Hierarchical Bayesian Nonparametric Models with Applications

Yee Whye Teh
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR, United Kingdom

Michael I. Jordan
Department of Statistics
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720, USA

February 14, 2009

Abstract

Hierarchical modeling is a fundamental concept in Bayesian statistics. The basic idea is that parameters are endowed with distributions which may themselves introduce new parameters, and this construction recurses. In this review we discuss the role of hierarchical modeling in Bayesian nonparametrics, focusing on models in which the infinite-dimensional parameters are treated hierarchically. For example, we consider a model in which the base measure for a Dirichlet process is itself treated as a draw from another Dirichlet process. This yields a natural recursion that we refer to as a hierarchical Dirichlet process. We also discuss hierarchies based on the Pitman-Yor process and on completely random processes. We demonstrate the value of these hierarchical constructions in a wide range of practical applications, in problems in computational biology, computer vision and natural language processing.

1 Introduction

Hierarchical modeling is a fundamental concept in Bayesian statistics. The basic idea is that parameters are endowed with distributions which may themselves introduce new parameters, and this construction recurses. A common motif in hierarchical modeling is that of the conditionally independent hierarchy, in which a set of parameters are coupled by making their distributions depend

on a shared underlying parameter. These distributions are often taken to be identical, based on an assertion of exchangeability and an appeal to de Finetti's theorem.

Hierarchies help to unify statistics, providing a Bayesian interpretation of frequentist concepts such as shrinkage and random effects. Hierarchies also provide ways to specify non-standard distributional forms, obtained as integrals over underlying parameters. They play a role in computational practice in the guise of variable augmentation. These advantages are well appreciated in the world of parametric modeling, and few Bayesian parametric modelers fail to make use of some aspect of hierarchical modeling in their work.

Nonparametric Bayesian models also typically include many classical finite-dimensional parameters, including scale and location parameters, and hierarchical modeling concepts are often invoked in specifying distributions for these parameters. For example, the Dirichlet process $\mathrm{DP}(\alpha, G_0)$ involves a concentration parameter α, which is generally given a prior distribution in nonparametric (and semiparametric) models that make use of the Dirichlet process. Moreover, the base measure, G_0, is often taken to be a parametric distribution and its parameters are endowed with prior distributions as well.

In this chapter we discuss a more thoroughgoing exploitation of hierarchical modeling ideas in Bayesian nonparametric statistics. The basic idea is that rather than treating distributional parameters such as G_0 parametrically, we treat them nonparametrically. In particular, the base measure G_0 in the Dirichlet process can itself be viewed as a random draw from some distribution on measures—specifically it can be viewed as a draw from the Dirichlet process. This yields a natural recursion that we refer to as a *hierarchical Dirichlet process*. Our focus in this chapter is on nonparametric hierarchies of this kind, where the tools of Bayesian nonparametric modeling are used recursively.

The motivations for the use of hierarchical modeling ideas in the nonparametric setting are at least as strong as they are in the parametric setting. In particular, nonparametric models involve large numbers of degrees of freedom, and hierarchical modeling ideas provide essential control over these degrees of freedom. Moreover, hierarchical modeling makes it possible to take the building blocks provided by simple stochastic processes such as the Dirichlet process and construct models that exhibit richer kinds of probabilistic structure. This breathes life into the nonparametric framework.

The chapter is organized as follows. In Section 2, we discuss the hierarchical Dirichlet process, showing how it can be used to link multiple Dirichlet processes. We present several examples of real-world applications in which such models are natural. Section 3 shows how the hierarchical Dirichlet process can be used to build nonparametric hidden Markov models; these are hidden Markov models in which the cardinality of the state space is unbounded. We also discuss extensions to nonparametric hidden Markov trees and nonparametric probabilistic context free grammars. In Section 4 we consider a different nonparametric hierarchy based on the Pitman-Yor model, showing that it is natural in domains such as natural language processing in which data often exhibit power-law behavior. Section 5 discusses the beta process, an alternative to the Dirichlet process

which yields sparse featural representations. We show that the counterpart of the Chinese restaurant process is a distribution on sparse binary matrices known as the Indian buffet process. We also consider hierarchical models based on the beta process. In Section 6, we consider some semiparametric models that are based on nonparametric hierarchies. Finally, in Section 7 we present an overview of some of the algorithms that have been developed for posterior inference in hierarchical Bayesian nonparametric models.

In all of these cases, we use practical applications to motivate these constructions and to make our presentation concrete. Our applications range from problems in biology to computational vision to natural language processing. Several of the models that we present provide state-of-the-art performance in these application domains. This wide range of successful applications serves notice as to the growing purview of Bayesian nonparametric methods.

2 Hierarchical Dirichlet Processes

The Dirichlet process (DP) is useful in models for which a component of the model is a discrete random variable of unknown cardinality. The canonical example of such a model is the DP mixture model, where the discrete variable is a cluster indicator. The *hierarchical Dirichlet process* (HDP) is useful in problems in which there are multiple groups of data, where the model for each group of data incorporates a discrete variable of unknown cardinality, and where we wish to tie these variables across groups (Teh et al., 2006). For example, the HDP mixture model allows us to share clusters across multiple clustering problems.

The basic building block of a hierarchical Dirichlet process is a recursion in which the base measure G_0 for a Dirichlet process $G \sim \mathrm{DP}(\alpha, G_0)$ is itself a draw from a Dirichlet process: $G_0 \sim \mathrm{DP}(\gamma, H)$. This recursive construction has the effect of constraining the random measure G to place its atoms at the discrete locations determined by G_0. The major application of such a construction is to the setting of conditionally independent hierarchical models of grouped data.

More formally, consider an indexed collection of DPs, $\{G_j\}$, one for each of a countable set of groups and defined on a common probability space (Θ, Ω). The hierarchical Dirichlet process ties these random measures probabilistically by letting them share their base measure and letting this base measure be random:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_j \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \qquad\qquad \text{for } j \in \mathcal{J}, \end{aligned} \tag{1}$$

where $\mathcal{J}$ is the index set. This conditionally independent hierarchical model induces sharing of atoms among the random measures G_j since each inherits its set of atoms from the same G_0. To understand the precise nature of the sharing induced by the HDP it is helpful to consider representations akin to the stick-breaking and Chinese restaurant representations of the DP. We consider these representations in the next three subsections before turning to a discussion of applications of the HDP.

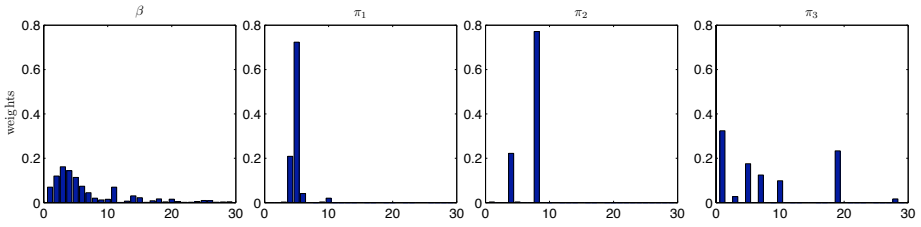

Figure 1: The HDP stick-breaking construction. The left panel depicts a draw of $\boldsymbol{\beta}$, and the remaining panels depict draws of $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$ and $\boldsymbol{\pi}_3$ conditioned on $\boldsymbol{\beta}$.

Note that the recursive construction of the HDP can be generalized to arbitrary hierarchies in the obvious way. Each G_j is given a DP prior with base measure $G_{\mathrm{pa}(j)}$, where $\mathrm{pa}(j)$ is the parent index of j in the hierarchy. As in the two-level hierarchy in Eq. (1), the set of atoms at the top level is shared throughout the hierarchy, while the multi-level hierarchy allows for a richer dependence structure on the weights of the atoms. Section 4 presents an instance of such a hierarchy in the setting of Pitman-Yor processes.

Other ways to couple multiple Dirichlet processes have been proposed in the literature; in particular the *dependent Dirichlet process* of MacEachern et al. (2001) provides a general formalism. Ho et al. (2006) gives a complementary view of the HDP and its Pitman-Yor generalizations in terms of coagulation operators. See Teh et al. (2006) and Chapter **??** for overviews.

2.1 Stick-Breaking Construction

In this section we develop a stick-breaking construction for the HDP. This representation provides a concrete representation of draws from an HDP and it provides insight into the sharing of atoms across multiple DPs.

We begin with the stick-breaking representation for the random base measure G_0, where $G_0 \sim \mathrm{DP}(\gamma, H)$. Given that this base measure is distributed according to a DP, we have (Sethuraman, 1994; Ishwaran and James, 2001, also see Section **??** in Chapter **??**):

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}, \tag{2}$$

where

$$\begin{aligned} v_k \,|\, \gamma &\sim \mathrm{Beta}(1, \gamma) \qquad\qquad \text{for } k = 1, \ldots, \infty \\ \beta_k &= v_k \prod_{l=1}^{k-1} (1 - v_l) \\ \theta_k^{**} \,|\, H &\sim H. \end{aligned} \tag{3}$$

We refer to the joint distribution on the infinite sequence $(\beta_1, \beta_2, \ldots)$ as the GEM(γ) distribution (Pitman, 2002) ("GEM" stands for Griffiths, Engen and McCloskey).

The random measures G_j are also distributed (conditionally) according to a DP. Moreover, the support of each G_j is contained within the support of G_0. Thus the stick-breaking representation for G_j is a reweighted sum of the atoms in G_0:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^{**}}. \tag{4}$$

The problem reduces to finding a relationship between the weights $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ and $\boldsymbol{\pi}_j = (\pi_{j1}, \pi_{j2}, \ldots)$. Let us interpret these weight vectors as probability measures on the discrete space $\{1, \ldots, \infty\}$. Taking partitions over integers induced by partitions on Θ, the defining property of the DP (Ferguson, 1973) implies:

$$\boldsymbol{\pi}_j \,|\, \alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}). \tag{5}$$

Some algebra then readily yields the following explicit construction for $\boldsymbol{\pi}_j$ conditioned on $\boldsymbol{\beta}$:

$$\begin{aligned} v_{jk} \,|\, \alpha, \beta_1, \ldots, \beta_k &\sim \mathrm{Beta}\left(\alpha\beta_k, \alpha\left(1 - \sum_{l=1}^{k} \beta_l\right)\right) \qquad \text{for } k = 1, \ldots, \infty \\ \pi_{jk} &= v_{jk} \prod_{l=1}^{k-1} (1 - v_{jl}). \end{aligned} \tag{6}$$

Figure 1 shows a sample draw of $\boldsymbol{\beta}$ along with draws from $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$ and $\boldsymbol{\pi}_3$ given $\boldsymbol{\beta}$.

From Eq. (3) we see that the mean of β_k is $\mathsf{E}[\beta_k] = \gamma^{k-1}(1+\gamma)^{-k}$ which decreases exponentially in k. The mean for $\boldsymbol{\pi}_j$ is simply its base measure $\boldsymbol{\beta}$; thus $\mathsf{E}[\pi_{jk}] = \mathsf{E}[\beta_k] = \gamma^{k-1}(1+\gamma)^{-k}$ as well. However the law of total variance shows that π_{jk} has higher variance than β_k: $\mathsf{Var}[\pi_{jk}] = \mathsf{E}[\frac{\beta_k(1-\beta_k)}{1+\alpha}] + \mathsf{Var}[\beta_k] > \mathsf{Var}[\beta_k]$. The higher variance is reflected in Figure 1 by the sparser nature of $\boldsymbol{\pi}_j$ relative to $\boldsymbol{\beta}$.

2.2 Chinese Restaurant Franchise

The Chinese restaurant process (CRP) describes the marginal probabilities of the DP in terms of a random partition obtained from a sequence of customers sitting at tables in a restaurant. There is an analogous representation for the HDP which we refer to as a *Chinese restaurant franchise* (CRF). In a CRF the metaphor of a Chinese restaurant is extended to a set of restaurants, one for each index in $\mathcal{J}$. The customers in the jth restaurant sit at tables in the same manner as the CRP, and this is done independently in the restaurants. The

coupling among restaurants is achieved via a franchise-wide menu. The first customer to sit at a table in a restaurant chooses a dish from the menu and all subsequent customers who sit at that table inherit that dish. Dishes are chosen with probability proportional to the number of tables (franchise-wide) which have previously served that dish.

More formally, label the ith customer in the jth restaurant with a random variable θ_{ji} that is distributed according to G_j. Similarly, let θ^*_{jt} denote a random variable corresponding to the tth table in the jth restaurant; these variables are drawn independently and identically distributed (iid) according to G_0. Finally, the dishes are iid variables θ^{**}_k distributed according to the base measure H. We couple these variables as follows. Each customer sits at one table and each table serves one dish; let customer i in restaurant j sit at table t_{ji}, and let table t serve dish k_{jt}. Then let $\theta_{ji} = \theta^*_{jt_{ji}} = \theta^{**}_{k_{jt_{ji}}}$.

Let n_{jtk} be the number of customers in restaurant j seated around table t and being served dish k, let m_{jk} be the number of tables in restaurant j serving dish k, and let K be the number of unique dishes served in the entire franchise. We denote marginal counts with dots; e.g., $n_{j\cdot k}$ is the number of customers in restaurant j served dish k.

To show that the CRF captures the marginal probabilities of the HDP, we integrate out the random measures G_j and G_0 in turn from the HDP. We start by integrating out the random measure G_j; this yields a set of conditional distributions for the θ_{ji} described by a Pólya urn scheme:

$$\theta_{ji} \,|\, \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}} + \frac{\alpha}{\alpha + n_{j\cdot\cdot}} G_0. \tag{7}$$

A draw from this mixture can be obtained by drawing from the terms on the right-hand side with probabilities given by the corresponding mixing proportions. If a term in the first summation is chosen then the customer sits at an already occupied table: we increment $n_{jt\cdot}$, set $\theta_{ji} = \theta^*_{jt}$ and let $t_{ji} = t$ for the chosen t. If the second term is chosen then the customer sits at a new table: we increment $m_{j\cdot}$ by one, set $n_{jm_{j\cdot}\cdot} = 1$, draw $\theta^*_{jm_{j\cdot}} \sim G_0$, set $\theta_{ji} = \theta^*_{jm_{j\cdot}}$ and $t_{ji} = m_{j\cdot}$.

Notice that each θ^*_{jt} is drawn iid from G_0 in the Pólya urn scheme in Eq. (7), and this is the only reference to G_0 in that equation. Thus we can readily integrate out G_0 as well, obtaining a Pólya urn scheme for the θ^*_{jt}:

$$\theta^*_{jt} \,|\, \theta^*_{11}, \ldots, \theta^*_{1m_{1\cdot}}, \ldots, \theta^*_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma + m_{\cdot\cdot}} \delta_{\theta^{**}_k} + \frac{\gamma}{\gamma + m_{\cdot\cdot}} H, \tag{8}$$

where we have presumed for ease of notation that $\mathcal{J} = \{1, \ldots, |\mathcal{J}|\}$. As promised, we see that the kth dish is chosen with probability proportional to the number of tables franchise-wide that previously served that dish ($m_{\cdot k}$).

