Hierarchical Dirichlet Process: A Gentle Introduction

Xiaodong Yu
University of Maryland, College Park

September 13, 2009

1 Introduction

This technical report contains my notes on Teh's technical report [3] and tutorial [2], and Chapter 2 of Sudderth's thesis [1]. Many figures are credited to Teh and Sudderth. It focuses on the motivation and ideas on developing the Hierarchical Dirichlet Process (HDP) mixture model, rather than the mathematical correctness or rigorousness. Thus I try to explain the concepts and ideas in an illustrative way to make it easy to understand, for myself, at least :).

2 Introducing HDP from the perspective of sharing clusters among groups

A HDP mixture is built upon multiple DP mixtures, where each group has a DP mixture. The goal of HDP is beyond the discovering of the clusters from each single group, which is the task of DP mixture models. In many application domains, we are more interested in the relationship among the clusters from different groups. In these cases, we want to see how clusters are shared among multiple DP mixtures. Typical applications of HDP include document modeling and topic discovery in natural language process domain, object and scene modeling and recognition in computer vision domain, etc. Table 1 summarizes the relationship among DP mixture model, HDP mixture model and their finite counter parts.

Now consider how to share clusters among different groups. In the mixture models, since clusters are represented by their parameters, sharing clusters is then equivalent to sharing cluster parameters θ_k. Recall that the cluster parameters θ_k is drawn from a Dirichlet process in a DP mixture model. Can we let these Dirichlet process, G_j, be generated from the common distribution G_0 with common parameter α_0 (Figure 1.a)? If G_0 is a continuous distribution, this idea will not work: each drawing of G_j from H will be distinctive, thus this no way to force sharing θ_k among different G_j.

	DP mixture model	HDP mixture model
tasks	finding clusters in a dataset	sharing clusters among multiple groups
example applications	data clustering	discovering topics in a document corpus
finite model	Finite mixture model	Finite hierarchical mixture model with Latent Dirichlet Allocation as a special case

Table 1: Relationship among DP mixture model, HDP mixture model, Finite mixture model and LDA

The solution is thus to force G_0 to be a discrete function, then drawing of G_j from G_0 will have chances to share common θ_k's. On the other hand, we do not want to limit the number of clusters in our application for various reasons. Thus a straight-forward solution is to let G_0 be a Dirichlet process with its own prior distribution (Figure 1.b). This leads to HDP mixture model:

$$G_0|H \sim DP(\gamma, H) \quad (2.1)$$
$$G_j|G_0 \sim DP(\alpha_0, G_0) \quad (2.2)$$
$$\phi_{ji}|G_j \sim G_j \quad (2.3)$$
$$x_{ji}|\phi_{ji} \sim F(\phi_{ji}) \quad (2.4)$$

The graphical representation of HDP mixture model is illustrated in Figure 2. This figure also shows an example of HDP mixture model, which includes shared, infinite 1-D Gaussian mixtures. All clusters have unit variances and there is only one cluster parameter, i.e., θ_k is the cluster mean. $H(\lambda)$ is a conjugate, Gaussian prior on cluster means. G_0 draws a global sample of θ_k, with weight β, G_1 and G_2 reuse these samples but with different weights $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ respectively. For particular observation x_{ji}, a cluster parameter ϕ_{ji} is first drawn from G_j, then x_{ji} is drawn from $\mathcal{N}(\phi_{ji}, 1)$. In this example, ϕ_{11} and ϕ_{22} take the same value of θ_2, and shows the sharing of cluster parameters.

3 Introducing HDP from the perspective of explicit stick-breaking construction

Making G_0 discrete forces sharing clusters between G_j, because atoms θ_k in all G_j's are from G_0 whereas different G_j has different weights on these atoms, as illustrated in Figure 2. But how these weights are computed? This question

