Notes on Hierarchical Dirichlet Process for Topic Models

- Tim Hopper

- 2015-09-21

- Qadium

Yee Whye Teh et al's 2005 paper *Hierarchical Dirichlet Processes* describes a nonparametric prior for grouped clustering problems. For example, the HDP helps in generalizing the latent Dirichlet allocation model to the case the number of topics in the data are discovered by the inference algorithm instead of being specified as a parameter of the model.

The authors describe three MCMC-based algorithms for fitting HDP based models. Unfortunately, the algorithms are described somewhat vaguely and in general terms. A fair bit of mathematical leg work is required before the HDP algorithms can be applied to the specific case of nonparametric latent Dirichlet allocation.

Here are some notes I've compiled in my effort to understand these algorithms.

HDP-LDA Generative Model

The generative model for Hierarchical Dirichlet Process Latent Dirichlet Allocation is as follows:

$$\begin{aligned} H &\sim \text{Dirichlet}(\beta) \\ G_0 \,|\, \gamma, H &\sim \text{DP}(\gamma, H) \\ G_j \,|\, \alpha_0, G_0 &\sim \text{DP}(\alpha_0, G_0) \\ \theta_{ji} \,|\, G_j &\sim G_j \\ x_{ij} \,|\, \theta_{ji} &\sim \text{Categorical}(\theta_{ji}) \end{aligned} \tag{1}$$

- H is Dirichlet distribution whose dimension is the size of the vocabulary, i.e. it is distribution over an uncountably-number of term distributions (topics).
- G_0 is a distribution over a countably-infinite number of categorical term distributions, i.e. topics.
- For each document j, G_j is a is a distribution over a countably-infinite number of categorical term distributions, i.e. topics.
- θ_{ji} is a categorical distribution over terms, i.e. a topic.
- x_{ij} is a term.

To see code for sampling from this generative model, see this post.

Chinese Restaurant Franchise Approach

Instead of the above Dirichlet process model, we can think of an identical "Chinese Restaurant Franchise" model.

Each θ_{ji} is a customer in restaurant j. Each customer is sitting at a table, and each table has multiple customers.

There is a global menu of K dishes that the restaurants serve, $\phi_1, \ldots, \phi_K$.

Some other definitions:

- ψ_{jt} is the dish served at table t in restaurant j; i.e. each ψ_{jt} corresponds to some ϕ_k.
- t_{ji} is the index of the ψ_{jt} associated with θ_{ji}.
- k_{jt} is the index of ϕ_k associated with ψ_{jt}.

Customer i in restaurant j sits at table t_{ji} while table t in restaurant j serves dish k_{jt}.

There are two arrays of count variables we will want to track:

- n_{jtk} is the number of customers in restaurant j at table t eating dish k.
- m_{jk} is the number of tables in restaurant j serving dish k (multiple tables may serve the same dish).

To summarize:

x_{ij} are observed data (words). We assume $x_{ij} \sim F(\theta_{ij})$. Further, we assume θ_{ji} is associated with table t_{ji}, that is $\theta_{ji} = \psi_{jt_{ji}}$. Further, we assume the topic for table j is indexed by k_{jt}, i.e. $\psi_{jt} = \phi_{k_{jt}}$. Thus, if we know t_{ji} (the table assignment for x_{ij}) and k_{jt} (the dish assignment for table t) for all i, j, t, we could determine the remaining parameters by sampling.

Gibbs Sampling

Teh et al describe three Gibbs samplers for this model. The first and third are most applicable to the LDA application. The section helps with more complication applications of the LDA algorithm (e.g. the hidden Markov model).

5.3 Posterior sampling by direct assignment

Section 5.3 describes a direct assignment Gibbs sampler that directly assigns words to topics by augmenting the model with an assignment variable z_{ji} that is equivalent to $k_{jt_{ji}}$. This also requires a count variable m_{jk}: the number of tables in document/franchise j assigned to dish/topic k. This sampler requires

less "bookkeeping" than the algorithm from 5.1, however it requires expensive simulation or computation of recursively computed Stirling numbers of the first kind.

My notes below are derive the necessary equations for Gibbs sampling for the algorithm in section 5.1. However, they also provide most of the derivations needed for 5.3.

5.1 Posterior sampling in the Chinese restaurant franchise

Section 5.1 describes "Posterior sampling in the Chinese restaurant franchise". Given observed data $\mathbf{x}$ (i.e. documents), we sample over the index variables t_{ji} (associating tables with customers/words) and k_{jt} (associating tables with dishes/topics). Given these variables, we can reconstruct the distribution over topics for each document and distribution over words for each topic.