The CRF is useful in understanding scaling properties of the clustering induced by an HDP. In a DP the number of clusters scales logarithmically (Antoniak, 1974). Thus $m_{j\cdot} \in \mathsf{O}(\alpha \log \frac{n_{j\cdot\cdot}}{\alpha})$ where $m_{j\cdot}$ and $n_{j\cdot\cdot}$ are respectively the

total number of tables and customers in restaurant j. Since G_0 is itself a draw from a DP, we have that $K \in \mathsf{O}(\gamma \log \sum_j \frac{m_{j\cdot}}{\gamma}) = \mathsf{O}(\gamma \log(\frac{\alpha}{\gamma} \sum_j \log \frac{n_{j\cdot\cdot}}{\alpha}))$. If we assume that there are J groups and that the groups (the customers in the different restaurants) have roughly the same size N, $n_{j\cdot\cdot} \in \mathsf{O}(N)$, we see that $K \in \mathsf{O}(\gamma \log \frac{\alpha}{\gamma} J \log \frac{N}{\alpha}) = \mathsf{O}(\gamma \log \frac{\alpha}{\gamma} + \gamma \log J + \gamma \log \log \frac{N}{\alpha})$. Thus the number of clusters scales doubly logarithmically in the size of each group, and logarithmically in the number of groups. The HDP thus expresses a prior belief that the number of clusters grows very slowly in N. If this prior belief is inappropriate for a given problem, there are alternatives; in particular, in Section 4.3.1 we discuss a hierarchical model that yields power-law scaling.

2.3 Posterior Structure of the HDP

The Chinese restaurant franchise is obtained by integrating out the random measures G_j and then integrating out G_0. Integrating out the random measures G_j yields a Chinese restaurant for each group as well as a sequence of iid draws from the base measure G_0, which are used recursively in integrating out G_0. Having obtained the CRF, it is of interest to derive conditional distributions that condition on the CRF; this not only illuminates the combinatorial structure of the HDP but it also prepares the ground for a discussion of inference algorithms (see Section 7), where it can be useful to instantiate the CRF explicitly.

The state of the CRF consists of the dish labels $\boldsymbol{\theta}^{**} = \{\theta^{**}_k\}_{k=1,\ldots,K}$, the table t_{ji} at which the ith customer sits, and the dish k_{jt} served at the tth table. As functions of the state of the CRF, we also have the numbers of customers $\mathbf{n} = \{n_{jtk}\}$, the numbers of tables $\mathbf{m} = \{m_{jk}\}$, the customer labels $\boldsymbol{\theta} = \{\theta_{ji}\}$ and the table labels $\boldsymbol{\theta}^* = \{\theta^*_{jt}\}$. The relationship between the customer labels and the table labels is given as follows: $\theta^*_{jt} = \theta^{**}_{jk_{jt}}$ and $\theta_{ji} = \theta^*_{jt_{ji}}$.

Consider the distribution of G_0 conditioned on the state of the CRF. G_0 is independent from the rest of the CRF when we condition on the iid draws $\boldsymbol{\theta}^*$, because the restaurants interact with G_0 only via the iid draws. The posterior thus follows from the usual posterior for a DP given iid draws:

$$G_0 \,|\, \gamma, H, \boldsymbol{\theta}^* \sim \mathrm{DP}\left(\gamma + m_{\cdot\cdot}, \frac{\gamma H + \sum_{k=1}^K m_{\cdot k} \delta_{\theta^{**}_k}}{\gamma + m_{\cdot\cdot}}\right). \tag{9}$$

Note that values for $\mathbf{m}$ and $\boldsymbol{\theta}^{**}$ are determined given $\boldsymbol{\theta}^*$, since they are simply the unique values and their counts among $\boldsymbol{\theta}^*$[1]. A draw from Eq. (9) can be constructed as follows (using the defining property of a DP):

$$\begin{aligned} \beta_0, \beta_1, \ldots, \beta_K \,|\, \gamma, G_0, \boldsymbol{\theta}^* &\sim \mathrm{Dirichlet}(\gamma, m_{\cdot 1}, \ldots, m_{\cdot K}) \\ G'_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_0 &= \beta_0 G'_0 + \sum_{k=1}^K \beta_k \delta_{\theta^{**}_k}. \end{aligned} \tag{10}$$

[1]Here we make the simplifying assumption that H is a continuous distribution so that draws from H are unique. If H is not continuous then additional bookkeeping is required.

We see that the posterior for G_0 is a mixture of atoms corresponding to the dishes and an independent draw from $\mathrm{DP}(\gamma, H)$.

Conditioning on this draw of G_0 as well as the state of the CRF, the posteriors for the G_j are independent. In particular, the posterior for each G_j follows from the usual posterior for a DP, given its base measure G_0 and iid draws $\boldsymbol{\theta}_j$:

$$G_j \,|\, \alpha, G_0, \boldsymbol{\theta}_j \sim \mathrm{DP}\left(\alpha + n_{j\cdot\cdot}, \frac{\alpha G_0 + \sum_{k=1}^{K} n_{j\cdot K}\delta_{\theta_k^{**}}}{\alpha + n_{j\cdot\cdot}}\right). \tag{11}$$

Note that $\mathbf{n}_j$ and $\boldsymbol{\theta}^{**}$ are simply the unique values and their counts among the $\boldsymbol{\theta}_j$. Making use of the decomposition of G_0 into G_0' and atoms located at the dishes $\boldsymbol{\theta}^{**}$, a draw from Eq. (11) can thus be constructed as follows:

$$\begin{aligned} \pi_{j0}, \pi_{j1}, \ldots, \pi_{jK} \,|\, \alpha, \boldsymbol{\theta}_j &\sim \mathrm{Dirichlet}(\alpha\beta_0, \alpha\beta_1 + n_{j\cdot 1}, \ldots, \alpha\beta_K + n_{j\cdot K}) \\ G_j' \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha\beta_0, G_0') \\ G_j &= \pi_{j0}G_j' + \sum_{k=1}^{K} \pi_{jk}\delta_{\theta_k^{**}}. \end{aligned} \tag{12}$$

We see that G_j is a mixture of atoms at θ_k^{**} and an independent draw from a DP, where the concentration parameter depends on β_0.

The posterior over the entire HDP is obtained by averaging the conditional distributions of G_0 and G_j over the posterior state of the Chinese restaurant franchise given $\boldsymbol{\theta}$.

This derivation shows that the posterior for the HDP can be split into a "discrete part" and a "continuous part." The discrete part consists of atoms at the unique values $\boldsymbol{\theta}^{**}$, with different weights on these atoms for each DP. The continuous part is a separate draw from an HDP with the same hierarchical structure as the original HDP and global base measure H, but with altered concentration parameters. The continuous part consists of an infinite series of atoms at locations drawn iid from H. Although we have presented this posterior representation for a two-level hierarchy, the representation extends immediately to general hierarchies.

2.4 Applications of the HDP

In this section we consider several applications of the HDP. These models use the HDP at different depths in an overall Bayesian hierarchy. In the first example the random measures obtained from the HDP are used to generate data directly, and in the second and third examples these random measures generate latent parameters.

2.4.1 Information Retrieval

The growth of modern search engines on the World Wide Web has brought new attention to a classical problem in the field of information retrieval (IR)—how

should a collection of documents be represented so that relevant documents can be returned in response to a query? IR researchers have studied a wide variety of representations and have found empirically that a representation known as *term frequency-inverse document frequency*, or "tf-idf," yields reasonably high-quality rankings of documents (Salton and McGill, 1983). The general intuition is that the relevance of a document to a query should be proportional to the frequency of query terms it contains ("term frequency"), but that query terms that appear in many documents should be downweighted since they are less informative ("inverse document frequency").

Cowans (2004, 2006) has shown that the HDP provides statistical justification for the intuition behind tf-idf. Let x_{ji} denote the ith word in the jth document in some corpus of documents, where the range of x_{ji} is a discrete *vocabulary* Θ. Consider the following simple model for documents:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) && \\ G_j \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) && \text{for } j \in \mathcal{J} \\ x_{ji} \,|\, G_j &\sim G_j && \text{for } i = 1, \ldots, n_j, \end{aligned} \tag{13}$$

where H is the global probability measure over the vocabulary Θ and where n_j is the number of words in the jth document. (Note that $n_j = n_{j\cdot\cdot}$ where the latter refers to the general notation introduced in Section 2.2; here and elsewhere we use n_j as a convenient shorthand.) In this model, G_j is a discrete measure over the vocabulary associated with document j and G_0 is a discrete measure over the vocabulary that acts to tie together word usages across the corpus. The model is presented as a graphical model in the left panel of Figure 2.

Integrating out G_0 and the G_j as discussed in Section 2.2, we obtain the following marginal probabilities for words $\theta \in \Theta$ in the jth document:

$$\begin{aligned} p_j(\theta) &= \frac{n'_{j\theta} + \alpha p_0(\theta)}{n'_{j\cdot} + \alpha} \\ p_0(\theta) &= \frac{m'_{\cdot\theta} + \gamma H(\theta)}{m'_{\cdot\cdot} + \gamma}, \end{aligned} \tag{14}$$

where $n'_{j\theta}$ is the term frequency—the number of occurrences of θ in document j—and $m'_{j\theta}$ is the number of tables serving dish θ in restaurant j in the CRF representation. (Note that the need for the specialized "prime" notation in this case is driven by the fact that Θ is a discrete space in this example. In particular, for each $\theta \in \Theta$ there may be multiple k such that $\theta^{**}_k = \theta$. The term frequency $n'_{j\theta} = \sum_{k:\theta^{**}_k = \theta} n_{j\cdot k}$ is the number of customers eating dish θ regardless of which menu entry they picked. Similarly, $m'_{j\theta} = \sum_{k:\theta^{**}_k = \theta} m_{jk\cdot}$.)

If we make the approximation that the number of tables serving a particular dish in a particular restaurant is at most one, then $m'_{\cdot\theta}$ is the document frequency—the number of documents containing word θ in the corpus. We now rank documents by computing a "relevance score" $R(j, Q)$—the log probability

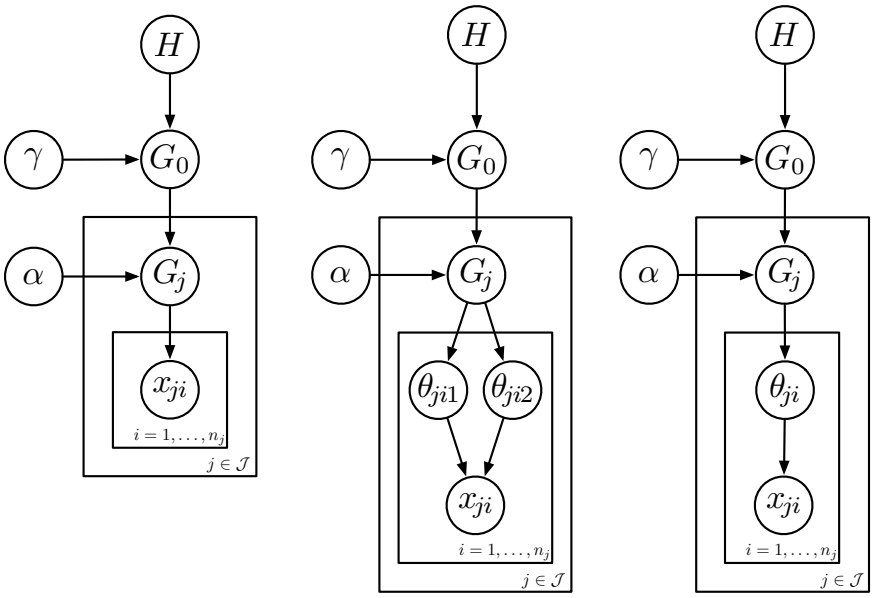

Figure 2: Graphical representations of HDP-based models. Left: An HDP model for information retrieval. Center: An HDP mixture model for haplotype phasing. Right: The HDP-LDA model for topic or admixture modeling.

of a query Q under each document j:

$$R(j,Q) = \sum_{\theta \in Q} \log p_j(\theta) \tag{15}$$
$$= \sum_{\theta \in Q} \left(\log \left(1 + \frac{n'_{j\theta}}{\alpha \frac{m'_{\cdot\theta} + \gamma H(\theta)}{m'_{\cdot\cdot} + \gamma}} \right) - \log(n'_{j\cdot} + \alpha) + \log(\alpha p_0(\theta)) \right).$$

In this score the first term is akin to a tf-idf score, the second term is a normalization penalizing large documents, and the third term can be ignored as it does not depend on document identity j. Thus we see that a simple application of the HDP provides a principled justification for the use of inverse document frequency and document length normalization. Moreover, in small-scale experiments, Cowans (2004, 2006) found that this score improves upon state-of-the-art relevance scores (Robertson et al., 1992; Hiemstra and Kraaij, 1998).

2.4.2 Multi-Population Haplotype Phasing

We now consider a class of applications in which the HDP provides a distribution on latent parameters rather than on the observed data.

Haplotype phasing is an interesting problem in statistical genetics that can be formulated as a mixture model (Stephens et al., 2001). Consider a set of M binary markers along a chromosome. Chromosomes come in pairs for humans, so let θ_{i1} and θ_{i2} denote the binary-valued vectors of markers for a pair of chromosomes for the ith individual. These vectors are referred to as *haplotypes*, and the elements of these vectors are referred to as *alleles*. A *genotype* x_i is a vector which records the unordered pair of alleles for each marker; that is, the association of alleles to chromosome is lost. The *haplotype phasing* problem is to restore haplotypes (which are useful for predicting disease associations) from genotypes (which are readily assayed experimentally whereas haplotypes are not).

Under standard assumptions from population genetics, we can write the probability of the ith genotype as a mixture model:

$$p(x_i) = \sum_{\theta_{i1}, \theta_{i2} \in \mathcal{H}} p(\theta_{i1}) p(\theta_{i2}) p(x_i \,|\, \theta_{i1}, \theta_{i2}), \tag{16}$$

where $\mathcal{H}$ is the set of haplotypes in the population and where $p(x_i \,|\, \theta_{i1}, \theta_{i2})$ reflects the loss of order information as well as possible measurement error. Given that the cardinality of $\mathcal{H}$ is unknown, this problem is naturally formulated as a DP mixture modeling problem where a "cluster" is a haplotype (Xing et al., 2007).

Let us now consider a multi-population version of the haplotype phasing problem in which the genotype data can be classified into (say) Asian, European and African subsets. Here it is natural to attempt to identify the haplotypes in each population and to share these haplotypes among populations. This can be

achieved with the following HDP mixture model:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) && (17)\\ G_j \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) && \text{for each population } j \in \mathcal{J} \\ \theta_{ji1}, \theta_{ji2} \,|\, G_j &\overset{\text{iid}}{\sim} G_j && \text{for each individual } i = 1, \ldots, n_j \\ x_{ji} \,|\, \theta_{ji1}, \theta_{ji2} &\sim F_{\theta_{ji1}, \theta_{ji2}}, \end{aligned}$$

where $\theta_{ji1}, \theta_{ji2}$ denote the pair of haplotypes for the ith individual in the jth population. The model is presented as a graphical model in the center panel of Figure 2. Xing et al. (2006) showed that this model performs effectively in multi-population haplotype phasing, outperforming methods that lump together the multiple populations or treat them separately.

