Supplement to Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process

Chong Wang
Computer Science Department
Princeton University
chongw@cs.princeton.edu

David M. Blei
Computer Science Department
Princeton University
blei@cs.princeton.edu

Abstract

In this supplement, we provide the details of computing the conditional density function for a term in deriving the Gibbs sampler for the sparse topic models.

1 Model notations

For the convenience of the reader, we list the primary variables that are used in deriving the Gibbs sampler for sparseTM. Note that marginal counts are represented with dots.

b_{kv} The term selector of term v for topic k – if and only if $b_{kv} = 1$, topic k will have the chance to include term v in it.

$\boldsymbol{\beta}_k$ The topic distribution for topic k. This is integrated out in the Gibbs sampler.

π_k The expected proportion of "1"s in b_k.

w_{di} The ith word in document d. This is an observed variable.

z_{di} The topic assignment for word w_{di}.

n_{dk} The number of customers in restaurant d eating dish k is denoted n_{dk}. (This is the number of words in document d that are assigned to topic k.) Further, $n_{d\cdot}$ denotes the number of customers in restaurant d. (This is the total number of words in document d.)

$n_k^{(v)}$ The number of times that term v has been assigned to topic k. In addition, $n_k^{(\cdot)}$ denotes the number of times that all the terms have been assigned to topic k.

u The index of a new topic in the sampling process.

2 Computing the conditional density

Recall that in the main text, the key for developing a collapsed Gibbs sampler is to compute the conditional density of w_{di} under the topic component k given all data items except w_{di} as after integrating out $\boldsymbol{\beta}_k$ and $\boldsymbol{b}_k$,

$$\begin{aligned} f_k^{-w_{di}}(w_{di} = v|\pi_k) &\triangleq p(w_{di} = v|\{w_{d'i'}, z_{d'i'} : z_{d'i'} = k, d'i' \neq di\}, \pi_k) \qquad (1) \\ &= \int d\boldsymbol{\beta}_k \; p(w_{di} = v|\boldsymbol{\beta}_k) p(\boldsymbol{\beta}_k|\{w_{d'i'}, z_{d'i'} : z_{d'i'} = k, d'i' \neq di\}, \pi_k). \\ &= \int d\boldsymbol{\beta}_k \; p(w_{di} = v|\boldsymbol{\beta}_k) \sum_{\boldsymbol{b}_k} p(\boldsymbol{\beta}_k, \boldsymbol{b}_k|\{w_{d'i'}, z_{d'i'} : z_{d'i'} = k, d'i' \neq di\}, \pi_k). \end{aligned}$$

With a little abuse of notation, we simplify $f_k^{-w_{di}}(w_{di} = v|\pi_k)$ to $f_k^{-w}(w = v|\pi_k)$.

2.1 $k = u$, i.e. k is a new topic

Since no words are assigned to this new topic, $f_u^{-w}(w = v|\pi_k)$ is just the marginal probability after integrating out $\boldsymbol{\beta}_k$ and $\boldsymbol{b}_k$. Thus, we have

$$f_u^{-w}(w = v|\pi_u) = \int d\boldsymbol{\beta}_u \, p(w_{di} = v|\boldsymbol{\beta}_u) \sum_{\boldsymbol{b}_u} p(\boldsymbol{\beta}_u|\boldsymbol{b}_u)p(\boldsymbol{b}_u|\pi_u). \tag{2}$$

The conditional probability for $\boldsymbol{b}_u$ given π_u is $p(\boldsymbol{b}_u|\pi_u) = \prod_{v=1}^{V} \pi^{b_{uv}}\pi^{1-b_{uv}} = \pi_u^{a_u}(1-\pi_u)^{V-a_u}$, where $a_u = \sum_{v=1}^{V} b_{uv}$. For the pseudo term $V+1$, the only case that contributes to the result is $a_u = 0$, i.e. $\boldsymbol{b}_u = [0, \cdots, 0, 1]$, thus

$$f_u^{-w}(w = V+1|\pi_u) = (1-\pi_u)^V, \tag{3}$$

By symmetry, for any of the "in vocabulary" terms $v \in \mathcal{V}$,

$$f_u^{-w}(w = v|\pi_u) = \left(1 - (1-\pi_u)^V\right)/V. \tag{4}$$

2.2 $k \neq u$, i.e. k is not a new topic

The derivation above only works for the new topic. Now we consider the general case where a topic already has some word assignments. We assume that we already *exclude* word w in the following descriptions. The following derivation is similar to [2].

