdot} + K\alpha}$$

where $n^{(d)}_{-i,\cdot}$ is the total number of topics assigned to document d excluding current one.

Algorithm

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \frac{n^{(d)}_{-i,j} + \alpha}{n^{(d)}_{-i,\cdot} + K\alpha}$$

Need to record four count variables:

- document-topic count $n^{(d)}_{-i,j}$
- document-topic sum $n^{(d)}_{-i,\cdot}$ (actually a constant)
- topic-term count $n^{(w_i)}_{-i,j}$
- topic-term sum $n^{(\cdot)}_{-i,j}$

Parameter Estimation

To obtain ϕ, and θ, two ways, (draw one sample of z or draw multiple samples of z to calculate the average)

$$\phi_{j,w} = \frac{n_w^{(j)} + \beta}{\sum_{w=1}^{V} n_w^{(j)} + V\beta}$$

$$\theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{\sum_{z=1}^{K} n_z^{(d)} + K\alpha}$$

where $n_w^{(j)}$ is the freqency of word assigned to topic j, and $n_z^{(d)}$ is the number of words assigned to topic z.

Comment

- Compared with VB, Gibbs Sampling is easy to implement.
- Easy to extend.
- More efficient. Faster to obtain good approximation.