2.4.3 Topic Modeling

A *topic model* or *mixed membership model* is a generalization of a finite mixture model in which each data point is associated with multiple draws from a mixture model, not a single draw (Blei et al., 2003; Erosheva, 2003). As we will see, while the DP is the appropriate tool to extend finite mixture models to the nonparametric setting, the appropriate tool for nonparamametric topic models is the HDP.

To motivate the topic model formulation, consider the problem of modeling the word occurrences in a set of newspaper articles (e.g., for the purposes of classifying future articles). A simple clustering methodology might attempt to place each article in a single cluster. But it would seem more useful to be able to cross-classify articles according to "topics"; for example, an article might be mainly about Italian food, but it might also refer to health, history and the weather. Moreover, as this example suggests, it would be useful to be able to assign numerical values to the degree to which an article treats each topic.

Topic models achieve this goal as follows. Define a *topic* to be a probability distribution across a set of *words* taken from some vocabulary W. A *document* is modeled as a probability distribution across topics. In particular, let us assume the following generative model for the words in a document. First choose a probability vector $\boldsymbol{\pi}$ from the K-dimensional simplex, and then repeatedly (1) select one of the K topics with probabilities given by the components of $\boldsymbol{\pi}$ and (2) choose a word from the distribution defined by the selected topic. The vector $\boldsymbol{\pi}$ thus encodes the expected fraction of words in a document that are allocated to each of the K topics. In general a document will be associated with multiple topics.

Another natural example of this kind of problem arises in statistical genetics. Assume that for each individual in a population we can assay the state of each of M *markers*, and recall that the collection of markers for a single individual is referred to as a *genotype*. Consider a situation in which K *subpopulations* which have hitherto remained separate are now thoroughly mixed (i.e., their mating patterns are that of a single population). Individual genotypes will now have portions that arise from the different subpopulations. This

is referred to as "admixture." We can imagine generating a new admixed genotype by fixing a distribution $\boldsymbol{\pi}$ across subpopulations and then repeatedly (1) choosing a subpopulation according to $\boldsymbol{\pi}$ and (2) choosing the value of a marker (an "allele") from the subpopulation-specific distribution on the alleles for that marker. This formulation is essentially isomorphic to the document modeling formulation. (The difference is that in the document setting the observed words are generally assumed to be exchangeable, whereas in the genetics setting each marker has its own distribution over alleles).

To fully specify a topic model we require a distribution for $\boldsymbol{\pi}$. Taking this distribution to be symmetric Dirichlet, we obtain the *latent Dirichlet allocation* (LDA) model, developed by Blei et al. (2003) and Pritchard et al. (2000) as a model for documents and admixture, respectively. This model has been widely used not only in the fields of information retrieval and statistical genetics, but also in computational vision, where a "topic" is a distribution across visual primitives, and an image is modeled as a distribution across topics (Fei-Fei and Perona, 2005).

Let us now turn to the problem of developing a Bayesian nonparametric version of LDA in which the number of topics is allowed to be open-ended. As we have alluded to, this requires the HDP, not merely the DP. To see this, consider the generation of a single word in a given document. According to LDA, this is governed by a finite mixture model, in which one of K topics is drawn and then a word is drawn from the corresponding topic distribution. Generating all of the words in a single document requires multiple draws from this finite mixture. If we now consider a different document, we again have a finite mixture, with the same mixture components (the topics), but with a different set of mixing proportions (the document-specific vector $\boldsymbol{\pi}$). Thus we have multiple finite mixture models. In the nonparametric setting they must be linked so that the same topics can appear in different documents.

We are thus led to the following model, which we refer to as HDP-LDA:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) && (18)\\ G_j \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) && \text{for each document } j \in \mathcal{J} \\ \theta_{ji} \,|\, G_j &\sim G_j && \text{for each word } i = 1, \ldots, n_j \\ x_{ji} \,|\, \theta_{ji} &\sim F_{\theta_{ji}}, \end{aligned}$$

where x_{ji} is the ith word in document j, H is the prior distribution over topics and $F_{\theta_{ji}}$ is the distribution over words. The model is presented as a graphical model in the right panel of Figure 2. Note that the atoms present in the random distribution G_0 are shared among the random distributions G_j. Thus, as desired, we have a collection of tied mixture models, one for each document.

Topic models can be generalized in a number of other directions. For example, in applications to document modeling it is natural to ask that topics occur at multiple levels of resolution. Thus, at a high level of resolution, we might wish to obtain topics that give high probability to words that occur throughout the documents in a corpus, while at a lower level we might wish to find topics that are focused on words that occur in specialized subsets of the documents.

A Bayesian nonparametric approach to obtaining this kind of *abstraction hierarchy* has been presented by Blei et al. (2004). In the model presented by these authors, topics are arranged into a tree, and a document is modeled as a path down the tree. This is achieved by defining the tree procedurally in terms of a linked set of Chinese restaurants.

3 Hidden Markov Models with Infinite State Spaces

Hidden Markov models (HMMs) are widely used to model sequential data and time series data (Rabiner, 1989). An HMM is a doubly-stochastic Markov chain in which a state sequence, $\theta_1, \theta_2, \ldots, \theta_\tau$, is drawn according to a Markov chain on a discrete state space Θ with transition kernel $\pi(\theta_t, \theta_{t+1})$. A corresponding sequence of observations, $x_1, x_2, \ldots, x_\tau$, is drawn conditionally on the state sequence, where for all t the observation x_t is conditionally independent of the other observations given the state θ_t. We let $F_{\theta_t}(x_t)$ denote the distribution of x_t conditioned on the state θ_t; this is referred to as the "emission distribution."

In this section we show how to use Bayesian nonparametric ideas to obtain an "infinite HMM"—an HMM with a countably infinite state space (Beal et al., 2002; Teh et al., 2006). The idea is similar in spirit to the passage from a finite mixture model to a DP mixture model. However, as we show, the appropriate nonparametric tool is the HDP, not the DP. The resulting model is thus referred to as the *hierarchical Dirichlet process hidden Markov model* (HDP-HMM). We present both the HDP formulation and a stick-breaking formulation in this section; the latter is particularly helpful in understanding the relationship to finite HMMs. It is also worth noting that a Chinese restaurant franchise (CRF) representation of the HDP-HMM can be developed, and indeed Beal et al. (2002) presented a precursor to the HDP-HMM that was based on an urn model akin to the CRF.

To understand the need for the HDP rather than the DP, note first that a classical HMM specifies a set of finite mixture distributions, one for each value of the current state θ_t. Indeed, given θ_t, the observation x_{t+1} is chosen by first picking a state θ_{t+1} and then choosing x_{t+1} conditional on that state. Thus the transition probability $\pi(\theta_t, \theta_{t+1})$ plays the role of a mixing proportion and the emission distribution F_{θ_t} plays the role of the mixture component. It is natural to consider replacing this finite mixture model by a DP mixture model. In so doing, however, we must take into account the fact that we obtain a *set* of DP mixture models, one for each value of the current state. If these DP mixture models are not tied in some way, then the set of states accessible in a given value of the current state will be disjoint from those accessible for some other value of the current state. We would obtain a branching structure rather than a chain structure. The solution to this problem is straightforward—we use the HDP to tie the DPs.

More formally, let us consider a collection of random transition kernels,

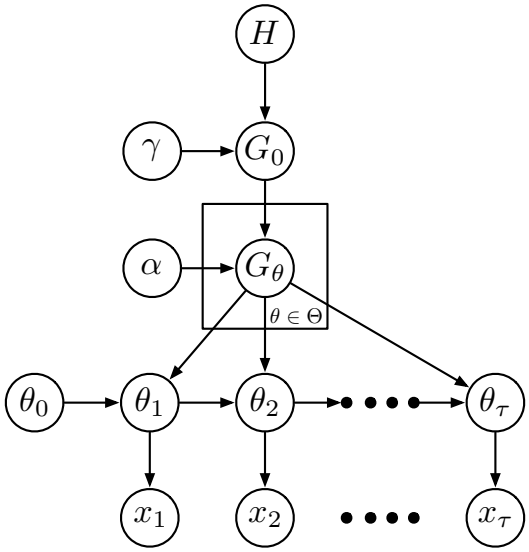

Figure 3: HDP hidden Markov model.

$\{G_\theta : \theta \in \Theta\}$, drawn from an HDP:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_\theta \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \qquad\qquad \text{for } \theta \in \Theta, \end{aligned} \tag{19}$$

where H is a base measure on the probability space $(\Theta, \mathcal{T})$. As we shall see, the random base measure G_0 allows the transitions out of each state to share the same set of next states. Let $\theta_0 = \theta_0^{**} \in \Theta$ be a predefined initial state. The conditional distributions of the sequence of latent state variables $\theta_1, \ldots, \theta_\tau$ and observed variables $x_1, \ldots, x_\tau$ are:

$$\begin{aligned} \theta_t \,|\, \theta_{t-1}, G_{\theta_{t-1}} &\sim G_{\theta_{t-1}} \qquad\qquad \text{for } t = 1, \ldots, \tau \\ x_t \,|\, \theta_t &\sim F_{\theta_t}. \end{aligned} \tag{20}$$

A graphical model representation for the HDP-HMM is shown in Figure 3.

We have defined a probability model consisting of an uncountable number of DPs, which may raise measure-theoretic concerns. These concerns can be dealt with, however, essentially due to the fact that the sample paths of the HDP-HMM only ever encounter a finite number of states. To see this more clearly, and to understand the relationship of the HDP-HMM to the parametric HMM, it is helpful to consider a stick-breaking representation of the HDP-HMM. This representation is obtained directly from the stick-breaking representation of the

underlying HDP:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}} \tag{21}$$
$$G_{\theta_l^{**}} = \sum_{k=1}^{\infty} \pi_{\theta_l^{**}k} \delta_{\theta_k^{**}} \qquad \text{for } l = 0, 1, \ldots, \infty,$$

where

$$\begin{aligned} \theta_k^{**} \,|\, H &\sim H && \text{for } k = 1, \ldots, \infty \\ \boldsymbol{\beta} \,|\, \gamma &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\pi}_{\theta_k^{**}} \,|\, \alpha, \boldsymbol{\beta} &\sim \mathrm{DP}(\alpha, \boldsymbol{\beta}). \end{aligned} \tag{22}$$

The atoms θ_k^{**} are shared across G_0 and the transition distributions $G_{\theta_l^{**}}$. Since all states visited by the HMM are drawn from the transition distributions, the states possibly visited by the HMM with positive probability (given G_0) will consist only of the initial state θ_0^{**} and the atoms $\theta_1^{**}, \theta_2^{**}, \ldots$. Relating to the parametric HMM, we see that the transition probability from state θ_l^{**} to state θ_k^{**} is given by $\pi_{\theta_l^{**}k}$ and the distribution on the observations is given by $F_{\theta_k^{**}}$.

This relationship to the parametric HMM can be seen even more clearly if we identify the state θ_k^{**} with the integer k, for $k = 0, 1, \ldots, \infty$, and if we introduce integer-valued variables z_t to denote the state at time t. In particular, if $\theta_t = \theta_k^{**}$ is the state at time t, we let z_t take on value k, and write $\boldsymbol{\pi}_k$ instead of $\boldsymbol{\pi}_{\theta_k^{**}}$. The HDP-HMM can now be expressed as:

$$\begin{aligned} z_t \,|\, z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} &\sim \boldsymbol{\pi}_{z_{t-1}} \\ x_t \,|\, z_t, \theta_{z_t}^{**} &\sim F_{\theta_{z_t}^{**}}, \end{aligned} \tag{23}$$

with priors on the parameters and transition probabilities given by Eq. (23). This construction shows explicitly that the HDP-HMM can be interpreted as an HMM with a countably infinite state space.

A difficulty with the HDP-HMM as discussed thus far is that it tends to be poor at capturing state persistence; it has a tendency to create redundant states and rapidly switch among them. This may not be problematic for applications in which the states are nuisance variables and it is overall predictive likelihood that matters, but it can be problematic for segmentation or parsing applications in which the states are the object of inference and when state persistence is expected. This problem can be solved by giving special treatment to self-transitions. In particular, let G_θ denote the transition kernel associated with state θ. Fox et al. (2009) proposed the following altered definition of G_θ (compare to Eq. (19)):

$$G_\theta \,|\, \alpha, \kappa, G_0, \theta \sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha G_0 + \kappa \delta_\theta}{\alpha + \kappa}\right), \tag{24}$$

where δ_θ is a point mass at θ and where κ is a parameter that determines the extra mass placed on a self-transition. To see in more detail how this affects state persistence, consider the stick-breaking weights $\boldsymbol{\pi}_{\theta_k^{**}}$ associated with one of the countably many states θ_k^{**} that can be visited by the HMM. The stick-breaking representation of $G_{\theta_k^{**}}$ is altered as follows (compare to Eq. (23)):

$$\boldsymbol{\pi}_{\theta_k^{**}} \mid \alpha, \boldsymbol{\beta}, \kappa \sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_{\theta_k^{**}}}{\alpha + \kappa}\right). \tag{25}$$

Fox et al. (2009) further place a vague gamma prior on $\alpha + \kappa$ and a beta prior on $\kappa/(\alpha + \kappa)$. The hyperparameters of these distributions allow prior control of state persistence. See also Beal et al. (2002), who develop a related prior within the framework of their hierarchical urn scheme.

3.1 Applications of the HDP-HMM

In the following sections we describe a number of applications and extensions of the HDP-HMM. An application that we will not discuss but is worth mentioning is the application of HDP-HMMs to the problem of modeling recombination hotspots and ancestral haplotypes for short segments of single nucleotide polymorphisms (Xing and Sohn, 2007).

3.1.1 Speaker Diarization

Speech recognition has been a major application area for classical parametric HMMs (Huang et al., 2001). In a typical application, several dozen states are used, roughly corresponding to the number of phoneme-like segments in speech. The observations x_t are spectral representations of speech over short time slices.

In many applications, however, the number of states is more fundamentally part of the inferential problem and it does not suffice to simply fix an arbitrary value. Consider an audio recording of a meeting in which the number of people participating in the meeting is unknown a priori. The problem of *speaker diarization* is that of segmenting the audio recording into time intervals associated with individual speakers (Wooters and Huijbregts, 2007). Here it is natural to consider an HDP-HMM model, where a state corresponds to an individual speaker and the observations are again short-term spectral representations. Posterior inference in the HDP-HMM yields estimates of the spectral content of each speaker's voice, an estimate of the number of speakers participating in the meeting, and a diarization of the audio stream.

Such an application of the HDP-HMM has been presented by Fox et al. (2009), who showed that the HDP-HMM approach yielded a state-of-the-art diarization method. A noteworthy aspect of their work is that they found that the special treatment of self-transitions discussed in the previous section was essential; without this special treatment the HDP-HMM's tendency to rapidly switch among redundant states led to poor speaker diarization performance.