Notes on Implementing Algorithm 5.1

Teh et al's original HDP paper is sparse on details with regard to applying these samplers to the specific case of nonparametric LDA. For example, both samplers require computing the conditional distribution of word x_{ji} under topic k given all data items except x_{ji}: $f_k^{x_{ji}}(x_{ji})$) (eq. 30).

A Blogger Formerly Known As Shuyo has a brief post where he states (with little-to-no derivation) the equations specific to the LDA case. Below I attempt provide some of those derivations in pedantic detail.

As stated above, in the case of topic modeling, H is a Dirichlet distribution over terms distributions and F is a multinomial distribution over terms.

By definition,

$$h(\phi_k) = \frac{1}{Z} \prod_v [\phi_k]_v^{\beta - 1} \tag{2}$$

and

$$f(x_{ji} = v \,|\, \phi_k) = \phi_{kv}. \tag{3}$$

Equation (30): $f_k^{x_{ji}}(x_{ji})$

For convenience, define $v = x_{ji}$ (word i in document j), define $k = k_{jt_{ji}}$ (topic assignment for the table in document j containing word i), and

$$n_{kv}^{-ji} = |\{x_{mn} \,|\, k_{mt_{mn}} = k,\, x_{mn} = v,\, (m,n) \neq (j,i)\}| \tag{4}$$

(the number of times the term x_{ji}, besides x_{ji} itself, is generated by the same topic as was x_{ji}).

First look at the term (for fixed k):

$$\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'} \mid \phi_k) = \prod_{j'} \prod_{i' \neq i, z_{j'i'}=k} [\phi_k]_{x_{j'i'}} \tag{5}$$

$[\phi_k]_v$ is the probability that term v is generated by topic k. The double sums run over every word generated by topic k in every document. Since $[\phi_k]_{x_{j'i'}}$ is fixed for a given word w, we could instead do a product over the each word of the vocabulary:

$$\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'} \mid \phi_k) = \prod_{w \in \mathcal{V}} [\phi_k]_w^{n_{kw}^{-ji}} \tag{6}$$

We use this early on in the big derivation below.

Also, note that

$$\int \phi_{kv}^{n_{kv}^{-ji}+\beta} \prod_{w \neq v} \phi_{kw}^{n_{kw}^{-ji}+\beta-1} d\phi_k \text{ and } \int \prod_w \phi_{kw}^{n_{kw}^{-ji}+\beta-1} d\phi_k \tag{7}$$

are the normalizing coefficients for Dirichlet distributions.

Equation (30) in Teh's paper is:

$$\begin{aligned}
f_k^{-x_{ji}}(x_{ji}) &= \frac{\int f(x_{ji} \mid \phi_k) \left[\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'} \mid \phi_k)\right] h(\phi_k) d\phi_k}{\int \left[\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'} | \phi_k)\right] h(\phi_k) d\phi_k} \\
&= \frac{\int f(x_{ji} \mid \phi_k) \left[\prod_{j'} \prod_{i' \neq i, z_{j'i'}=k} \phi_{kv}\right] \cdot h(\phi_k) d\phi_k}{\int \left[\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'} | \phi_k)\right] h(\phi_k) d\phi_k} \\
&\propto \frac{\int \phi_{kv} \prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\beta-1} d\phi_k}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\beta-1} d\phi_k} \\
&= \frac{\int \phi_{kv} \cdot \phi_{kv}^{n_{kv}^{-ji}} \cdot \prod_{w \neq v} \phi_{kw}^{n_{kw}^{-ji}} \cdot \phi_{kv}^{\beta-1} \cdot \prod_{w \neq v} \phi_{kw}^{\beta-1} d\phi_k}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\beta-1} d\phi_k} \\
&= \int \phi_{kv}^{n_{kv}^{-ji}+\beta} \prod_{w \neq v} \phi_{kw}^{n_{kw}^{-ji}+\beta-1} d\phi_k \cdot \frac{1}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}+\beta-1} d\phi_k}.
\end{aligned} \tag{8}$$