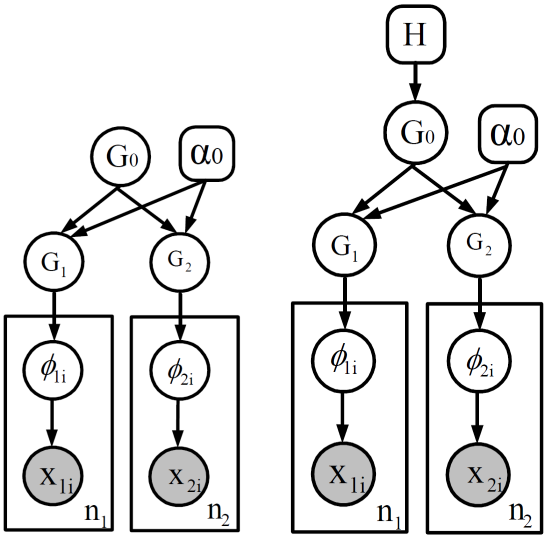

Figure 1: Two methods to extend DP mixture to deal with data of multiple groups: (a) Generating G_j from a continuous H and (b) Generating G_j from a discrete distribution G_0, which is generated from a Dirichlet process H.

can be answered when we write out the explicit formulas of G_0 and G_j:

$$\begin{aligned} G_0(\theta) &= \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \\ \beta_k &= \beta'_k \prod_{\ell=1}^{k-1} (1-\beta'_\ell) \\ \beta'_k &\sim \text{Beta}(1, \gamma) \end{aligned} \tag{3.1}$$

$$\begin{aligned} G_j(\theta) &= \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k) \\ \pi_{jk} &= \pi'_{jk} \prod_{\ell=1}^{k-1} (1-\pi'_{j\ell}) \\ \pi'_{jk} &\sim \text{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{\ell=1}^{k} \beta_\ell\right)\right) \end{aligned} \tag{3.2}$$

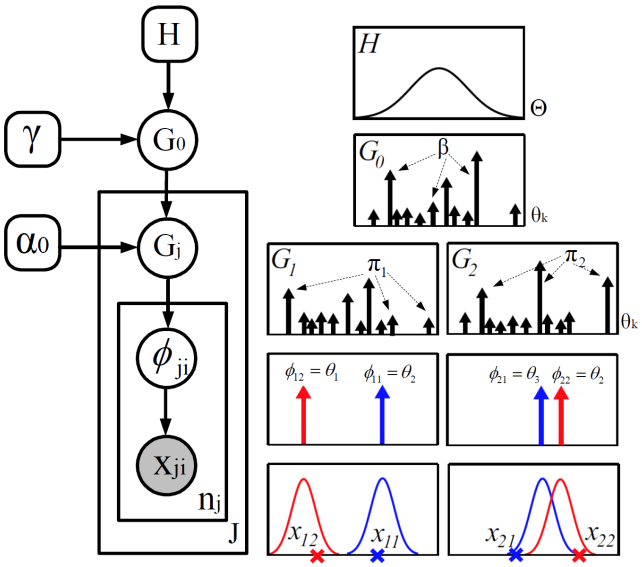

Figure 2: The graphic representations for the HDP mixture model in the Pólya urn scheme, and an example of HDP mixture model. The example is adopted from Sudderth's thesis [1]. See Section 2 for detailed discussion of this example.

Explicitly representing G_0 and G_j leads to the stick-breaking representation for HDP mixture model.

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \text{GEM}(\gamma) \\ \boldsymbol{\pi}_j|\alpha_0, \boldsymbol{\beta} &\sim \text{DP}(\alpha_0, \boldsymbol{\beta}) \\ z_{ji}|\boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \theta_k|H &\sim H \\ x_{ji}|z_{ji}, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}). \end{aligned} \tag{3.3}$$

Figure 3 illustrates the graphical representations of DP mixture model in stick-breaking construction. Compared with the finite hierarchical mixture model, it is easy to see that HDP is its infinite limit. This representation explicitly shows the global cluster parameter θ_k in the graphical representation. Its influence on observation x_{ji} is via the group-specified mixture weights $\boldsymbol{\pi}_j$. In the next section, the representation based on Chinese restaurant franchise will directly shows how the global clusters influences x_{ji}.