3.1.2 Word Segmentation

As another application of the HDP-HMM to speech, consider the problem of segmenting an audio stream into a sequence of words. Speech is surprisingly continuous with few obvious breaks between words and the problem of *word segmentation*—that of identifying coherent segments of "words" and their boundaries in continuous speech—is nontrivial. Goldwater et al. (2006b) proposed a statistical approach to word segmentation based upon the HDP-HMM. The latent states of the HMM correspond to words. An HDP-HMM rather than a parametric HMM is required for this problem, since there are an unbounded number of potential words.

In the model, an utterance is viewed as a sequence of phonemes, $\rho_1, \rho_2, \ldots, \rho_\tau$. The sequence is modeled by an HDP-HMM in which words are the latent states. A word is itself a sequence of phonemes. The model specification is as follows. First, the number of words n is drawn from a geometric distribution. Then a sequence of n words, $\theta_1, \theta_2, \ldots, \theta_n$, is drawn from an HDP-HMM:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) && (26)\\ G_\theta \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) && \text{for } \theta \in \Theta \\ \theta_i \,|\, \theta_{i-1}, G_{\theta_{i-1}} &\sim G_{\theta_{i-1}} && \text{for } i = 1, \ldots, n. \end{aligned}$$

where $\theta_0 \sim G_\emptyset$ is a draw from an initial state distribution. Each G_θ is the transition distribution over next words, given the previous word θ. This is defined for every possible word θ, with Θ the set of all possible words (including the empty word θ_0 which serves as an initial state for the Markov chain). The base measure H over words is a simple independent phonemes model: the length of the word, $l \geq 1$, is first drawn from another geometric distribution, then each phoneme r_i is drawn independently from a prior over phonemes:

$$H(\theta = (r_1, r_2, \ldots, r_l)) = \eta_0(1-\eta_0)^{l-1} \prod_{t=1}^{l} H_0(r_t), \tag{27}$$

where H_0 is a probability measure over individual phonemes. The probability of the observed utterance is then a sum over probabilities of sequences of words such that their concatenation is $\rho_1, \rho_2, \ldots, \rho_\tau$.

Goldwater et al. (2006b) have shown that this HDP-HMM approach leads to significant improvements in segmentation accuracy.

3.1.3 Trees and Grammars

A number of other structured probabilistic objects are amenable to a nonparametric treatment based on the HDP. In this section we briefly discuss some recent developments which go beyond the chain-structured HMM to consider objects such as trees and grammars.

A *hidden Markov tree* (HMT) is a directed tree in which the nodes correspond to states, and in which the probability of a state depends (solely) on its unique

parent in the tree. To each state there is optionally associated an observation, where the probability of the observation is conditionally independent of the other observations given the state (Chou et al., 1994).

We can generalize the HDP-HMM to a *hierarchical Dirichlet process hidden Markov tree* (HDP-HMT) model in which the number of states is unbounded. This is achieved by a generalization of the HDP-HMM model in which the transition matrix along each edge of the HMT is replaced with sets of draws from a DP (one draw for each row of the transition matrix) and these DPs are tied with the HDP. This model has been applied to problems in image processing (denoising, scene recognition) in which the HDP-HMT is used to model correlations among wavelet coefficients in multiresolution models of images (Kivinen et al., 2007a,b).

As a further generalization of the HDP-HMM, several groups have considered nonparametric versions of probabilistic grammars (Johnson et al., 2007; Liang et al., 2007; Finkel et al., 2007). These grammars consist of collections of rules, of the form $A \to BC$, where this transition from a symbol A to a pair of symbols BC is modeled probabilistically. When the number of grammar symbols is unknown a priori, it is natural to use the HDP to generate symbols and to tie together the multiple occurrences of these symbols in a parse tree.

4 Hierarchical Pitman-Yor Processes

As discussed in Chapter **??**, a variety of alternatives to the DP have been explored in the Bayesian nonparametrics literature. These alternatives can provide a better fit to prior beliefs than the DP. It is therefore natural to consider hierarchical models based on these alternatives. In this section we shall describe one such hierarchical model, the hierarchical Pitman-Yor (HPY) process, which is based on the Pitman-Yor process (also known as the two-parameter Poisson-Dirichlet process). We briefly describe the Pitman-Yor process here; Section **??** in Chapter **??** as well as Perman et al. (1992), Pitman and Yor (1997) and Ishwaran and James (2001) present further material on the Pitman-Yor process. In Section 4.3.1 we describe an application of the HPY process to language modeling. Section 4.3.2 presents a spatial extension of the HPY process and an application to image segmentation.

4.1 Pitman-Yor Processes

The Pitman-Yor process is a two-parameter generalization of the DP, with a discount parameter $0 \leq d < 1$ and a concentration parameter $\alpha > -d$. When $d = 0$ the Pitman-Yor process reduces to a DP with concentration parameter α. We write $G \sim \mathrm{PY}(d, \alpha, H)$ if G is a Pitman-Yor process with the given parameters and base measure H. The stick-breaking construction and the Chinese restaurant process have natural generalizations in the Pitman-Yor process. A draw G

from the Pitman-Yor process has the following stick-breaking construction:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^*}, \tag{28}$$

where the atoms θ_k^* are drawn iid from H, and the weights are obtained as follows:

$$\begin{aligned} v_k \,|\, d, \alpha &\sim \text{Beta}(1-d, \alpha + kd) \qquad\qquad \text{for } k = 1, \ldots, \infty \\ \beta_k &= v_k \prod_{l=1}^{k-1} (1 - v_l). \end{aligned} \tag{29}$$

We refer to the joint distribution over $\beta_1, \beta_2, \ldots$ as the GEM(d, α) distribution, this being a two-parameter generalization of the one-parameter GEM(α) associated with the DP. Suppose that H is a smooth distribution and let $\theta_1, \theta_2, \ldots$ be iid draws from G. Marginalizing out G, the distribution of θ_i conditioned on $\theta_1, \ldots, \theta_{i-1}$ follows a generalization of the Pólya urn scheme:

$$\theta_i \,|\, \theta_1, \ldots, \theta_{i-1}, d, \alpha, H \sim \sum_{t=1}^{K} \frac{n_t - d}{\alpha + i - 1} \delta_{\theta_t^*} + \frac{\alpha + Kd}{\alpha + i - 1} H, \tag{30}$$

where θ_t^* is the tth unique value among $\theta_1, \ldots, \theta_{i-1}$, there being n_t occurrences of θ_t^*, and K such unique values. In the Chinese restaurant analogy, each θ_i is a customer, θ_t^* corresponds to a table, and customer i sits at table t if $\theta_i = \theta_t^*$. There are two salient properties of this generalized Chinese restaurant process. First, the rich-gets-richer property of the original Chinese restaurant process is preserved, which means that there are a small number of large tables. Second, there are a large number of small tables since the probability of occupying new tables grows along with the number of occupied tables, and the discount d decreases the probabilities of new customers sitting at small tables.

When $0 < d < 1$ the Pitman-Yor process yields power-law behavior (Pitman, 2002; Goldwater et al., 2006a; Teh, 2006a, see also Chapter **??**). It is this power-law behavior which makes the Pitman-Yor process more suitable than the DP for many applications involving natural phenomena. The power-law nature of the Pitman-Yor process can be expressed in several ways. First, under Eq. (29) we have $\mathsf{E}[\beta_k] \in \mathsf{O}(k^{-1/d})$ if $0 < d < 1$, which indicates that cluster sizes decay according to a power law. Second, Zipf's Law can be derived from the Chinese restaurant process; that is, the proportion of tables with n customers scales as $\mathsf{O}(n^{-1-d})$. Finally the Chinese restaurant process also yields Heaps' Law, where the total number of tables in a restaurant with n customers scales as $\mathsf{O}(n^d)$. Note that the discount parameter d is the key parameter governing the power-law behavior. These various power laws are illustrated in Figure 4.

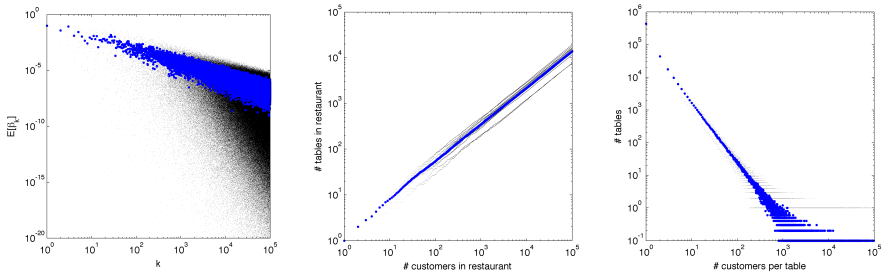

Figure 4: Power-law behavior of the Pitman-Yor process. Left: $\mathsf{E}[\beta_k]$ vs. k. Middle: number of tables in restaurant vs. number of customers. Right: number of tables vs. number of customers at each table. Each plot shows the results of 10 draws (small dots) and their mean (large dots). The log-log plots are well approximated by straight lines, indicating power laws.

4.2 Hierarchical Pitman-Yor Processes

The hierarchical Pitman-Yor (HPY) process is defined in the obvious manner:

$$\begin{aligned} G_0 \,|\, \eta, \gamma, H &\sim \mathrm{PY}(\eta, \gamma, H) \\ G_j \,|\, d, \alpha, G_0 &\sim \mathrm{PY}(d, \alpha, G_0) \qquad \text{for } j \in \mathcal{J}, \end{aligned} \tag{31}$$

where G_0 is the common base measure shared across the different Pitman-Yor processes G_j, and is itself given a Pitman-Yor process prior. Similarly to the HDP, this hierarchical construction generalizes immediately to a multiple-level hierarchy.

Recall that one of the useful facts about the HDP is that it can be represented using both a stick-breaking representation and a Chinese restaurant franchise representation. It would be of interest to consider generalizations of these objects to the HPY process. As we shall see in the following, the Chinese restaurant franchise can be readily generalized to an HPY analog. Unfortunately, however, there is no known analytic form for the stick-breaking representation of the HPY process.

Recall that in the Chinese restaurant franchise representation, each G_j corresponds to a restaurant, draws $\theta_{ji} \sim G_j$ correspond to customers, tables t in restaurant j corresponds to draws $\theta^*_{jt} \sim G_0$, and dishes correspond to draws $\theta^{**}_k \sim H$. Let n_{jtk} be the number of customers in restaurant j seated at table t and eating dish k, m_{jk} be the number of tables in restaurant j serving dish k, and K be the number of dishes served throughout the franchise. The conditional distributions given by the Chinese restaurant franchise for the HPY process are

as follows:

$$\theta_{ji} \,|\, \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha, d, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot} - d}{\alpha + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}} + \frac{\alpha + m_{j\cdot} d}{\alpha + n_{j\cdot\cdot}} G_0 \quad (32)$$

$$\theta^*_{jt} \,|\, \theta^*_{11}, \ldots, \theta^*_{1m_{1\cdot}}, \ldots, \theta^*_{j,t-1}, \gamma, \eta, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k} - \eta}{\gamma + m_{\cdot\cdot}} \delta_{\theta^{**}_k} + \frac{\gamma + K\eta}{\gamma + m_{\cdot\cdot}} H, \quad (33)$$

which is a natural generalization of the CRF for the HDP (cf. Eq. (7) and Eq. (8)).

4.3 Applications of the Hierarchical Pitman-Yor Process

In this section we describe an application of the HPY process to language modeling and another application to image segmentation.

4.3.1 Language Modeling

Statistical models of sentences in a natural language (e.g. English) are an indispensable component of many systems for processing linguistic data, including speech recognition, handwriting recognition and machine translation systems (Manning and Schütze, 1999). In this section we describe an application of the hierarchical Pitman-Yor process in statistical language modeling.

Most statistical language models treat sentences as drawn from Markov models of fixed order larger than one. That is, the probability of a sentence consisting of a sequence of words $(\theta_1, \theta_2, \ldots, \theta_\tau)$ is modeled as

$$p(\theta_1, \ldots, \theta_\tau) = \prod_{t=1}^{\tau} p(\theta_t \,|\, \theta_{t-n+1}, \ldots, \theta_{t-1}), \quad (34)$$

where for simplicity $\theta_{-n+2}, \ldots, \theta_0$ are special "start-of-sentence" symbols, and $n \geq 2$ is one plus the order of the Markov model. Such models are known as *n-gram models*. In typical applications $n = 3$, corresponding to a second-order Markov model and a context consisting of just the previous two words.

In natural languages the size of the vocabulary typically consists of more than 10^4 words. This means that in a 3-gram model the number of parameters is in excess of 10^{12}, making maximum likelihood estimation infeasible. In fact a naïve prior treating parameters corresponding to different contexts independently performs badly as well—it is important to model dependencies across different contexts for a language model to be successful. In the language modeling community such dependencies are achieved by a variety of heuristic *smoothing algorithms*, which combine the counts associated with different contexts in various ways (Chen and Goodman, 1999).

It is also possible to take a hierarchical Bayesian point of view on smoothing, and indeed such an approach was considered in a parametric setting by MacKay and Peto (1994). However, word occurrences in natural languages tend to follow

power laws, and a nonparametric model such as the HPY process provides a more natural prior for this domain (Teh, 2006a,b; Goldwater et al., 2006a). Indeed, the most successful heuristic smoothing methods are closely related to an HPY model.

Given a context $\mathbf{u}$ consisting of a sequence of words, let $G_{\mathbf{u}}$ be the distribution over the next word following the context $\mathbf{u}$. That is, $G_{\mathbf{u}}(\theta) = p(\theta_t = \theta \,|\, \theta_{t-n+1}, \ldots, \theta_{t-1} = \mathbf{u})$ in Eq. (34). We place a Pitman-Yor prior on $G_{\mathbf{u}}$, with base measure $G_{\mathrm{pa}(\mathbf{u})}$, where $\mathrm{pa}(\mathbf{u})$ is the context with the first word dropped from $\mathbf{u}$:

$$G_{\mathbf{u}} \,|\, d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{\mathrm{pa}(\mathbf{u})} \sim \mathrm{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{\mathrm{pa}(\mathbf{u})}). \tag{35}$$

The parameters of the Pitman-Yor process depend on the length of the context $|\mathbf{u}|$. We recursively place a Pitman-Yor prior on $G_{\mathrm{pa}(\mathbf{u})}$, dropping words from the front of the context until $G_\emptyset$, the distribution over next words given the empty context $\emptyset$. Finally we place a Pitman-Yor prior on $G_\emptyset$:

$$G_\emptyset \,|\, d_0, \alpha_0, G_0 \sim \mathrm{PY}(d_0, \alpha_0, G_0), \tag{36}$$

where G_0 is the uniform distribution over the vocabulary. The structure of this hierarchical prior reflects the notion that more recent words in the context are more informative in predicting the next word.