Recognizing these integrals as those that occur in the Dirichlet distribution, we have,

$$
\begin{aligned}
f_k^{-x_{ji}}(x_{ji}) &= \frac{\Gamma\left(\beta + n_{kv}^{-ji} + 1\right) \prod_{w \neq v} \Gamma\left(\beta + n_{kw}^{-ji}\right)}{\Gamma\left(\sum_{w \neq v}\left[\beta + n_{kw}^{-ji}\right] + (\beta + n_{kv}^{-ji} + 1)\right)} \cdot \frac{1}{\int \prod_w \left(\phi_{kw}\right)^{n_{kw}^{-ji} + \beta - 1} d\phi_k} \\
&= \frac{\Gamma\left(\beta + n_{kv}^{-ji} + 1\right) \prod_{w \neq v} \Gamma\left(\beta + n_{kw}^{-ji}\right)}{\Gamma\left(\sum_{w \in \mathcal{V}}\left[\beta + n_{kw}^{-ji}\right] + 1\right)} \cdot \frac{1}{\int \prod_w \left(\phi_{kw}\right)^{n_{kw}^{-ji} + \beta - 1} d\phi_k} \\
&= \frac{\Gamma\left(\beta + n_{kv}^{-ji} + 1\right) \prod_{w \neq v} \Gamma\left(\beta + n_{kw}^{-ji}\right)}{\Gamma\left(\sum_{w \in \mathcal{V}}\left[\beta + n_{kw}^{-ji}\right] + 1\right)} \cdot \frac{\Gamma\left(\sum_{w \in \mathcal{V}}\left[\beta + n_{kw}^{-ji}\right]\right)}{\prod_w \Gamma\left(\beta + n_{kw}^{-ji}\right)} \\
&= \frac{\Gamma\left(\beta + n_{kv}^{-ji} + 1\right) \prod_{w \neq v} \Gamma\left(\beta + n_{kw}^{-ji}\right)}{\Gamma\left(V\beta + n_{k\cdot}^{-ji} + 1\right)} \cdot \frac{\Gamma\left(V\beta + n_{k\cdot}^{-ji}\right)}{\prod_w \Gamma\left(\beta + n_{kw}^{-ji}\right)} \\
&= \frac{\Gamma\left(\beta + n_{kv}^{-ji} + 1\right) \Gamma\left(V\beta + n_{k\cdot}^{-ji}\right)}{\Gamma\left(V\beta + n_{k\cdot}^{-ji} + 1\right)} \cdot \frac{\prod_{w \neq v} \Gamma\left(\beta + n_{kw}^{-ji}\right)}{\prod_w \Gamma\left(\beta + n_{kw}^{-ji}\right)} \\
&= \frac{\Gamma\left(\beta + n_{kn}^{-ji} + 1\right) \Gamma\left(V\beta + n_{k\cdot}^{-ji}\right)}{\Gamma\left(V\beta + n_{k\cdot}^{-ji} + 1\right) \Gamma\left(\beta + n_{kv}^{-ji}\right)} \\
&= \frac{\Gamma\left(\beta + n_{kn}^{-ji} + 1\right)}{\Gamma\left(\beta + n_{kv}^{-ji}\right)} \cdot \frac{\Gamma\left(V\beta + n_{k\cdot}^{-ji}\right)}{\Gamma\left(V\beta + n_{k\cdot}^{-ji} + 1\right)} \\
&= \frac{\beta + n_{kv}^{-ji}}{V\beta + n_{k\cdot}^{-ji}}
\end{aligned}
\tag{9}
$$

This is validated in the appendix of this paper.

Equation: $f_{k^{\mathbf{new}}}^{x_{ji}}(x_{ji})$

We also need the prior density of x_{ji} to compute the likelihood that x_{ji} will be seated at a new table.

$$
\begin{aligned}
f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) &= \int f(x_{ji} \mid \phi) h(\phi) d\phi_k \\
&= \int \phi_v \cdot \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \prod_w \phi_w^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \int \phi_v \cdot \prod_w \phi_w^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \int \phi_v^{\beta} \cdot \prod_{w \neq v} \phi_w^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \cdot \frac{\Gamma(\beta+1) \prod_{w \neq v} \Gamma(\beta)}{\Gamma(V\beta+1)} \\
&= \frac{\Gamma(V\beta)}{\Gamma(V\beta+1)} \cdot \frac{\beta \prod_w \Gamma(\beta)}{\prod_w \Gamma(\beta)} \\
&= \frac{1}{V\beta} \cdot \beta \\
&= \frac{1}{V}
\end{aligned}
\tag{10}
$$

Equation (31): $p(x_{ji} \mid \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k})$