To illustrate the algorithm, we need to introduce some more notations. Let $B_k \triangleq \{v : n_k^{(v)} > 0, v \in \mathcal{V}\}$, indicating the set of terms having non-zero assignment to topic k. We can safely assume $B_k \neq \Phi$. (Here, Φ represents an empty set.) Since if $B_k = \Phi$, we return the "new topic" case we just discussed. In addition, let A_k be an arbitrary subset of $\mathcal{V} = \{1, 2, \cdots, V\}$. and $\mathcal{E}_k$ be the set that contains all possible A_ks. The cardinality of $\mathcal{E}_k$ is $|\mathcal{E}_k| = 2^V$. In the sequel, we implicitly assume $A_k \in \mathcal{E}_k$. Under such assumption, we define $p(A_k|\pi_k) \triangleq p(b_{kv} = 1, v \in A_k, b_{kv'} = 0, v' \notin A_k|\pi_k) = \pi_k^{|A_k|}(1-\pi_k)^{V-|A_k|}$, i.e. the probability of those selectors in A_k being "on". Let $S_k \triangleq \{w_{di}, z_{di} : z_{di} = k\}$ be the set of word assignments for topic k, where we already exclude word w. Note that we can write $f_k^{-w}(w = v|\pi_k)$ as (according to Equation 1)

$$f_k^{-w}(w = v|\pi_k) = \int d\boldsymbol{\beta}_k \, p(w_{di} = v|\boldsymbol{\beta}_k)p(\boldsymbol{\beta}_k|, S_k, \pi_k) = \mathbb{E}[\beta_{kv}|S_k, \pi_k]. \tag{5}$$

Now, we compute the posterior of $\boldsymbol{\beta}_k$ given S_k and π_k,

$$\begin{aligned}
&p(\boldsymbol{\beta}_k|S_k, \pi_k) \propto p(S_k|\boldsymbol{\beta}_k)p(\boldsymbol{\beta}_k|\pi_k)\\
&= p(S_k|\boldsymbol{\beta}_k) \sum_{\boldsymbol{b}_k} p(\boldsymbol{\beta}_k|\boldsymbol{b}_k)p(\boldsymbol{b}_k|\pi_k)\\
&\propto \prod_{v \in B_k} \beta_{kv}^{n_k^{(v)}} \sum_{A_k : A_k \supset B_k} p(A_k|\pi_k)\frac{\Gamma(|A_k|\gamma)}{\Gamma^{|A_k|}(\gamma)} \prod_{v \in A_k} \beta_{kv}^{\gamma-1} \prod_{v \notin A_k} 1_{\beta_{kv}=0}\\
&= \sum_{A_k : A_k \supset B_k} p(A_k|\pi_k)\frac{\Gamma(|A_k|\gamma)}{\Gamma^{|A_k|}(\gamma)} \prod_{v \in A_k} \beta_{kv}^{n_k^{(v)}+\gamma-1} \prod_{v \notin A_k} 1_{\beta_{kv}=0},\\
&= \sum_{A_k : A_k \supset B_k} p(A_k|\pi_k)\frac{\Gamma(|A_k|\gamma)}{\Gamma^{|A_k|}(\gamma)} \frac{\prod_{v \in A_k}\Gamma(n_k^{(v)}+\gamma)}{\Gamma(n_k^{(\cdot)}+|A_k|\gamma)} \left(\frac{\Gamma(n_k^{(\cdot)}+|A_k|\gamma)}{\prod_{v \in A_k}\Gamma(n_k^{(v)}+\gamma)} \prod_{v \in A_k} \beta_{kv}^{n_k^{(v)}+\gamma-1} \prod_{v \notin A_k} 1_{\beta_{kv}=0} \right),
\end{aligned}$$

where $1_{\beta_{kv}=0}$ is an indicator function. We see that $p(\boldsymbol{\beta}_k|S_k, \pi_k)$ is a mixture of Dirichlet distributions with $2^{V-|B_k|}$ components. For a particular setting of A_k ($A_k \supset B_k$), let

$$\begin{aligned}
q(A_k|\pi_k) &\triangleq p(A_k|\pi_k)\frac{\Gamma(|A_k|\gamma)}{\Gamma^{|A_k|}(\gamma)} \frac{\prod_{v \in A_k}\Gamma(n_k^{(v)}+\gamma)}{\Gamma(n_k^{(\cdot)}+|A_k|\gamma)}\\
&\propto p(A_k|\pi_k)\frac{\Gamma(|A_k|\gamma)}{\Gamma(n_k^{(\cdot)}+|A_k|\gamma)}.
\end{aligned} \tag{6}$$