Teh (2006a,b) applied the HPY language model to a 14-million word corpus, and found that it produces state-of-the-art prediction results, closely matching results using interpolated and modified Kneser-Ney, two of the most widely-used smoothing algorithms (Chen and Goodman, 1998). Moreover, the HPY language model has been shown to outperform modified Kneser-Ney in the context of an application to dialog transcription (Huang and Renals, 2007). These results are unsurprising, as Teh (2006a,b) and Goldwater et al. (2006a) showed that interpolated Kneser-Ney can be derived as an approximation to the CRF representation of the HPY language model. In particular, interpolated Kneser-Ney assumes that the number of tables in each restaurant serving each dish is at most one. This is the same approximation as in Section 2.4.1.

4.3.2 Image Segmentation

Models based on the Pitman-Yor process have also had impact in the field of image processing, a field that shares with the language modeling domain the fact that power laws characterize many of the statistics within the domain. In particular, using a database of images that were manually segmented and labeled by humans (Oliva and Torralba, 2001), Sudderth and Jordan (2009) have shown that both the segment sizes and the label occurrences (e.g., "sky," "grass") follow long-tailed distributions that are well captured by the Pitman-Yor process. This suggests considering models in which the marginal distributions at each site in an image are governed by Pitman-Yor processes. Moreover, to share information across a collection of images it is natural to consider HPY priors. In this section we describe a model based on such an HPY prior (Sudderth

and Jordan, 2009). Our focus is the problem of image segmentation, where the observed data are a collection of images (an image is a collection of gray-scale or color values at each point in a two-dimensional grid) and the problem is to output a partition of each image into segments (a segment is a coherent region of the image, as defined by human labelings).

Let us consider a generative model for image texture and color, simplifying at first in two ways: (1) we focus on a single image and (2) we neglect the issue of spatial dependency within the image. Thus, for now we focus simply on obtaining Pitman-Yor marginal statistics for segment sizes and segment labels within a single image. Let us suppose that the image is represented as a large collection of *sites*, where a site is a local region in the image (often referred to as a *pixel* or a *super-pixel*). Let $\boldsymbol{\pi} \sim \text{GEM}(d, \alpha)$ be a draw from the two-parameter GEM distribution. For each site i, let t_i denote the segment assignment of site i, where $t_i \sim \text{Discrete}(\boldsymbol{\pi})$ are independent draws from $\boldsymbol{\pi}$. Given a large number of sites of equal size, the total area assigned to segment t will be roughly π_t, and segment sizes will follow Pitman-Yor statistics.

We also assign a label to each segment, again using a two-parameter GEM distribution. In particular, let $\boldsymbol{\beta} \sim \text{GEM}(\eta, \gamma)$ be a distribution across labels. For each segment t we label the segment by drawing $k_t \sim \text{Discrete}(\boldsymbol{\beta})$ independently. We also let θ_k^{**} denote an "appearance model"[2] for label type k, where the θ_k^{**} are drawn from some prior distribution H. Putting this together, the label assigned to site i is denoted k_{t_i}. The visual texture and color at site i are then generated by a draw from the distribution $\theta_{k_{t_i}}^{**}$.

To obtain a spatially dependent Pitman-Yor process, Sudderth and Jordan (2009) adapt an idea of Duan et al. (2007), who used a latent collection of Gaussian processes to define a spatially dependent set of draws from a Dirichlet process. In particular, to each index t we associate a zero-mean Gaussian process, u_t. At a given site i, we thus have an infinite collection of Gaussian random variables, $\{u_{ti}\}_{t=1,\ldots,\infty}$. By an appropriate choice of thresholds for this infinite sequence of Gaussian variables, it is possible to mimic a draw from the distribution $\boldsymbol{\pi}$ (by basing the selection on the first Gaussian variable in the sequence that is less than its threshold). Indeed, for a single site, this is simply a change-of-variables problem from a collection of beta random variables to a collection of Gaussian random variables. The Gaussian process framework couples the choice of segments at nearby sites via the covariance function. Figure 5 gives an example of three draws from this model, showing the underlying random distribution $\boldsymbol{\pi}$ (truncated to four values), the corresponding collection of draws from Gaussian processes (again truncated), and the resulting segmented image.

This framework applies readily to multiple images by coupling the label distribution $\boldsymbol{\beta}$ and appearance models θ_k^{**} across multiple images. Letting $j \in \mathcal{J}$ index the images in the collection, we associate a segment distribution $\boldsymbol{\pi}_j$ with each image and associate a set of Gaussian processes with each image to describe

[2]This parameter is generally a multinomial parameter encoding the probabilities of various discrete-valued texture and color descriptors.

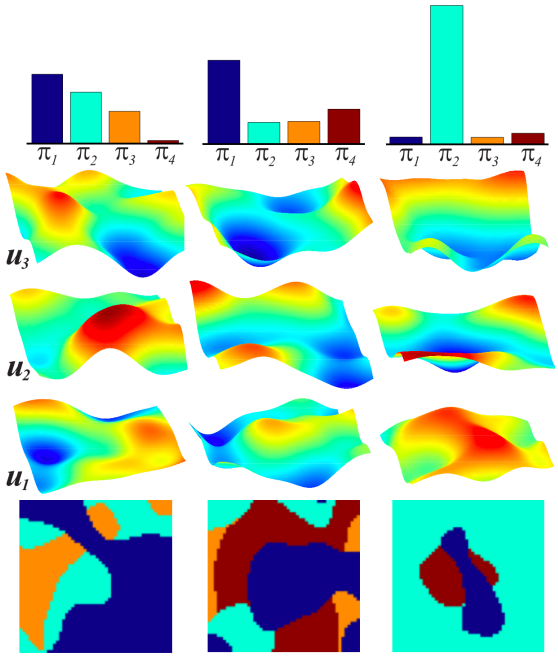

Figure 5: Draws from dependent Pitman-Yor processes. Top: the random proportions $\boldsymbol{\pi}_j$. Middle: draws from Gaussian processes, one for each entry in $\boldsymbol{\pi}_j$. Bottom: resulting segmentation.

the segmentation of that image.

The image segmentation problem can be cast as posterior inference in this HPY-based model. Given an image represented as a collection of texture and color descriptors, we compute the maximum a posteriori set of segments for the sites. Sudderth and Jordan (2009) have shown that this procedure yields a state-of-the-art unsupervised image segmentation algorithm.

5 The Beta Process and the Indian Buffet Process

The DP mixture model embodies the assumption that the data can be partitioned or clustered into discrete classes. This assumption is made particularly clear in the Chinese restaurant representation, where the table at which a data point sits indexes the class (the mixture component) to which it is assigned. If we represent the restaurant as a binary matrix in which the rows are the data points and the columns are the tables, we obtain a matrix with a single one in each row and all other elements equal to zero.

A different assumption that is natural in many settings is that objects can be

described in terms of a collection of binary *features* or *attributes*. For example, we might describe a set of animals with features such as `diurnal/nocturnal`, `avian/non-avian`, `cold-blooded/warm-blooded`, etc. Forming a binary matrix in which the rows are the objects and the columns are the features, we obtain a matrix in which there are multiple ones in each row. We will refer to such a representation as a *featural representation*.

A featural representation can of course be converted into a set of clusters if desired: if there are K binary features, we can place each object into one of 2^K clusters. In so doing, however, we lose the ability to distinguish between classes that have many features in common and classes that have no features in common. Also, if K is large, it may be infeasible to consider models with 2^K parameters. Using the featural representation, we might hope to construct models that use on the order of K parameters to describe 2^K classes.

In this section we discuss a Bayesian nonparametric approach to featural representations. In essence, we replace the Dirichlet/multinomial probabilities that underlie the Dirichlet process with a collection of beta/Bernoulli draws. This is achieved via the *beta process*, a stochastic process whose realizations provide a countably infinite collection of coin-tossing probabilities. We also discuss some other representations of the beta process that parallel those for the DP. In particular we describe a stick-breaking construction as well as an analog of the Chinese restaurant process known as the *Indian buffet process*.

5.1 The Beta Process and the Bernoulli Process

The beta process is an instance of a general class of stochastic processes known as *completely random measures* (Kingman, 1967, see also Chapter **??**). The key property of completely random measures is that the random variables obtained by evaluating a random measure on disjoint subsets of the probability space are mutually independent. Moreover, draws from a completely random measure are discrete (up to a fixed deterministic component). Thus we can represent such a draw as a weighted collection of atoms on some probability space, as we do for the DP. (Note, however, that the DP is not a completely random measure because the weights are constrained to sum to one for the DP; thus, the independence assertion does not hold for the DP. The DP can be obtained by *normalizing* a completely random measure (specifically the gamma process; see **??**).

Applications of the beta process in Bayesian nonparametric statistics have mainly focused on its use as a model for random hazard functions (Hjort, 1990, see also Chapter **??**). In this case, the probability space is the real line and it is the cumulative integral of the sample paths that is of interest (yielding a random, nondecreasing step function). In the application of the beta process to featural representations, on the other hand, it is the realization itself that is of interest and the underlying space is no longer restricted to be the real line.

Following Thibaux and Jordan (2007), let us thus consider a general probability space (Θ, Ω) endowed with a finite *base measure* B_0 (note that B_0 is not a probability measure; it does not necessarily integrate to one). Intuitively we

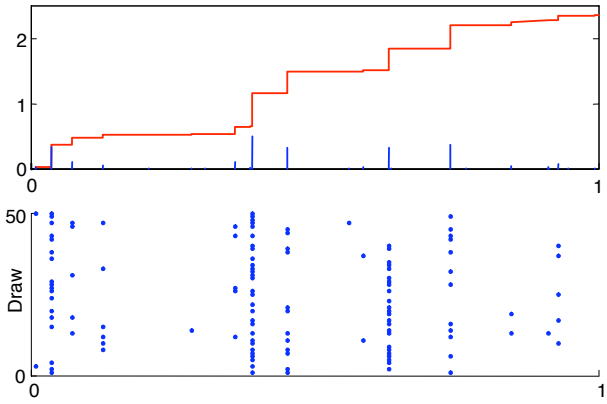

Figure 6: (a) A draw $B \sim \mathrm{BP}(1, U[0,1])$. The set of blue spikes is the sample path and the red curve is the corresponding cumulative integral $\int_{-\infty}^{x} B(d\theta)$. (b) 100 samples from $\mathrm{BeP}(B)$, one sample per row. Note that a single sample is a *set* of unit-weight atoms.

wish to partition Θ into small regions, placing atoms into these regions according to B_0 and assigning a weight to each atom, where the weight is a draw from a beta distribution. A similar partitioning occurs in the definition of the DP, but in that case the aggregation property of Dirichlet random variables immediately yields a consistent set of marginals and thus an easy appeal to Kolmogorov's theorem. Because the sum of two beta random variables is not a beta random variable, the construction is somewhat less straightforward in the beta process case.

The general machinery of completely random processes deals with this issue in an elegant way. Consider first the case in which B_0 is absolutely continuous and define the *Lévy measure* on the product space $[0,1] \otimes \Theta$ in the following way:

$$\nu(d\omega, d\theta) = c\omega^{-1}(1-\omega)^{c-1} d\omega B_0(d\theta), \tag{37}$$

where $c > 0$ is a *concentration parameter*. Now sample from a nonhomogeneous Poisson process with the Lévy measure ν as its rate measure. This yields a set of atoms at locations $(\omega_1, \theta_1), (\omega_2, \theta_2) \ldots$. Define a realization of the beta process as:

$$B = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \tag{38}$$

where δ_{θ_k} is an atom at θ_k with ω_k its mass in B. We denote this stochastic process as $B \sim \mathrm{BP}(c, B_0)$. Figure 6(a) provides an example of a draw from $\mathrm{BP}(1, U[0,1])$, where $U[0,1]$ is the uniform distribution on $[0,1]$.

We obtain a countably infinite set of atoms from this construction because the Lévy measure in Eq. (37) is σ-finite with infinite mass. Indeed, consider

partitioning the product space $[0,1] \otimes \Theta$ into stripes having equal integral under this density. These stripes have the same finite rate under the Poisson process, and there are an infinite number of such stripes. Note also that the use of a limiting form of the beta density implies that most of the atoms are associated with very small weights. Campbell's Theorem shows that the sum of these weights is finite with probability one, since $\int \omega \nu(d\omega, d\theta) < \infty$.

If B_0 contains atoms, then these are treated separately. In particular, denote the measure of the kth atom as q_k (assumed to lie in $(0,1)$). The realization B necessarily contains that atom, with the corresponding weight ω_k defined as an independent draw from $\text{Beta}(cq_k, c(1-q_k))$. The overall realization B is a sum of the weighted atoms coming from the continuous component and the discrete component of B_0.

Let us now define a *Bernoulli process* $\text{BeP}(B)$ with an atomic base measure B as a stochastic process whose realizations are collections of atoms of unit mass on Θ. Atoms can only appear at the locations of atoms of B. Whether or not an atom appears is determined by independent tosses of a coin, where the probability of success is the corresponding weight of the atom in B. After n draws from $\text{BeP}(B)$ we can fill a binary matrix that has n rows and an infinite number of columns (corresponding to the atoms of B arranged in some order). Most of the entries of the matrix are zero while a small (finite) number of the entries are equal to one. Figure 6(b) provides an example.

The beta process and the Bernoulli process are *conjugate*. Consider the specification:

$$\begin{aligned} B \,|\, c, B_0 &\sim \text{BP}(c, B_0) \\ Z_i \,|\, B &\sim \text{BeP}(B), \qquad\qquad \text{for } i = 1, \ldots, n, \end{aligned} \tag{39}$$

where $Z_1, \ldots, Z_n$ are conditionally independent given B. The resulting posterior distribution is itself a beta process, with updated parameters:

$$B \,|\, Z_1, \ldots, Z_n, c, B_0 \sim \text{BP}\left(c+n, \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^{n} Z_i\right). \tag{40}$$

This formula can be viewed as an analog of standard finite-dimensional beta/Bernoulli updating. Indeed, given a prior $\text{Beta}(a,b)$, the standard update takes the form $a \to a + \sum_i z_i$ and $b \to b + n - \sum_i z_i$. In Eq. (40), c plays the role of $a+b$ and cB_0 is analogous to a.

5.2 The Indian Buffet Process

Recall that the Chinese restaurant process can be obtained by integrating out the Dirichlet process and considering the resulting distribution over partitions. In the other direction, the Dirichlet process is the random measure that is guaranteed (by exchangeability and De Finetti's theorem) to underlie the Chinese restaurant process. In this section we discuss the analog of these relationships for the beta process.

We begin by defining a stochastic process known as the *Indian buffet process* (IBP). The IBP was originally defined directly as a distribution on (equivalence classes of) binary matrices by Griffiths and Ghahramani (2006) and Ghahramani et al. (2007). The IBP is an infinitely exchangeable distribution on these equivalence classes, thus it is of interest to discover the random measure that must underlie the IBP according to De Finetti's Theorem. Thibaux and Jordan (2007) showed that the underlying measure is the beta process; that is, the IBP is obtained by integrating over the beta process B in the hierarchy in Eq. (39).