These last two derivations give us Equation (31), the likelihood that $t_{ji} = t^{new}$:

$$
\begin{aligned}
p(x_{ji} \mid \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k}) = \sum_{k=1}^{K} \left[\frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \cdot \frac{\beta + n_{kv}^{-ji}}{V\beta + n_{k\cdot}^{-ji}} \right] \\
+ \frac{\gamma}{m_{\cdot\cdot} + \gamma} \cdot \frac{1}{V}
\end{aligned}
\tag{11}
$$

Equation (32): $p(t_{ji} = t \mid \mathbf{t}^{-ji}, \mathbf{k})$

From this, we know the conditional distribution of t_{ji} is:

$$
p(t_{ji} = t \mid \mathbf{t}^{-ji}, \mathbf{k}) \propto \begin{cases} n_{jt\cdot}^{-ji} \cdot \frac{\beta + n_{k_{jt}v}^{-ji}}{V\beta + n_{k_{jt}\cdot}^{-ji}} & \text{if } t \text{ previously used,} \\ \alpha_0 \cdot p(x_{ji} \mid \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k}) & \text{if } t = t^{\text{new}}. \end{cases} \tag{12}
$$

Equation (33): $p(k_{jt^{\mathbf{new}}} = k \mid \mathbf{t}, \mathbf{k}^{-jt^{\mathbf{new}}})$

If the sampled value of t_{ji} is t^{new}, we sample a dish $k_{jt^{\text{new}}}$ for the table with:

$$p(k_{jt^{\text{new}}} = k \,|\, \mathbf{t}, \mathbf{k}^{-jt^{\text{new}}}) \propto \begin{cases} m_{\cdot k} \cdot \frac{\beta + n_{kv}^{-ji}}{V\beta + n_{k\cdot}^{-ji}} & \text{if } k \text{ previously used,} \\ \frac{\gamma}{V} & \text{if } t = k^{\text{new}}. \end{cases} \tag{13}$$

Equation (34): $p(k_{jt} = k \,|\, \mathbf{t}, \mathbf{k}^{-jt})$

We need to sample k_{jt} (the dish/topic for table t in restaurant j):

$$p(k_{jt} = k \,|\, \mathbf{t}, \mathbf{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} \cdot f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) & \text{if } k \text{ previously used,} \\ \gamma \cdot f_{k^{\text{new}}}^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) & \text{if } t = k^{\text{new}}. \end{cases} \tag{14}$$

where $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$ is the "conditional density of $\mathbf{x}_{jt}$ given all data items associated with mixture component k leaving out $\mathbf{x}_{jt}$" (Teh, et al). ($\mathbf{x}_{jt}$ is every customer in restaurant j seated at table t). $m_{\cdot k}^{-jt}$ is the number of tables (in all franchises) serving dish k when we remove table jt.

This requires $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$; this is different from Equation (30), though they look quite similar.

$$\begin{aligned} f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) &= \frac{\int \prod_{x_{ji} \in \mathbf{x}_{jt}} f(x_{ji} \,|\, \phi_k) \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'} = k} f(x_{j'i'} \,|\, \phi_k) \right] h(\phi_k) d\phi_k}{\int \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'} = k} f(x_{j'i'} | \phi_k) \right] h(\phi_k) d\phi_k} \\ &= \frac{\int \prod_{x_{ji} \in \mathbf{x}_{jt}} \phi_{k x_{ji}} \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'} = k} \phi_{k x_{j'i'}} \right] \prod_w \phi_{kw}^{\beta - 1} d\phi_k}{\int \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'} = k} f(x_{j'i'} | \phi_k) \right] \prod_w \phi_{kw}^{\beta - 1} d\phi_k} \\ &= \frac{\int \prod_{x_{ji} \in \mathbf{x}_{jt}} \phi_{k x_{ji}} \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'} = k} \phi_{k x_{j'i'}} \right] \prod_w \phi_{kw}^{\beta - 1} d\phi_k}{\int \left[\prod_{x_{j'i'} \notin \mathbf{x_{jt}}, z_{j'i'} = k} \phi_{k x_{j'i'}} \right] \prod_w \phi_{kw}^{\beta - 1} d\phi_k} \end{aligned} \tag{15}$$

