Homework 4
Inference and Representation

Aodong Li
NetID: al5350
al5350@nyu.edu

June 27, 2018

1. (a) We have

$$
\begin{aligned}
p(\theta|\alpha) &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \\
p(X = x|\theta) &= \prod_{n=1}^{N} \sum_{k=1}^{K} \theta_k 1[x_i = k] \\
&= \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} 1[x_i = k]}, \\
p(\theta|x, \alpha) &\propto p(\theta, x, \alpha) \\
&= p(x|\theta)p(\theta|\alpha) \\
&\propto \left(\prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} 1[x_i = k]} \right) \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \\
&\propto \prod_{k=1}^{K} \theta_k^{\alpha_k + \sum_{i=1}^{N} 1[x_i = k] - 1},
\end{aligned}
$$

where it can be seen that $p(\theta|x, \alpha)$ is a Dirichlet distribution with parameters $\alpha_k' = \alpha_k + \sum_{i=1}^{N} 1[x_i = k]$. So the Dirichlet distribution is the conjugate prior for the Categorical distribution.

(b) From (a), we have $p(\theta|x) = Dir(\alpha')$ and

$$\begin{aligned}
p(x_{new}|x,\alpha) &= \int_\theta p(x_{new}|\theta)p(\theta|x)\mathrm{d}\ \theta \\
&= \frac{1}{B(\alpha')}\int_\theta \prod_{k=1}^{K} \theta_k^{\alpha_k+\sum_{i=1}^{N}1[x_i=k]+1[x_{new}=k]-1}\mathrm{d}\ \theta \\
&= \frac{B(\alpha'')}{B(\alpha')},
\end{aligned}$$

where $\alpha''_k = \alpha'_k + 1[x_{new} = k]$.

2. (a) See Jupyter notebook for the implementation of topic models of collaborative filtering.

   (b) We can first write the joint distribution as

$$p(w,z,\theta;\alpha,\beta) = \prod_{d=1}^{M} p(\theta_d;\alpha)\prod_{n=1}^{N} p(w_{di}|z_{di};\beta)p(z_{di}|\theta_d).$$

In the following, we will leave out of the hyperparameters to relieve the cluster of notations.

The full conditional of z_{di} can be represented as

$$\begin{aligned}
p(z_{di}|w_{di},\theta_d;\beta) &\propto p(z_{di},w_{di},\theta_d) \\
&= p(\theta_d)p(z_{di}|\theta_d)p(w_{di}|z_{di};\beta) \\
&\propto \left(\prod_{t=1}^{T}\theta_{dt}^{\alpha_t-1}\right)\left(\prod_{t=1}^{T}\theta_{dt}^{1[z_{di}=t]}\right)\left(\prod_{w=1}^{W}\beta_{z_{di}w}^{1[w_{di}=w]}\right) \\
&\propto \left(\prod_{t=1}^{T}\theta_{dt}^{1[z_{di}=t]}\right)\left(\prod_{w=1}^{W}\beta_{z_{di}w}^{1[w_{di}=w]}\right).
\end{aligned}$$

The full conditional of θ_d is

$$\begin{aligned}
p(\theta_d|z_d;\alpha) &\propto p(\theta_d,z_d;\alpha) \\
&= p(\theta_d;\alpha)p(z_d|\theta_d) \\
&= \left(\prod_{t=1}^{T}\theta_{dt}^{\alpha_t-1}\right)\left(\prod_{t=1}^{T}\theta_{dt}^{\sum_{i=1}^{N}1[z_{di}=t]}\right) \\
&= \prod_{t=1}^{T}\theta_{dt}^{\alpha_t+n_{dt}-1},
\end{aligned}$$

where n_{dt} is the number of words belonging to topic t for each document d.

(c)

$$
\begin{aligned}
p(z_d|w_d;\alpha,\beta) &\propto p(z_d, w_d;\alpha,\beta) \\
&= \int_{\theta_d} p(z_d, w_d, \theta_d;\alpha,\beta)\mathrm{d}\ \theta_d \\
&= \int_{\theta_d} p(\theta_d;\alpha)\prod_{i=1}^{N} p(z_{di}|\theta_d)p(w_{di}|z_{di};\beta)\mathrm{d}\ \theta_d \\
&= \frac{1}{B(\alpha)}\int_{\theta_d}\left(\prod_{t=1}^{T}\theta_{dt}^{\alpha_t-1}\right)\left(\prod_{t=1}^{T}\theta_{dt}^{n_{dt}}\right)\left(\prod_{t=1}^{T}\prod_{w=1}^{W}\beta_{tw}^{n_{dw}}\right)\mathrm{d}\ \theta_d \\
&= \frac{B(\alpha')}{B(\alpha)}\left(\prod_{t=1}^{T}\prod_{w=1}^{W}\beta_{tw}^{n_{dw}}\right),
\end{aligned}
$$

where $\alpha'_t = \alpha_t + n_{dt}$ and n_{dt} is the number of words belonging to topic t for each document d.

So to derive a Gibbs sampler, we have

$$
\begin{aligned}
p(z_{di}|z_{d-i}, w_d) &= \frac{p(z,w)}{p(z_{-i},w)} \\
&= \frac{p(z)}{p(z_{-i})}\frac{p(w|z)}{p(w_{-i}|z_{-i})p(w_i)} \\
&\propto \frac{p(z)}{p(z_{-i})}\frac{p(w|z)}{p(w_{-i}|z_{-i})} \\
&= \frac{B(\alpha.)}{B(\alpha_{-i})}\beta_{z_i,w} \\
&= (n_{dk}^{-i} + \alpha_k)\beta_{z_i,w}
\end{aligned}
$$

(d) See jupyter notebook.