The IBP is defined as follows. Consider an Indian buffet with a countably-infinite number of dishes and customers that arrive in sequence in the buffet line. Let Z^* denote a binary-valued matrix in which the rows are customers and the columns are the dishes, and where $Z^*_{nk} = 1$ if customer n samples dish k. The first customer samples Poisson(α) dishes, where $\alpha = B_0(\Theta)$ is the total mass of B_0. A subsequent customer n samples dish k with probability $\frac{m_k}{c+n-1}$, where m_k is the number of customers who have previously sampled dish k; that is, $Z^*_{nk} \sim \text{Bernoulli}(\frac{m_k}{c+n-1})$. Having sampled from the dishes previously sampled by other customers, customer n then goes on to sample an additional number of new dishes determined by a draw from a Poisson($\frac{c}{c+n-1}\alpha$) distribution.

To derive the IBP from the beta process, consider first the distribution Eq. (40) for $n = 0$; in this case the base measure is simply B_0. Drawing from $B \sim \text{BP}(B_0)$ and then drawing $Z_1 \sim \text{BeP}(B)$ yields atoms whose locations are distributed according to a Poisson process with rate B_0; the number of such atoms is Poisson(α). Now consider the posterior distribution after $Z_1, \ldots, Z_{n-1}$ have been observed. The updated base measure is $\frac{c}{c+n-1}B_0 + \frac{1}{c+n-1}\sum_{i=1}^{n-1} Z_i$. Treat the discrete component and the continuous component separately. The discrete component, $\frac{1}{c+n-1}\sum_{i=1}^{n-1} Z_i$, can be reorganized as a sum over the unique values of the atoms; let m_k denote the number of times the kth atom appears in one of the previous Z_i. We thus obtain draws $\omega_k \sim \text{Beta}((c+n-1)q_k, (c+n-1)(1-q_k))$, where $q_k = \frac{m_k}{c+n-1}$. The expected value of ω_k is $\frac{m_k}{c+n-1}$ and thus (under Bernoulli sampling) this atom appears in Z_n with probability $\frac{m_k}{c+n-1}$. From the continuous component, $\frac{c}{c+n-1}B_0$, we generate Poisson($\frac{c}{c+n-1}\alpha$) new atoms. Equating "atoms" with "dishes," and rows of Z^* with draws Z_n, we have obtained exactly the probabilistic specification of the IBP.

5.3 Stick-Breaking Constructions

The stick-breaking representation of the DP is an elegant constructive characterization of the DP as a discrete random measure (Chapter **??**). This construction can be viewed in terms of a metaphor of breaking off lengths of a stick, and it can also be interpreted in terms of a size-biased ordering of the atoms. In this section, we consider analogous representations for the beta process. Draws $B \sim \text{BP}(c, B_0)$ from the beta process are discrete with probability one, which gives hope that such representations exist. Indeed, we will show that there are two stick-breaking constructions of B, one based on a size-biased ordering of the atoms (Thibaux and Jordan, 2007), and one based on a stick-breaking representation known as the inverse Lévy measure (Wolpert and Ickstadt, 1998).

The size-biased ordering of Thibaux and Jordan (2007) follows straightforwardly from the discussion in Section 5.2. Recall that the Indian buffet process is defined via a sequence of draws from Bernoulli processes. For each draw, a Poisson number of new atoms are generated, and the corresponding weights in the base measure B have a beta distribution. This yields the following truncated representation:

$$B_N = \sum_{n=1}^{N} \sum_{k=1}^{K_n} \omega_{nk} \delta_{\theta_{nk}}, \tag{41}$$

where

$$\begin{aligned} K_n \,|\, c, B_0 &\sim \text{Poisson}(\tfrac{c}{c+n-1}\alpha) && \text{for } n = 1, \ldots, \infty \\ \omega_{nk} \,|\, c &\sim \text{Beta}(1, c+n-1) && \text{for } k = 1, \ldots, K_n \\ \theta_{nk} \,|\, B_0 &\sim B_0/\alpha. \end{aligned} \tag{42}$$

It can be shown that this size-biased construction B_N converges to B with probability one. The expected total weight contributed at step N is $\frac{c\alpha}{(c+N)(c+N-1)}$, while the expected total weight remaining, in $B - B_N$, is $\frac{c\alpha}{c+N}$. The expected total weight remaining decreases to zero as $N \to \infty$, but at a relatively slow rate. Note also that we are not guaranteed that atoms contributed at later stages of the construction will have small weight—the sizes of the weights need not be in decreasing order.

The stick-breaking construction of Teh et al. (2007) can be derived from the inverse Lévy measure algorithm of Wolpert and Ickstadt (1998). This algorithm starts from the Lévy measure of the beta process, and generates a sequence of weights of decreasing size using a nonlinear transformation of a one-dimensional Poisson process to one with uniform rate. In general this approach does not lead to closed forms for the weights; inverses of the incomplete Beta function need to be computed numerically. However for the one-parameter beta process (where $c = 1$) we do obtain a simple closed form:

$$B_K = \sum_{k=1}^{K} \omega_k \delta_{\theta_k}, \tag{43}$$

where

$$\begin{aligned} v_k \,|\, \alpha &\sim \text{Beta}(1, \alpha) && \text{for } k = 1, \ldots, \infty \\ \omega_k &= \prod_{l=1}^{k} (1 - v_l) \\ \theta_k \,|\, B_0 &\sim B_0/\alpha. \end{aligned} \tag{44}$$

Again $B_K \to B$ as $K \to \infty$, but in this case the expected weights decrease exponentially to zero. Further, the weights are generated in strictly decreasing order, so we are guaranteed to generate the larger weights first.

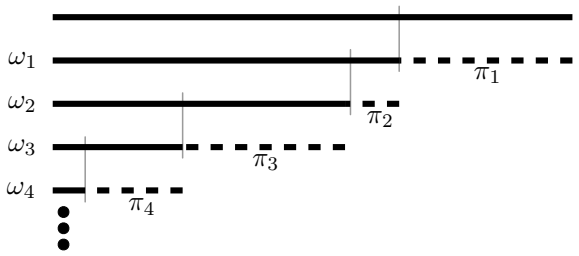

Figure 7: Stick-breaking construction for the DP and the one-parameter BP. The lengths π_i are the weights for the DP and the lengths ω_i are the weights for the BP.

The stick-breaking construction for the one-parameter beta process has an intriguing connection to the stick-breaking construction for the DP. In particular, both constructions use the same beta-distributed breakpoints v_k; the difference is that for the DP we use the lengths of the sticks just broken off as the weights while for the beta process we use the remaining lengths of the sticks. This is depicted graphically in Figure 7.

5.4 Hierarchical Beta Processes

Recall the construction of the hierarchical Dirichlet process: a set of Dirichlet processes are coupled via a random base measure. A similar construction can be carried out in the case of the beta process: let the common base measure for a set of beta processes be drawn from an underlying beta process (Thibaux and Jordan, 2007). Under this hierarchical Bayesian nonparametric model, the featural representations that are chosen for one group will be related to the featural representations that are used for other groups.

We accordingly define a *hierarchical beta process* (HBP) as follows:

$$\begin{aligned} B_0 \mid \kappa, A &\sim \mathrm{BP}(\kappa, A) && \\ B_j \mid c, B_0 &\sim \mathrm{BP}(c, B_0) && \text{for } j \in \mathcal{J} \\ Z_{ji} \mid B_j &\sim \mathrm{BeP}(B_j) && \text{for } i = 1, \ldots, n_j, \end{aligned} \tag{45}$$

where $\mathcal{J}$ is the set of groups and there are n_j individuals in group j. The hyperparameter c controls the degree of coupling among the groups: larger values of c yield realizations B_j that are closer to B_0 and thus a greater degree of overlap among the atoms chosen in the different groups.

As an example of the application of the HBP, Thibaux and Jordan (2007) considered the problem of document classification, where there are $|\mathcal{J}|$ groups of documents and where the goal is to classify a new document into one of these groups. In this case, Z_{ji} is a binary vector that represents the presence or absence in the ith document of each of the words in the vocabulary Θ. The HBP yields a form of regularization in which the group-specific word probabilities

are shrunk towards each other. This can be compared to standard Laplace smoothing, in which word probabilities are shrunk towards a fixed reference point. Such a reference point can be difficult to calibrate when there are rare words in a corpus, and Thibaux and Jordan (2007) showed empirically that the HBP yielded better predictive performance than Laplace smoothing.

5.5 Applications of the Beta Process

In the following sections we describe a number of applications of the beta process to hierarchical Bayesian featural models. Note that this is a rather different class of applications than the traditional class of applications of the beta process to random hazard functions.

5.5.1 Sparse Latent Variable Models

Latent variable models play an essential role in many forms of statistical analysis. Many latent variable models take the form of a regression on a latent vector; examples include principal component analysis, factor analysis and independent components analysis. Paralleling the interest in the regression literature in sparse regression models, one can also consider sparse latent variable models, where each observable is a function of a relatively small number of latent variables. The beta process provides a natural way of constructing such models. Indeed, under the beta process we can work with models that define a countably-infinite number of latent variables, with a small, finite number of variables being *active* (i.e., non-zero) in any realization.

Consider a set of n observed data vectors, $x_1, \ldots, x_n$. We use a beta process to model a set of latent features, $Z_1, \ldots, Z_n$, where we capture interactions among the components of these vectors as follows:

$$\begin{aligned} B \,|\, c, B_0 &\sim \mathrm{BP}(c, B_0) \\ Z_i \,|\, B &\sim \mathrm{BeP}(B) \qquad\qquad \text{for } i = 1, \ldots, n. \end{aligned} \tag{46}$$

As we have seen, realizations of beta and Bernoulli processes can be expressed as weighted sums of atoms:

$$\begin{aligned} B &= \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k} \\ Z_i &= \sum_{k=1}^{\infty} Z_{ik}^* \delta_{\theta_k}. \end{aligned} \tag{47}$$

We view θ_k as parametrizing feature k, while Z_i denotes the features that are active for item i. In particular, $Z_{ik}^* = 1$ if feature k is active for item i. The data point x_i is modeled as follows:

$$\begin{aligned} y_{ik} \,|\, H &\sim H \qquad\qquad \text{for } k = 1, \ldots, \infty \\ x_i \,|\, Z_i, \boldsymbol{\theta}, \mathbf{y}_i &\sim F_{\{\theta_k, y_{ik}\}_{k: Z_{ik}^* = 1}}, \end{aligned} \tag{48}$$

where y_{ik} is the value of feature k if it is active for item i, and the distribution $F_{\{\theta_k, y_{ik}\}_{k:Z^*_{ik}=1}}$ depends only on the active features, their values, and their parameters.

Note that this approach defines a latent variable model with an infinite number of sparse latent variables, but for each data item only a finite number of latent variables are active. The approach would often be used in a predictive setting in which the latent variables are integrated out, but if the sparseness pattern is of interest per se, it is also possible to compute a posterior distribution over the latent variables.

There are several specific examples of this sparse latent variable model in the literature. One example is an independent components analysis model with an infinite number of sparse latent components (Knowles and Ghahramani, 2007; Teh et al., 2007), where the latent variables are real-valued and x_i is a noisy observation of the linear combination $\sum_k Z^*_{ik} y_{ik} \theta_k$. Another example is the "noisy-or" model of Wood et al. (2006), where the latent variables are binary and are interpreted as presence or absence of diseases, while the observations x_i are binary vectors indicating presence or absence of symptoms.

5.5.2 Relational Models

The beta process has also been applied to the modeling of relational data (also known as dyadic data). In the relational setting, data are relations among pairs of objects (Getoor and Taskar, 2007); examples include similarity judgments between two objects, protein-protein interactions, user choices among a set of options, and ratings of products by customers.

We first consider the case in which there is a single set of objects and relations are defined among pairs of objects in that set. Formally, define an observation as a relation x_{ij} between objects i and j in a collection of n objects. Each object is modeled using a set of latent features as in Eq. (46) and Eq. (47). The observed relation x_{ij} between objects i and j then has a conditional distribution that is dependent only on the features active in objects i and j. For example, Navarro and Griffiths (2007) modeled subjective similarity judgments between objects i and j as normally distributed with mean $\sum_{k=1}^{\infty} \theta_k Z^*_{ik} Z^*_{jk}$; note that this is a weighted sum of features active in both objects. Chu et al. (2006) modeled high-throughput protein-protein interaction screens where the observed binding affinity of proteins i and j is related to the number of overlapping features $\sum_{k=1}^{\infty} Z^*_{ik} Z^*_{jk}$, with each feature interpreted as a potential protein complex consisting of proteins containing the feature. Görür et al. (2006) proposed a nonparametric *elimination by aspects* choice model where the probability of a user choosing object i over object j is modeled as proportional to a weighted sum, $\sum_{k=1}^{\infty} \theta_k Z^*_{ik}(1 - Z^*_{jk})$, across features active for object i that are not active for object j. Note that in these examples, the parameters of the model, θ_k, are the atoms of the beta process.

Relational data involving separate collections of objects can be modeled with the beta process as well. Meeds et al. (2007) modeled movie ratings, where the collections of objects are movies and users, and the relational data consists of

ratings of movies by users. The task is to predict the ratings of movies not yet rated by users, using these predictions to recommend new movies to users. These tasks are called *recommender systems* or *collaborative filtering*. Meeds et al. (2007) proposed a featural model where movies and users are modeled using separate IBPs. Let Z^* be the binary matrix of movie features and Y^* the matrix of user features. The rating of movie i by user j is modeled as normally distributed with mean $\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\theta_{kl}Z^*_{ik}Y^*_{jl}$. Note that this dyadic model cannot be represented using two independent beta processes, since there is a parameter θ_{kl} for each combination of features in the two IBPs. The question of what random measure underlies this model is an interesting one.

6 Semiparametric Models

The nonparametric priors introduced in previous sections can be combined with more traditional finite-dimensional priors, as well as hierarchies of such priors. In the resulting *semiparametric* models, the object of inference may be the finite-dimensional parameter, with the nonparametric component treated as a nuisance parameter to be integrated out. In other cases, the finite-dimensional parameters are to be integrated out and aspects of the nonparametric component are the inferential focus. In this section we describe two such semiparametric models based on the HDP. The first model couples the stick-breaking representation of the HDP with a Gaussian hierarchy, while the other is based on the Chinese restaurant franchise representation of the HDP.

6.1 Hierarchical DPs with Random Effects

An important characteristic of the HDP is that the same atoms appear in different DPs, allowing clusters to be shared across the different groups. The *hierarchical DP with random effects* (HDP+RE) model of Kim and Smyth (2007) generalizes the HDP by allowing atoms in different DPs to differ from each other to better capture group-specificity of cluster parameters. This model is based on the stick-breaking representation for HDPs. We begin with the standard representation for the common random base measure $G_0 \sim \mathrm{DP}(\gamma, H)$:

$$\begin{aligned} \boldsymbol{\beta} \,|\, \gamma &\sim \mathrm{GEM}(\gamma) && (49)\\ \theta^{**}_k \,|\, H &\sim H && \text{for } k = 1, \ldots, \infty \\ G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\theta^{**}_k}. \end{aligned}$$

For each group $j \in \mathcal{J}$, the weights and atoms for the group-specific G_j differ from G_0 in the following way:

$$\begin{aligned} \boldsymbol{\pi}_j \,|\, \boldsymbol{\beta} &\sim \mathrm{DP}(\alpha, \boldsymbol{\beta}) && \text{for } j \in \mathcal{J} \\ \theta^*_{jk} \,|\, \theta^{**}_k &\sim T_{\theta^{**}_k} && \text{for } k = 1, \ldots, \infty \\ G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta^*_{jk}}, \end{aligned} \tag{50}$$

where T_θ is a distribution centered at θ; for example, T_θ might be a normal distribution with mean θ.