The denominator is

$$
\begin{aligned}
\text{denominator} &= \int \left[\prod_{x_{j'i'} \notin \mathbf{x_{jt}}, z_{j'i'}=k} \phi_{k x_{j'i'}} \right] \prod_w \phi_{kw}^{\beta-1} d\phi_k \\
&= \int \left[\prod_w \phi_{kw}^{n_{kw}^{-jt}} \prod_w \phi_{kw}^{\beta-1} \right] d\phi_k \\
&= \int \left[\prod_w \phi_{kw}^{n_{kw}^{-jt}+\beta-1} \right] d\phi_k \\
&= \frac{\prod_w \Gamma\left(n_{kw}^{-jt}+\beta\right)}{\Gamma\left(\sum_w n_{kw}^{-jt}+\beta\right)} \\
&= \frac{\prod_w \Gamma\left(n_{kw}^{-jt}+\beta\right)}{\Gamma\left(n_{k\cdot}^{-jt}+V\beta\right)}
\end{aligned} \tag{16}
$$

The numerator is

$$
\begin{aligned}
\text{numerator} &= \int \prod_{x_{ji} \in \mathbf{x}_{jt}} \phi_{k x_{ji}} \left[\prod_{x_{j'i'} \notin \mathbf{x}_{jt}, z_{j'i'}=k} \phi_{k x_{j'i'}} \right] \prod_w \phi_{kw}^{\beta-1} d\phi_k \\
&= \int \prod_w \phi_{kw}^{n_{kw}^{-jt}+n_{\cdot w}^{jt}+\beta+1} d\phi_k \\
&= \frac{\prod_w \Gamma\left(n_{kw}^{-jt}+n_{\cdot w}^{jt}+\beta\right)}{\Gamma\left(\sum_w n_{kw}^{-jt}+n_{\cdot w}^{jt}+\beta\right)} \\
&= \frac{\prod_w \Gamma\left(n_{kw}^{-jt}+n_{\cdot w}^{jt}+\beta\right)}{\Gamma\left(n_{k\cdot}^{-jt}+n_{\cdot\cdot}^{jt}+\beta\right)}
\end{aligned} \tag{17}
$$

This gives us a closed form version of this conditional distribution:

$$
f_k^{-\mathbf{x_{jt}}}(\mathbf{x}_{jt}) = \frac{\prod_w \Gamma\left(n_{kw}^{-jt}+n_{\cdot w}^{jt}+\beta\right)}{\prod_w \Gamma\left(n_{kw}^{-jt}+\beta\right)} \frac{\Gamma\left(n_{k\cdot}^{-jt}+V\beta\right)}{\Gamma\left(n_{k\cdot}^{-jt}+n_{\cdot\cdot}^{jt}+\beta\right)}. \tag{18}
$$

We also need the conditional distribution of k is a new dish: $f_{k^{\text{new}}}^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$. Shuyo provides without derivation:

$$
\begin{aligned}
f_{k^{\text{new}}}^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) &= \int \left[\prod_{x_{ji} \in \mathbf{x_{jt}}} f(x_{ji} \,|\, \phi) \right] h(\phi) d\phi_k \\
&= \int \prod_{x_{ji} \in \mathbf{x_{jt}}} \phi_{x_{ji}} \cdot \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \prod_w \phi_w^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \int \prod_{x_{ji} \in \mathbf{x_{jt}}} \phi_{x_{ji}} \cdot \prod_w \phi_w^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \int \prod_{x_{ji} \in \mathbf{x_{jt}}} \phi_{x_{ji}}^{(\beta+1)-1} \cdot \prod_{x_{jt} \notin \mathbf{x_{jt}}} \phi_{x_{jt}}^{\beta-1} d\phi \\
&= \frac{\Gamma(V\beta)}{\prod_w \Gamma(\beta)} \cdot \frac{\prod_{x_{ji} \in \mathbf{x_{jt}}} \Gamma(\beta+1) \prod_{x_{ji} \notin \mathbf{x_{jt}}} \Gamma(\beta)}{\Gamma(V\beta + \sum_{x_{ji} \in \mathbf{x_{jt}}} 1)} \\
&= \frac{\Gamma(V\beta) \prod_w \Gamma(\beta + n_{\cdot w}^{jt})}{\Gamma(V\beta + n_{\cdot\cdot}^{jt}) \prod_w \Gamma(\beta)}.
\end{aligned}
\tag{19}
$$

Given these equations for $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$ and $f_{k^{\text{new}}}^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$, we can draw samples from $p(k_{jt} = k \,|\, \mathbf{t}, \mathbf{k}^{-jt})$ by enumeration over topics. We now have a complete Gibbs sampler for the Posterior sampling in the Chinese restaurant franchise in Teh, et al.