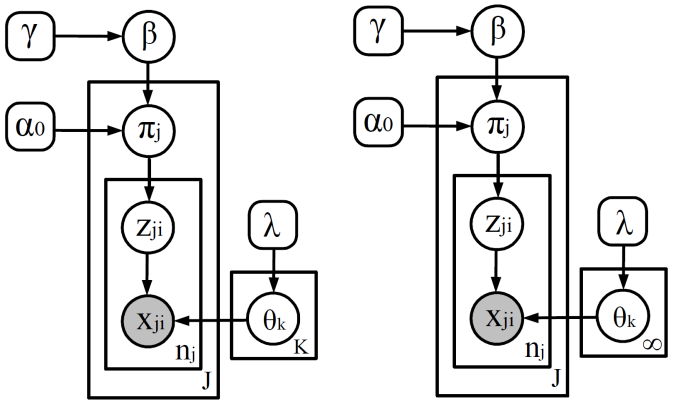

Figure 3: The graphic representation of finite hierarchical mixture model (a) and HDP mixture model (b) in stick-breaking construction.

4 Introducing HDP mixture model from the perspective of Chinese restaurant franchise metaphor

4.1 Chinese restaurant process

Chinese restaurant process is a metaphor to illustrate Dirichlet process. For a set of random variables $\phi_1, ..., \phi_i$ distributed according to $G \sim \mathrm{DP}(\alpha_0, G_0)$, the last ϕ_i has the following distribution conditioned on the previous variables:

$$\phi_i|\phi_1, .., \phi_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{n_k}{\alpha_0 + i - 1}\delta_{\theta_k} + \frac{\alpha_0}{\alpha_0 + i - 1}G_0. \tag{4.1}$$

This distribution can be described in a Chinese restaurant process metaphor:

- Imagine a Chinese restaurant that has unlimited number of tables $\theta_k, \quad k = 1, ..., \infty$.
- First customer sits at the first table.
- Suppose there are K tables occupied before the i-th customer comes. When the i-th customer comes, he can sit at:
  - Table $k \leq K$ with probability $\propto \frac{n_k}{\alpha_0+i-1}$,
  - A new table $K + 1$ with probability $\propto \frac{\alpha_0}{\alpha_0+n-1}$.

  In the first case, we set $\phi_i = \theta_k$; in the second case, we increase K to $K + 1$, draw a new sample $\theta_K \sim G_0$ and set $\phi_i = \theta_K$

variables	meaning	metaphor
ϕ_i	random variables $\phi_i \vert G \sim G$	customer i
θ_k	distinct values of ϕ_i in the given data set, $\theta_k \vert \alpha_0, G_0 \sim G_0$	table k
n_k	the number of ϕ_i associated to θ_k	the number of customers sitting around table k

Table 2: Variables involved in the Chinese restaurant process

An example that illustrates the above process in Figure 4 can be used to understand the meanings of the random variables, and the variables involved in this process are summarized in Table 2. The Chinese restaurant process illustrate the "cluster" property of the DP, i.e., the more customers sit at a table, the higher chance a new customer will choose to sit at this table and most probably, and thus only a limited number of tables will be occupied although there are unlimited number of tables in the restaurant. This property makes it feasible for us to sample from a DP mixture.

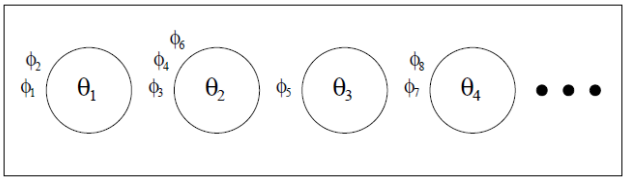

Figure 4: An example illustrating the Chinese restaurant process. This figure is adapted from Teh's technical report [3].

4.2 Chinese restaurant franchise

The Chinese restaurant franchise is essentially a two-level Chinese restaurant process:

- Within a restaurant, customers ϕ_{ji} choose tables ψ_{jt},
- Within all restaurants, tables ψ_{jt} choose dishes θ_k.

In both levels, the choosing follows the Chinese restaurant process as illustrated is the previous subsection.

The random variables and their analogies in the Chinese restaurant franchise metaphor are described in Table 3.