Kim and Smyth (2007) used the HDP+RE model to model bumps in functional magnetic resonance imaging (fMRI) data. fMRI analyses report areas of high metabolic activity in the brain that are correlated with external stimuli in an attempt to discover the function of local brain regions. Such areas of activities often show up as *bumps* in fMRI images, and each bump can be modeled well using a normal density. An fMRI image then consists of multiple bumps and can be modeled with a DP mixture. Each individual brain might have slightly different structure and might react differently to the same stimuli, while each fMRI machine has different characteristics. The HDP+RE model naturally captures such variations while sharing statistical strength across individuals and machines.

6.2 Analysis of Densities and Transformed DPs

In this section we describe another approach to introducing group-specific parameters within the DP framework. The common base measure G_0 is still given a DP prior as in Eq. (49), while the group-specific random measures are defined differently:

$$\begin{aligned} H_0 &= \sum_{k=1}^{\infty} \beta_k T_{\theta^{**}_k} \\ G_j \,|\, H_0 &\sim \mathrm{DP}(\alpha, H_0) && \text{for } j \in \mathcal{J}. \end{aligned} \tag{51}$$

In the particular case in which H and T_θ are normal distributions with fixed variances, this model has been termed the *analysis of densities* (AnDe) model by Tomlinson and Escobar (2003), who used it for sharing statistical strength among multiple density estimation problems.

Sudderth et al. (2008) called the model given by Eq. (49) and Eq. (51) a *transformed DP*. The transformed DP is very similar to an HDP, the difference being that the atoms in G_0 are replaced by distributions parametrized by the atoms. If these distributions are smooth the measures G_j will not share atoms as in the HDP. Instead each atom in G_j is drawn from $T_{\theta^{**}_k}$ with probability β_k. Identifying an atom of G_j with θ^{**}_k, the Chinese restaurant franchise representation for the HDP can be generalized to the transformed DP. We have customers (draws from G_j) going into restaurants (G_j) and sitting around tables (draws

from H_0), while tables are served dishes (atoms in G_0) from a franchise-wide menu (G_0). In the HDP the actual dish served at the different tables that order the same dish are identical. For the transformed DP the dishes that are served at different tables ordering the same dish on the menu can take on distinct values.

Sudderth et al. (2008) used the transformed DP as a model for visual scene analysis that can simultaneously segment, detect and recognize objects within the scenes. Each image is first preprocessed into a set of low-level descriptors of local image appearances, and Eq. (49) and Eq. (51) are completed with a mixture model for these descriptors:

$$\begin{aligned} \theta_{ji} \,|\, G_j &\sim G_j \qquad\qquad \text{for } j \in \mathcal{J} \text{ and } i = 1, \ldots, n_j \\ x_{ji} \,|\, \theta_{ji} &\sim F_{\theta_{ji}}, \end{aligned} \tag{52}$$

where x_{ji} is one of n_j image descriptors in image j and $F_{\theta_{ji}}$ is a distribution over image descriptors parameterized by θ_{ji}.

The Chinese restaurant franchise representation of the transformed DP translates to a hierarchical representation of visual scenes, with scenes consisting of multiple objects and objects consisting of descriptors of local image appearances. To see this, note that customers (x_{ji}) are clustered into tables (object instances), and tables are served dishes from a global menu (each object instance belongs to an object category). There could be multiple tables in the same restaurant serving variations (different "seasonings") on a given dish from the global menu. This corresponds to the fact that there could be multiple instances of the same object category in a single visual scene, with each instance being in a different location or having different poses or lighting conditions (thus yielding *transformed* versions of an object category template).

7 Inference for Hierarchical Bayesian Nonparametric Models

In this section we discuss algorithmic aspects of inference for the hierarchical Bayesian nonparametric models that we have discussed in earlier sections. Our treatment will be brief and selective; in particular, we focus on relatively simple algorithms that help to convey basic methodology and provide a sense of some of the options that are available. An underlying theme of this section is that the various mathematical representations available for nonparametric models—including stick-breaking representations, urn models and truncations—can be combined in various ways to yield a wide range of possible algorithmic implementations.

While we focus on sampling-based inference methods throughout this section, we also note that there is a growing literature on variational methods for inference in hierarchical Bayesian nonparametric models; examples include Liang et al. (2007); Sudderth and Jordan (2009) and Teh et al. (2008).

7.1 Inference for Hierarchical Dirichlet Processes

We begin by considering posterior inference for a simple HDP mixture model. In this model, the random measures G_j are drawn from an HDP model according to Eq. (1), and this HDP prior is completed as follows:

$$\begin{aligned} &\theta_{ji} \,|\, G_j \sim G_j && \text{for } i = 1, \ldots, n_j \\ &x_{ji} \,|\, \theta_{ji} \sim F_{\theta_{ji}}, \end{aligned} \tag{53}$$

where the ith observation in the jth group is denoted x_{ji} and where this observation is drawn from a distribution $F_{\theta_{ji}}$ indexed by θ_{ji}. The latent parameter θ_{ji} is drawn from G_j and can be viewed as indexing the mixture component associated with the data point x_{ji}. We shall assume that H is conjugate to F_θ for simplicity. Nonconjugate models can be treated by adapting techniques from the DP mixture literature (cf. Neal, 2000).

Teh et al. (2006) presented sampling algorithms for the HDP mixture model based on both the CRF representation and the stick-breaking representation. In the following section we describe the CRF-based sampler. We then turn to an alternative sampler that is based on the posterior representation of the HDP described in Section 2.3.

7.1.1 Chinese Restaurant Franchise Sampler

Recall our notation for the CRF representation of the HDP. Customer i in restaurant j is associated with an iid draw from G_j and sits at table t_{ji}. Table t in restaurant j is associated with an iid draw from G_0 and serves a dish k_{jt} from a franchise-wide menu. Dish k is associated with an iid draw from H. If H is an absolutely continuous measure then each such dish is unique with probability one. There are n_{jtk} customers in restaurant j sitting at table t and eating dish k, and there are m_{jk} tables in restaurant j serving dish k.

Given this setup, we describe a Gibbs sampler in which the table and dish assignment variables are iteratively sampled conditioned on the state of all other variables. The variables consist of $\{t_{ji}\}_{j\in\mathcal{J},i=1,\ldots,n_j}$ and $\{k_{jt}\}_{j\in\mathcal{J},t=1,\ldots,m_{j\cdot}}$. The parameters θ_{ji} are integrated out analytically (recall our assumption of conjugacy). Consider the assignment of customer i in restaurant j to a table t_{ji}. To resample t_{ji} we make use of exchangeability and imagine customer i being the last customer to enter restaurant j. The customer can sit at an already occupied table, can sit at a new table and be served an existing dish, or can sit at a new table and be served a new dish. The probabilities of these events are:

$$\begin{cases} t_{ji} = t & \text{with probability} \propto \frac{n_{jt\cdot}^{\neg ji}}{n_{j\cdot\cdot}^{\neg ji}+\alpha} f_{k_{jt}}(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{m_{\cdot k}^{\neg ji}}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_k(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k^{\text{new}} & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{\gamma}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_{k^{\text{new}}}(\{x_{ji}\}), \end{cases} \tag{54}$$

where t^{new} and k^{new} denote a new table and new dish, respectively, and where superscript $^{\neg ji}$ denotes counts in which customer i in restaurant j is removed from the CRF (if this empties a table we also remove that table from the CRF along with the dish served on it). The fractional terms are the conditional priors given by the CRF in Eq. (7) and Eq. (8), and $f_k(\{x_{ji}\})$ is defined using the following general notation:

$$f_k(\{x_{ji}\}_{ji\in D}) = \frac{\int h(\theta) \prod_{j'i'\in D_k\cup D} f_\theta(x_{j'i'})d\theta}{\int h(\theta) \prod_{j'i'\in D_k\setminus D} f_\theta(x_{j'i'})d\theta}, \tag{55}$$

where D is an arbitrary index set, where $D_k = \{j'i' : k_{j't_{j'i'}} = k\}$ denotes the set of indices of data items currently associated with dish k, and where $h(\cdot)$ and $f_\theta(\cdot)$ denote the densities of H and F_θ respectively. In particular, $f_k(\{x_{ji}\})$ is the marginal conditional probability of the singleton data point x_{ji} in cluster k, given all of the other data points currently assigned to cluster k.

The Gibbs update for the dish k_{jt} served at table t in restaurant j is derived similarly. The probabilities of the relevant events in this case are:

$$k_{jt} = \begin{cases} k & \text{with probability} \propto \frac{m_{\cdot k}^{\neg jt}}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_k(\{x_{ji} : t_{ji} = t\}) \\ k^{\text{new}} & \text{with probability} \propto \frac{\gamma}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_{k^{\text{new}}}(\{x_{ji} : t_{ji} = t\}). \end{cases} \tag{56}$$

While the computational cost of the Gibbs updates is generally dominated by the computation of the marginal conditional probabilities $f_k(\cdot)$, the number of possible events that can occur at one Gibbs step is one plus the total number of tables or dishes in all restaurants that are ancestors of j, and this number can be large in deep or wide hierarchies.

A drawback of the CRF sampler is that it couples sampling in the various restaurants (since all DPs are integrated out). This coupling makes deriving a CRF sampler for certain models (e.g. the HDP-HMM) difficult. An alternative is to construct samplers that use a mixed representation—some DPs in stick-breaking representation and some in CRP representation—and thereby decouple the restaurants (Teh et al., 2006).

The CRF-based sampler can be easily extended to arbitrary hierarchies. It can also be extended to the hierarchical Pitman-Yor process discussed in Section 4.

7.1.2 Posterior Representation Sampler

In Section 2.3 we showed that the posterior of the HDP consists of a discrete part corresponding to mixture components associated with data and a continuous part corresponding to components not associated with data. This representation can be used to develop a sampler which represents only the discrete part explicitly. In particular, referring to Eq. (10) and Eq. (12), the posterior

representation sampler maintains only the weights $\boldsymbol{\beta}$ and $\{\boldsymbol{\pi}_j\}_{j\in\mathcal{J}}$. (The atoms $\{\theta^{**}_k\}_{k=1,\ldots,K}$ can be integrated out in the conjugate setting.) We also make use of cluster index variables z_{ji}, defined so that $\theta_{ji} = \theta^{**}_{z_{ji}}$ (i.e., $z_{ji} = k_{jt_{ji}}$ in the CRF representation).

The sampler iterates between two phases: the sampling of the cluster indices $\{z_{ji}\}$, and the sampling of the weights $\boldsymbol{\beta}$ and $\{\boldsymbol{\pi}_j\}$. The sampling of the cluster indices is a simple variation on the Gibbs updates in the CRF sampler described in Section 7.1.1. In particular, we define the following Gibbs conditionals:

$$z_{ji} = \begin{cases} k & \text{with probability} \propto \pi_{jk} f_k(\{x_{ji}\}) \\ k^{\text{new}} & \text{with probability} \propto \pi_{j0} f_{k^{\text{new}}}(\{x_{ji}\}). \end{cases} \tag{57}$$

If a new component k^{new} is chosen, the corresponding atom is instantiated in the sampler. Specifically, the weights corresponding to this new atom can be generated as follows:

$$\begin{aligned} v_0 \,|\, \gamma &\sim \text{Beta}(\gamma, 1) \\ (\beta_0^{\text{new}}, \beta_{K+1}^{\text{new}}) &= (\beta_0 v_0, \beta_0(1-v_0)) \\ v_j \,|\, \alpha, \beta_0, v_0 &\sim \text{Beta}(\alpha\beta_0 v_0, \alpha\beta_0(1-v_0)) \qquad \text{for } j \in \mathcal{J} \\ (\pi_{j0}^{\text{new}}, \pi_{j\,K+1}^{\text{new}}) &= (\pi_{j0} v_j, \pi_{j0}(1-v_j)). \end{aligned} \tag{58}$$

Finally we set $z_{ji} = K+1$ and increment K.

The second phase resamples the weights $\{\boldsymbol{\pi}_j\}_{j\in\mathcal{J}}$ and $\boldsymbol{\beta}$ conditioned on the cluster indices $\{z_{ji}\}$. The approach is to first integrate out the random measures, leaving a CRP representation as in Section 2.2, then the weights $\{\boldsymbol{\pi}_j\}_{j\in\mathcal{J}}$ and $\boldsymbol{\beta}$ can be sampled conditionally on the state of the CRF using Eq. (10) and Eq. (12). Because we are conditioning on $\{z_{ji}\}$, and customers with different values of z_{ji} cannot be assigned to the same table, each restaurant effectively gets split into independent "sub-restaurants," one for each value of k. (See also the related direct assignment sampler in Teh et al. (2006).) Let $n_{j\cdot k}$ be the number of observations in group j assigned to component k, and let m_{jk} be the random number of tables in a sub-restaurant with $n_{j\cdot k}$ customers and concentration parameter $\alpha\beta_k$. The $\{m_{jk}\}$ are mutually independent and thus a draw for each of them can be simulated using the CRP. We can now sample the $\boldsymbol{\beta}$ and $\{\boldsymbol{\pi}_j\}$ using Eq. (10) and Eq. (12).

7.2 Inference for HDP Hidden Markov Models

The posterior representation sampler of Section 7.1.2 can also be used to derive a Gibbs sampler for the HDP-HMM. Consider the formulation of the HDP-HMM given in Eq. (23) and Eq. (24) where we make use of a sequence of latent indicator variables $z_1, \ldots, z_\tau$. We again assume that H is conjugate to F_θ. Note that the posterior of the HDP prior for the model (given $z_1, \ldots, z_\tau$) can be decomposed into a discrete part consisting of K atoms (corresponding to the K states currently visited by $z_1, \ldots, z_\tau$), as well as a continuous part consisting of unused atoms. The weights on the K atoms (equivalently the transition

probabilities among the K states currently used by the HDP-HMM) can be constructed from a CRF representation of the HDP:

$$\begin{aligned}(\beta_0, \beta_1, \ldots, \beta_K) &\sim \text{Dirichlet}(\gamma, m_{\cdot 1}, \ldots, m_{\cdot K}) \\ (\pi_{j0}, \pi_{j1}, \ldots, \pi_{jK}) &\sim \text{Dirichlet}(\alpha\beta_0, \alpha\beta_1 + n_{j\cdot 1}, \ldots, \alpha\beta_K + n_{j\cdot K}) \quad \text{for } j = 1, \ldots, K,\end{aligned} \tag{59}$$

where $n_{j\cdot k}$ is the number of transitions from state j to state k (equivalently the number of customers eating dish k in restaurant j), while $m_{\cdot k}$ is the number of tables serving dish k in the CRF representation of the HDP. The conditional probabilities for the Gibbs update of z_t are as follows:

$$z_t = \begin{cases} k & \text{with probability} \propto \pi_{z_{t-1}k}\pi_{kz_{t+1}} f_k(\{x_t\}) \\ k^{\text{new}} & \text{with probability} \propto \pi_{z_{t-1}0}\beta_{z_{t+1}} f_{k^{\text{new}}}(\{x_t\}). \end{cases} \tag{60}$$

The three factors on the right-hand side are the probability of transitioning into the current state, the probability of transitioning out of the current state, and the conditional probability of the current observation x_t respectively. The $\beta_{z_{t+1}}$ factor arises because transitions from the new state k^{new} have not been observed before so we need to use the conditional prior mean $\boldsymbol{\beta}$. The weights $\boldsymbol{\beta}$ and transition probabilities $\boldsymbol{\pi}_j$ can be updated as for the posterior representation sampler for plain HDPs.