We used fact that

$$\frac{\prod_{v\in A_k}\Gamma(n_k^{(v)}+\gamma)}{\Gamma^{|A_k|}(\gamma)} = \frac{\prod_{v\in B_k}\Gamma(n_k^{(v)}+\gamma)\Gamma^{|A_k|-|B_k|}(\gamma)}{\Gamma^{|A_k|}(\gamma)} = \frac{\prod_{v\in B_k}\Gamma(n_k^{(v)}+\gamma)}{\Gamma^{|B_k|}(\gamma)}, \tag{7}$$

which is *not* a function of A_k, and cancelled out in the normalization. We obtain

$$p(\boldsymbol{\beta}_k|S_k,\pi_k) \propto \sum_{A_k\in\mathcal{E}_k, A_k\supset B_k} q(A_k|\pi_k)\,\text{Dirichlet}\left(\mathbf{1}_{A_k}(\boldsymbol{n}_k+\gamma)\right),$$

where Dirichlet $(\mathbf{1}_{A_k}(\boldsymbol{n}_k+\gamma))$ is a Dirichlet distribution over the simplex defined by A_k (the vth entry is 0 if $v\notin A_k$) and $q(A_k|\pi_k)$ is the unnormalized weight. To compute f for $w=v$, we have

$$\begin{aligned}
f_k^{-w}(w=v|\pi_k) &= \mathbb{E}[\beta_{kv}|S_k,\pi_k] \\
&\propto \sum_{A_k\in\mathcal{E}_k, A_k\supset B_k} q(A_k|\pi_k)\frac{n_k^{(v)}+\gamma}{n_k^{(\cdot)}+|A_k|\gamma}1_{v\in A_k} \\
&= (n_k^{(v)}+\gamma)\sum_{A_k:A_k\supset B_k}\frac{p(A_k|\pi_k)\Gamma(|A_k|\gamma)}{\Gamma(n_k^{(\cdot)}+1+|A_k|\gamma)}1_{v\in A_k} \\
&= (n_k^{(v)}+\gamma)\pi_k^{|B_k|}\sum_{A_k:A_k\supset B_k}\frac{\pi_k^{|A_k|-|B_k|}(1-\pi_k)^{V-|A_k|}\Gamma(|A_k|\gamma)}{\Gamma(n_k^{(\cdot)}+1+|A_k|\gamma)}1_{v\in A_k} \\
&\propto (n_k^{(v)}+\gamma)\sum_{A_k:A_k\supset B_k}\frac{\pi_k^{|A_k|-|B_k|}(1-\pi_k)^{V-|A_k|}\Gamma(|A_k|\gamma)}{\Gamma(n_k^{(\cdot)}+1+|A_k|\gamma)}1_{v\in A_k}, \\
&= (n_k^{(v)}+\gamma)\mathbb{E}\left[\frac{\Gamma(|A_k|\gamma)1_{v\in A_k}}{\Gamma(n_k^{(\cdot)}+1+|A_k|\gamma)}|A_k\supset B_k,\pi_k\right],
\end{aligned} \tag{8}$$

where $1_{v\in A_k}$ is an indicator function and we use the fact that $\Gamma(x+1)=x\Gamma(x)$. And this can be further simplified by considering the value of v.

Case 1: $v\in B_k$**.** For $B_k\subset A_k$, $1_{v\in A_k}$ is always 1. Let $X=\sum_{v\notin B_k} b_{kv} = |A_k - |B_k|$, where we immediately have $X \mid \pi_k \sim$ Bionomial$(V-|B_k|,\pi_k)$. Consequently, we have,

$$\begin{aligned}
\mathbb{E}\left[X|\pi_k\right] &= (V-|B_k|)\pi_k, \\
\text{Var}\left[X|\pi_k\right] &= (V-|B_k|)\pi_k(1-\pi_k).
\end{aligned} \tag{9}$$