Formally, the Chinese restaurant franchise can be described as follows:

- Consider a Chinese restaurant franchise, whose J restaurants share a menu with unbounded number of dishes, $\theta_k, \quad k = 1, ..., \infty$.

variables	meaning	metaphor
ϕ_{ji}	random variables $\phi_{ji}\|G_j \sim G_j$	customer i in restaurant j
ψ_{jt}	distinct values of ϕ_{ji} in group j, $\psi_{jt}\|\alpha_0, G_0 \sim G_0$	table t in restaurant j
t_{ji}	index of ψ_{jt} associated to ϕ_{ji}, $t_{ji}\|\tilde{\boldsymbol{\pi}}_j \sim \tilde{\boldsymbol{\pi}}_j$	the table taken by customer i in restaurant j, .i.e. , Table(ϕ_{ji})=ψ_{jt}
n_{jt}	the number of ϕ_{ji} associated to ψ_{jt} in group j	the number of customers sitting around table t in restaurant j
θ_k	distinct values within all ψ_{jt}, $\theta_k\|H, \lambda \sim H(\lambda)$	dish k, which is shared within all restaurants
k_{jt}	index of θ_k associated to ψ_{jt}, $k_{jt}\|\boldsymbol{\beta} \sim \beta$	the dish ordered by table t in restaurant j, i.e., Dish(ψ_{jt}) = k_{jt}
m_{jk}	the number of ψ_{jt} associated to θ_k in group j	the number of tables ordered dish k in restaurant j
m_k	$\sum_k m_{jk}$, i.e., the number of ψ_{jt} associated to θ_k over all j	the total number of tables ordered dish k within all restaurants

Table 3: Variables involved in the Chinese restaurant franchise

- At each table of each restaurant, one dish is ordered from the public menu by the first customer who sits there, and it is shared among all customers who sit at that table. Multiple tables at multiple restaurants can serve the same dish.

- Suppose there are T_j tables occupied before the i-th customer comes into restaurant j and there are total K dishes has been ordered among all restaurants in the franchise in that moment. When the i-th customer comes into restaurant j, he can choose to sit at an occupied table or a new table according to the following probabilities:

  - Table $t \leq T_j$ with probability $\propto \frac{n_{jt}}{\alpha_0+i-1}$,
  - A new table $T_j + 1$ with probability $\propto \frac{\alpha_0}{\alpha_0+n-1}$.

  In the first case, we set $\phi_{ji} = \psi_{jt}$ and let $t_{ji} = t$ for the chosen t; in the second case, we increase T_j to $T_j + 1$, draw a new sample $\psi_{j,T_j} \sim G_0$ and set $\phi_{ji} = \psi_{j,T_j}$ and $t_{ji} = T_j$. This process of customers choose tables is essentially the same as the Chinese restaurant process, and it can be summarized in the following conditional probability:

$$\phi_{ji}|\phi_{j1}, .., \phi_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{\alpha_0 + i - 1}\delta_{\psi_{jt}} + \frac{\alpha_0}{\alpha_0 + i - 1}G_0. \quad (4.2)$$

- If he sits at an occupied table, he shares the dish that has been ordered at that table. If he sits at a new table, he order a dish for that table

according to its popularity among the whole franchise, while a new dish can also be tried, according to the following probabilities:

- dish $k \leq K$ $t \leq T_j$ with probability $\propto \frac{m_k}{\sum_k m_k + \gamma}$,
- a new dish $K+1$ with probability $\propto \frac{\gamma}{\sum_k m_k + \gamma}$.

In the first case, we set $\psi_{jt} = \theta_k$ and let $k_{jt} = k$ for the chosen k. In the second case, increase K to $K+1$, draw a new sample $\theta_K \sim H$ and set $\psi_{jt} = \theta_K$, $k_{jt} = K$. This process of new table choose dishes is actually another Chinese restaurant process, as long as we thing the table in this step as customer and the dish in this step as table. Similarly, this step can be summarized in the following conditional probability:

$$\psi_{jt}|\psi_{11}, \psi_{12}, .., \psi_{j1}, ..., \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k} + \frac{\gamma}{\sum_k m_k + \gamma} H. \tag{4.3}$$

A graphical example illustrating the Chinese restaurant franchise is in Figure 5.