This simple Gibbs sampler can converge very slowly due to strong dependencies among the latent states (Scott, 2002). To obtain a faster algorithm we would like to update the latent states in a block via the forward-backward algorithm for HMMs; the traditional form of this algorithm cannot, however, be applied directly to the HDP-HMM since there are an infinite number of possible states. The solution is to limit the number of states to a finite number so that the forward-backward algorithm becomes feasible. Fox et al. (2009) proposed doing this via a truncation of the stick-breaking process (cf. Ishwaran and James, 2001), while Van Gael et al. (2008) proposed a slice sampling approach which adaptively limits the number of states to a finite number (Neal, 2003; Walker, 2007).

7.3 Inference for Beta Processes

In this section we describe a Gibbs sampler for the beta process latent variable model described in Section 5.5.1. This sampler is based on the stick-breaking representation of the beta process.

Recall that the model is defined in terms of a set of feature weights $\{\omega_k\}_{k=1,\ldots,\infty}$ and the atoms (feature parameters) $\{\theta_k\}_{k=1,\ldots,\infty}$. Moreover, corresponding to each data item x_i, we have a set of binary feature "activities" $\{Z^*_{ik}\}_{k=1,\ldots,\infty}$ and latent feature values $\{y_{ik}\}_{k=1,\ldots,\infty}$. The observed data item x_i depends on $\{\theta_k, y_{ik}\}_{k:Z^*_{ik}=1}$.

The conditional distributions defining the model are given in Eq. (44) and Eq. (48), where $p(Z^*_{ik} = 1 \,|\, \omega_k) = \omega_k$. Gibbs sampling in this model is straightforward except for a few difficulties which we describe below along with their resolution.

The main difficulty with a Gibbs sampler is that there are an infinite number of random variables that need to be sampled. To circumvent this problem, Teh et al. (2007) propose to use slice sampling (Neal, 2003; Walker, 2007) to adaptively truncate the representation to a finite number of features. Consider an auxiliary variable s with conditional distribution:

$$s \mid Z^*, \{\omega_k\}_{k=1,\ldots,\infty} \sim \text{Uniform}\left[0, \min_{k:\exists i, Z^*_{ik}=1} \omega_k\right], \tag{61}$$

where the supremum in the range of s is the smallest feature weight ω_k among the currently active features. Conditioned on the current state of the other variables a new value for s can easily be sampled. Conditioned on s, features for which $\omega_k < s$ are forced to be inactive since making them active would make s lie outside its range. This means that we only need to update the finite number of features for which $\omega_k > s$. This typically includes all the active features, along with a small number of inactive features (needed for the sampler to explore the use of new features).

A related issue concerns the representation of the model within the finite memory of the computer. Using the auxiliary variable s it is clear that we need only represent features $1, \ldots, K$, where K is such that $\omega_{K+1} < s$; that is, the model is truncated after feature K. As the values of s and the feature weights change over the course of Gibbs sampling this value of K changes as well. If K is decreased we simply delete the last few features, while if K is increased we sample the variables ω_k, θ_k and y_{ik} corresponding to these new features from their conditional distributions given the current state of the represented features.

The final issue is the problem of sampling the feature weights $\omega_1, \ldots, \omega_K$. Unlike the case of DPs, it is easier in this case to work with the weights directly instead of the stick-breaking variables v_k. In particular, Teh et al. (2007) showed that the joint probability for the weights is:

$$p(\omega_1, \ldots, \omega_K) = \mathbb{I}(0 \leq \omega_K \leq \cdots \leq \omega_1 \leq 1)\alpha^K \omega_K^\alpha \prod_{k=1}^{K} \omega_k^{-1}, \tag{62}$$

where $\mathbb{I}(\cdot) = 1$ if the predicate is true and 0 otherwise. For $k = 1, \ldots, K-1$ the conditional probability of ω_k given the other variables can be computed from Eq. (62) and the conditional probability of $Z^*_{1k}, \ldots, Z^*_{nk}$ given ω_k. For ω_K we also have to condition on $Z^*_{ik} = 0$ for all i and $k > K$; this probability can be computed using the Lévy-Khintchine representation for the beta process (Teh et al., 2007).

7.4 Inference for Hierarchical Beta Processes

In this section we present an inference algorithm for the hierarchical beta process given in Eq. (45). The observed data are the variables Z_{ji}; these binary vectors denote (in the language of document classification) the presence or absence of words in document i of group j. The underlying measure space Θ is

interpreted as the vocabulary. (Each element in Θ is referred to as a "word"). Let $\theta_1, \ldots, \theta_K \in \Theta$ denote the words that are observed among the documents. That is, these are the $\theta \in \Theta$ such that $Z_{ji}(\theta) = 1$ for some i and j.

Because both the beta and Bernoulli processes are completely random measures, the posterior over B_0 and B_j for $j \in \mathcal{J}$ decomposes into a discrete part over the observed vocabulary $\{\theta_1, \ldots, \theta_K\}$ and a continuous part over $\Theta \backslash \{\theta_1, \ldots, \theta_K\}$. The discrete part further factorizes over each observed word θ_k. Thus it is sufficient to focus separately on inference for each observed word and for the continuous part corresponding to unobserved words.

For a fixed θ_k, let $a = A(\theta_k)$, $\omega_0 = B_0(\theta_k)$, $\omega_j = B_j(\theta_k)$ and $z_{ji} = Z_{ji}(\theta_k)$. The slice of the HBP corresponding to θ_k has the following joint distribution:

$$\begin{aligned} \omega_0 \,|\, c_0, a &\sim \text{Beta}(c_0 a, c_0(1-a)) && (63) \\ \omega_j \,|\, c_j, \omega_0 &\sim \text{Beta}(c_j \omega_0, c_j(1-\omega_0)) && \text{for } j \in \mathcal{J} \\ z_{ji} \,|\, \omega_j &\sim \text{Bernoulli}(\omega_j) && \text{for } i = 1, \ldots, n_j. \end{aligned}$$

Note that the prior over ω_0 is improper if A is continuous and $a = 0$. This beta hierarchy is a special case of the finite Dirichlet hierarchy of Eq. (10) and Eq. (12) and it is straightforward to use the posterior representation sampler described in Section 7.1 to sample from the posterior given the observed z_{ji}. Thibaux and Jordan (2007) described an alternative where the ω_j are integrated out and rejection sampling is used to sample from ω_0.

Finally, we consider the continuous part of the posterior. This component is not simply the prior, since we have to condition on the fact that no words in $\Theta \backslash \{\theta_1, \ldots, \theta_K\}$ have been observed among the documents. Thibaux and Jordan (2007) solved this problem by noting that the posterior factors over the levels indexed by n in the size-biased ordering in Eq. (42). Focusing on each level separately, they derived a posterior distribution on the number of atoms in each level, combining this with the posterior over the level-specific weights to obtain the overall posterior.

8 Discussion

Our goal in this chapter has been to place hierarchical modeling in the same central role in Bayesian nonparametrics that it plays in other areas of Bayesian statistics. Indeed, one of the principal arguments for hierarchical modeling in parametric statistics is that it provides control over the large numbers of degrees of freedom that arise, for example, in random effects models. Such an argument holds a fortiori in the nonparametric setting.

Nonparametric priors generally involve hyperparameters, some of which are finite-dimensional and some of which are infinite-dimensional. Sharing the finite-dimensional parameter among multiple draws from such a prior is a natural modeling strategy that mimics classical hierarchical modeling concepts. It is our contention, however, that this form of control is far too limited, and that the infinite-dimensional parameters should generally also be shared. We have

made this point principally by considering examples in applied problem domains. In domains such as computational vision, information retrieval and genetics, nonparametric models provide natural descriptions of the complex objects under study; in particular, it is natural to describe an image, a document or a genome as the realization of a stochastic process. Now, in considering collections of such objects it is natural to want to share details of the realization among the objects in the collection—we wish to share parts of objects, features, recurring phrases and motifs. This can be achieved by coupling multiple draws from a nonparametric prior via their infinite-dimensional parameters.

Another advantage of hierarchical modeling in the classical setting is that it expands the repertoire of distributional forms that can be considered. For example, heavy-tailed distributions can be obtained by placing a prior on the scale parameter of lighter-tailed distributions. Although this point has been little explored to date in the nonparametric setting, we expect that it will be a fruitful direction for further research. In particular, there are stringent computational constraints that limit the nonparametric repertoire, and hierarchical constructions offer one way forward. Indeed, as we have seen, computationally oriented constructions such as urn models and stick-breaking representations often carry over naturally to hierarchical nonparametric models.

Finally, it is worth noting a difficulty that is raised by hierarchical modeling. Although Bayesian hierarchies help to control hyperparameters, they do not remove the need to specify distributions for hyperparameters. Indeed, when hyperparameters are placed high in a hierarchy it can be difficult to give operational meaning to such hyperparameters. One approach to coping with this issue involves considering the marginal probabilities that are induced by a nonparametric prior. For example, we argued that the marginals induced by a Pitman-Yor prior exhibit long tails that provide a good match to the power-law behavior found in textual data and image statistics. Further research is needed to develop this kind of understanding for a wider range of hierarchical Bayesian nonparametric models and problem domains.

8.1 Acknowledgements

We would like to thank David Blei, Jan Gasthaus, Sam Gershman, Tom Griffiths, Kurt Miller, Vinayak Rao and Erik Sudderth for their helpful comments on the manuscript.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, volume 16.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chen, S. F. and Goodman, J. T. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Chen, S. F. and Goodman, J. T. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

Chou, K. C., Willsky, A. S., and Benveniste, A. (1994). Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, 39(3):464–478.

Chu, W., Ghahramani, Z., Krause, R., and Wild, D. L. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *BIOCOMPUTING: Proceedings of the Pacific Symposium*.

Cowans, P. (2004). Information retrieval using hierarchical Dirichlet processes. In *Proceedings of the Annual International Conference on Research and Development in Information Retrieval*, volume 27, pages 564–565.

Cowans, P. (2006). *Probabilistic Document Modelling*. PhD thesis, University of Cambridge.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825.

Erosheva, E. (2003). Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, volume 7, pages 501–510.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Finkel, J. R., Grenager, T., and Manning, C. D. (2007). The infinite tree. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Fox, E., Sudderth, E., Jordan, M. I., and Willsky, A. (2009). An HDP-HMM for systems with state persistence. In *Advances in Neural Information Processing Systems*, volume 21, Cambridge, MA. MIT Press.

Getoor, L. and Taskar, B., editors (2007). *Introduction to Statistical Relational Learning*. MIT Press.

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). Bayesian nonparametric latent feature models (with discussion and rejoinder). In *Bayesian Statistics*, volume 8.

Goldwater, S., Griffiths, T., and Johnson, M. (2006a). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, volume 18.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006b). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Görür, D., Jäkel, F., and Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the International Conference on Machine Learning*, volume 23.

Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18.

Hiemstra, D. and Kraaij, W. (1998). Twenty-one at TREC-7: Ad-hoc and cross-language track. In *Text REtrieval Conference*, pages 174–185.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294.

Ho, M. W., James, L. F., and Lau, J. W. (2006). Coagulation fragmentation laws induced by general coagulations of two-parameter Poisson-Dirichlet processes. http://arxiv.org/abs/math.PR/0601608.

Huang, S. and Renals, S. (2007). Hierarchical Pitman-Yor language models for ASR in meetings. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, volume 10.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing*. Prentice-Hall, Upper Saddle River, NJ.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems*, volume 19.

Kim, S. and Smyth, P. (2007). Hierarchical dirichlet processes with random effects. In *Advances in Neural Information Processing Systems*, volume 19.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kivinen, J., Sudderth, E., and Jordan, M. I. (2007a). Image denoising with nonparametric hidden Markov trees. In *IEEE International Conference on Image Processing (ICIP)*, San Antonio, TX.

Kivinen, J., Sudderth, E., and Jordan, M. I. (2007b). Learning multiscale representations of natural scenes using Dirichlet processes. In *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Knowles, D. and Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation*, volume 7 of *Lecture Notes in Computer Science*. Springer.

Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

MacEachern, S., Kottas, A., and Gelfand, A. (2001). Spatial nonparametric Bayesian models. Technical Report 01-10, Institute of Statistics and Decision Sciences, Duke University. http://ftp.isds.duke.edu/WorkingPapers/01-10.html.

MacKay, D. and Peto, L. (1994). A hierarchical Dirichlet language model. *Natural Language Engineering*.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems*, volume 19.

Navarro, D. J. and Griffiths, T. L. (2007). A nonparametric Bayesian method for inferring features from similarity judgements. In *Advances in Neural Information Processing Systems*, volume 19.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31:705–767.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39.

Pitman, J. (2002). Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley. Lecture notes for St. Flour Summer School.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A., and Lau, M. (1992). Okapi at TREC. In *Text REtrieval Conference*, pages 21–30.

Salton, G. and McGill, M. (1983). *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989.

Sudderth, E. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*, volume 21.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2008). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77.

Teh, Y. W. (2006a). A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore.

Teh, Y. W. (2006b). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Teh, Y. W., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems*, volume 20.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Tomlinson, G. and Escobar, M. (2003). Analysis of densities. Talk given at the Joint Statistical Meeting.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the International Conference on Machine Learning*, volume 25.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45.

Wolpert, R. L. and Ickstadt, K. (1998). Simulations of lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 227–242. Springer-Verlag.

Wood, F., Griffiths, T. L., and Ghahramani, Z. (2006). A non-parametric Bayesian method for inferring hidden causes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 22.

Wooters, C. and Huijbregts, M. (2007). The ICSI RT07s speaker diarization system. In *Lecture Notes in Computer Science*. Springer.

Xing, E. P., Jordan, M. I., and Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14:267–284.

Xing, E. P. and Sohn, K. (2007). Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2(2).

Xing, E. P., Sohn, K., Jordan, M. I., and Teh, Y. W. (2006). Bayesian multipopulation haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the International Conference on Machine Learning*, volume 23.