Since $|A_k| = |B_k|+X$, we can write $f_k^{-w}(w=v|\pi_k)$,

$$f_k^{-w}(w=v|\pi_k) \propto (n_k^{(v)}+\gamma)\mathbb{E}\left[g_{B_k}(X)|\pi_k\right], \tag{10}$$

where $g_{B_k}(x) = \frac{\Gamma((|B_k|+x)\gamma)}{\Gamma(n_k^{(\cdot)}+1+(|B_k|+x)\gamma)}$. In theory, these quantities can be computed in close form [2]. (In [2], a different prior for $\boldsymbol{b}$ was used.) However, this is not feasible in our case, since we need to evaluate it for every word in each Gibbs sampling. Thus, we propose an efficient approximation method as follows. According to the central limit theory, we know when $V-|B_k|$ is large, which is the case in our models, the distribution of X can be well approximated by a Gaussian distribution with mean and variance defined in equation 9. Then $\mathbb{E}\left[g_{B_k}(X)|\pi_k\right]$ can be approximated by the second order Taylor expansion,

$$\mathbb{E}\left[g_{B_k}(X)|\pi_k\right] \approx g_{B_k}(\mathbb{E}\left[X|\pi_k\right]) + \frac{1}{2}g''_{B_k}(\mathbb{E}\left[X|\pi_k\right])\text{Var}\left[X|\pi_k\right],$$

where $g''_{B_k}(x)$ is the second order derivative of $g_{B_k}(x)$.

Case 2: $v \notin B_k$**.** This also means $n_k^{(v)} = 0$ according to the definition of B_k. Let $\bar{B}_k = B_k \cup \{v\}$ and $\bar{X} = \sum_{v \notin \bar{B}_k} b_{kv} = |A_k| - |\bar{B}_k|$. Similarly we have

$$\begin{aligned}
\mathbb{E}\,[\bar{X}|\pi_k] &= (V - |\bar{B}_k|)\pi_k \\
\text{Var}\,[\bar{X}|\pi_k] &= (V - |\bar{B}_k|)\pi_k(1 - \pi_k) \\
f_k^{-w}(w = v|\pi_k) &\propto \pi_k \gamma \sum_{A_k : \bar{B}_k \subset A_k} \frac{\pi_k^{|A_k| - |\bar{B}_k|}(1 - \pi_k)^{V - |A_k|}\Gamma(|A_k|\gamma)1_{v \in A_k}}{\Gamma(n_k^{(\cdot)} + 1 + |A_k|\gamma)} \\
&= \pi_k \gamma \mathbb{E}\,[g_{\bar{B}_k}(\bar{X})|\pi_k], \qquad (11)
\end{aligned}$$

where we can employ a similar approximation method as we did to compute $\mathbb{E}\,[g_{B_k}(X)|\pi_k]$.

Given all these, we can approximate conditional density function f analytically.

From the derivation process, we have also verified the fact that

$$\mathbb{E}\,[g_{B_k}(X)|\pi_k] > \pi_k \mathbb{E}\,[g_{B_k}(\bar{X})|\pi_k]. \qquad (12)$$

In addition, for HDP-LDA, the corresponding f function can be similarly derived using Equation 5 by removing π_k and setting $\boldsymbol{b}_k = \mathbf{1}$, which is

$$f_k^{-w}(w = v)_{\text{HDP-LDA}} \propto n_k^{(v)} + \gamma.$$

Finally, we observe the following fact. Let v be a term that has word assignments ($n_k^{(v)} > 0$) and v' not ($n_k^{(v')} = 0$). For HDP-LDA, the probability ratio for these two terms given its f function is

$$\text{ratio}_{\text{HDP-LDA}} = \frac{n_k^{(v)} + \gamma}{\gamma}.$$

In sparseTM, the corresponding ratio (conditioned on π_k) given its f function is

$$\text{ratio}_{\text{sparseTM}} = \frac{(n_k^{(v)} + \gamma)\mathbb{E}\,[g_{B_k}(X)|\pi_k]}{\gamma \pi_k \mathbb{E}\,[g_{B_k}(X)|\pi_k]} > \text{ratio}_{\text{HDP-LDA}}.$$

This shows that sparseTM prefers more to those terms already with word assignments.

References

[1] Teh, Y. W., M. I. Jordan, M. J. Beal, et al. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[2] Friedman, N., Y. Singer. Efficient Bayesian parameter estimation in large discrete domains. In *NIPS*. 1999.