The Chinese restaurant franchise metaphor clearly shows the two-level cluster properties: the franchise level defines the global cluster θ_k, and the restaurant level defines the local cluster ψ_{jt}. Each local cluster contains a parameter ψ_{jt} copied from some global cluster θ_k, which we indicated by $k_{jt} \sim \boldsymbol{\beta}$. Each observation x_{ji} is associated to a parameter ϕ_{ji} copied from some local cluster ψ_{jt}, which we indicated by $t_{ji} \sim \tilde{\boldsymbol{\pi}}_j$.

Based on these explicit assignments of observations to local clusters and local clusters to global clusters, we can derive a HDP graphical representation based on Chinese restaurant franchise, as in Figure 6. In the right figure in Figure 6, parameters and variables are explicitly categorized into three groups: λ and θ_k define the global clusters, γ, $\boldsymbol{\beta}$ and k_{jt} are responsible for the global clusters weights, α_0, $\tilde{\boldsymbol{\pi}}_j$ and t_{ji} are responsible for the local clusters weights. At the global level, a global probability measure $G_0 \sim \mathrm{DP}(\gamma, H)$ is defined:

$$\begin{aligned} G_0(\theta) &\sim \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \\ \boldsymbol{\beta}|\gamma &\sim \mathrm{GEM}(\gamma) \\ \theta_k|\gamma &\sim H(\gamma) \\ k_{jt}|\boldsymbol{\beta} &\sim \boldsymbol{\beta}, \end{aligned} \tag{4.4}$$

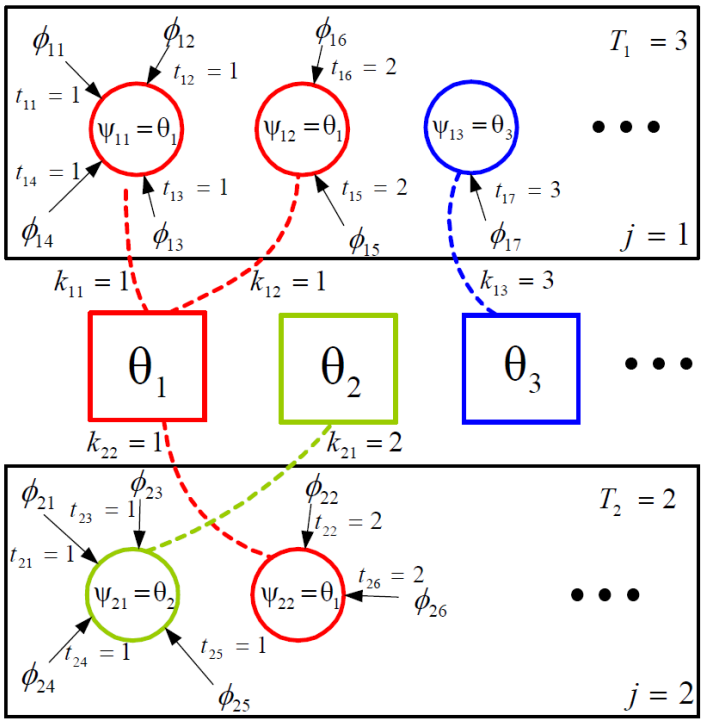

Figure 5: An example illustrating the Chinese restaurant franchise, where a franchise menu with dish θ_k (global clusters, denoted by squares at center) is shared among tables (local clusters, denoted by circles on top and bottom) in two restaurants (groups, denoted by large rectangles). All customers seated at a given table shared the same dish. k_{jt} indicates the dish ordered at table ψ_{jt}, t_{ji} indicates the table where customer ϕ_{ji} sit. This figure is adapted from Sudderth [1].

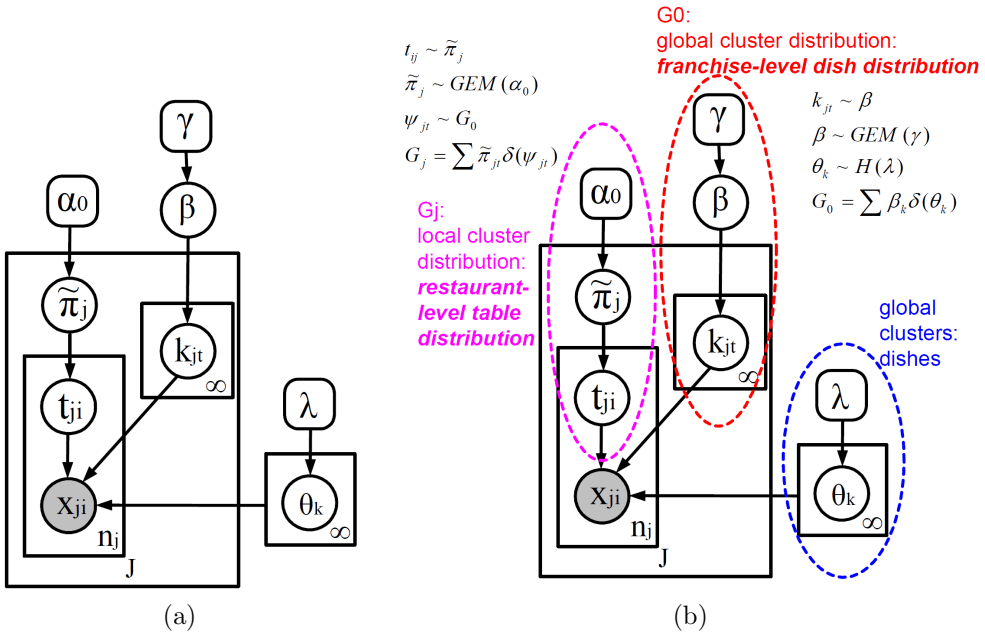

Figure 6: The graphical representation of HDP mixture model in the Chinese restaurant franchise representation.

and at the local level, group-specific mixture distribution $G_j \sim \mathrm{DP}(\alpha, G_0)$ is defined:

$$\begin{aligned} G_j(\theta) &\sim \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta(\theta, \psi_{jt}) \\ \tilde{\boldsymbol{\pi}}_j | \alpha_0 &= \mathrm{GEM}(\alpha_0) \\ \psi_{jt} | G_0 &\sim G_0 \\ t_{ji} | \tilde{\boldsymbol{\pi}}_j &\sim \tilde{\boldsymbol{\pi}}_j. \end{aligned} \tag{4.5}$$

Compare the G_j in (3.2) and (4.5), though they both are discrete distribution, their atoms and weights are different:

- The atoms of G_j in (3.2) are global clusters θ_k, while the atoms of G_j in (4.5) are local clusters ψ_{jt}.
- The atom weights in (3.2) specify the distribution of global clusters θ_k, $\boldsymbol{\pi}_j | \alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$, while the atoms weights in (4.5) specify the distribution of local clusters ψ_{jt}, $\tilde{\boldsymbol{\pi}}_j | \alpha_0 = \mathrm{GEM}(\alpha_0)$.

As we discussed above, ψ_{jt} is a copy from some θ_k which is indicated by k_{jt}. Thus, aggregating all ψ_{jt} whose values equal to θ_k in group j, we can get $\boldsymbol{\pi}_{jk}$ from $\tilde{\boldsymbol{\pi}}_{jt}$:

$$\pi_{jk} = \sum_{t | k_{jt} = k} \tilde{\pi}_{jt}. \tag{4.6}$$

Compare the HDP mixture model in stick-breaking representation in Figure 3.b and the HDP mixture model in Chinese restaurant franchise representation in Figure 6, we can find another difference between these two representations. In stick-breaking representation, the indicator z_{ji} directly indicates the *global* cluster assigned to x_{ji}, while in the Chinese restaurant franchise representation, the global cluster is indirectly indicated via local cluster indicator t_{ji}, taking $z_{ji} = k_{jt_{ji}}$.

References

[1] E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, MIT, 2006.

[2] Y. W. Teh. A Tutorial on Dirichlet Processes and Hierarchical Dirichlet Processes. available online at http://mlg.eng.cam.ac.uk/tutorials/07/ywt.pdf, May 2007.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. Technical Report 653, Department of Statistics, University of California at Berkeley, 2004.