Stick-breaking Construction for the Indian Buffet Process

Yee Whye Teh
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR, UK
ywteh@gatsby.ucl.ac.uk

Dilan Görür
MPI for Biological Cybernetics
Dept. Schölkopf
Spemannstrasse 38
72076 Tübingen, Germany
dilan.gorur@tuebingen.mpg.de

Zoubin Ghahramani
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ, UK
zoubin@eng.cam.ac.uk

Abstract

The Indian buffet process (IBP) is a Bayesian nonparametric distribution whereby objects are modelled using an unbounded number of latent features. In this paper we derive a stick-breaking representation for the IBP. Based on this new representation, we develop slice samplers for the IBP that are efficient, easy to implement and are more generally applicable than the currently available Gibbs sampler. This representation, along with the work of Thibaux and Jordan [17], also illuminates interesting theoretical connections between the IBP, Chinese restaurant processes, Beta processes and Dirichlet processes.

1 INTRODUCTION

The Indian Buffet Process (IBP) is a distribution over binary matrices consisting of $N > 0$ rows and an unbounded number of columns [6]. These binary matrices can be interpreted as follows: each row corresponds to an object, each column to a feature, and a 1 in entry (i, k) indicates object i has feature k. For example, objects can be movies like “Terminator 2”, “Shrek” and “Shanghai Knights”, while features can be “action”, “comedy”, “stars Jackie Chan”, and the matrix can be $[101; 010; 110]$ in Matlab notation.

Like the Chinese Restaurant Process (CRP) [1], the IBP provides a tool for defining nonparametric Bayesian models with latent variables. However, unlike the CRP, in which each object belongs to one *and only one* of infinitely many latent classes, the IBP allows each object to possess potentially *any combination* of infinitely many latent features. This added flexibility has resulted in a great deal of interest in the IBP, and the development of a range of interesting applications. These applications include models for choice behaviour [5], protein-protein interactions [2], the structure of causal graphs [19], dyadic data for collaborative filtering applications [10], and human similarity judgments [11].

In this paper, we derive a new, stick-breaking representation for the IBP, a development which is analogous to Sethuraman’s seminal stick-breaking representation for CRPs [15]. In this representation, as we will see in Section 3, the probability of each feature is represented explicitly by a stick of length between 0 and 1. Sethuraman’s representation paved the way for both novel samplers for and generalizations of CRPs [7]. Similarly, we show how our novel stick-breaking representation of the IBP can be used to develop new slice samplers for IBPs that are efficient, easy to implement and have better applicability to non-conjugate models (Sections 4, 5.2, 6). This new representation also suggests generalizations of the IBP (such as a Pitman-Yor variant, in Section 3.2). Moreover, although our stick-breaking representation of the IBP was derived from a very different model than the CRP, we demonstrate a surprising duality between the sticks in these two representations which suggests deeper connections between the two models (Section 3.2). The theoretical developments we describe here, which show a stick-breaking representation which is to the IBP what Sethuraman’s construction is to the CRP, along with the recent work of Thibaux and Jordan [17], showing that a particular subclass of Beta processes is to the IBP as the Dirichlet process is to the CRP, firmly establish the IBP in relation to the well-known classes of Bayesian nonparametric models.

2 INDIAN BUFFET PROCESSES

The IBP is defined as the limit of a corresponding distribution over matrices with K columns, as the number of columns $K \to \infty$. Let Z be a random binary $N \times K$ matrix, and denote entry (i, k) in Z by z_{ik}. For each feature k let μ_k be the prior probability that feature k is present in an object. We place a $\mathrm{Beta}(\frac{\alpha}{K}, 1)$ prior on μ_k, with α being the strength parameter of the IBP. The full model is:

$$\mu_k \sim \mathrm{Beta}(\tfrac{\alpha}{K}, 1) \qquad \text{indepedently } \forall k \qquad (1a)$$

$$z_{ik}|\mu_k \sim \mathrm{Bernoulli}(\mu_k) \qquad \text{indepedently } \forall i, k \qquad (1b)$$

Let us now consider integrating out the μ_k’s and taking the limit of $K \to \infty$ to obtain the IBP. For the first object, the chance of it having each particular feature k is independently $\frac{\alpha}{K}$ once μ_k is integrated out, thus the distribution over the number of features it has is $\text{Binomial}(\frac{\alpha}{K}, K)$. As $K \to \infty$, this approaches $\text{Poisson}(\alpha)$. For subsequent objects $i = 2, \ldots, N$, the probability of it also having a feature k already belonging to a previous object is $\frac{\frac{\alpha}{K}+m_{<i\,k}}{\frac{\alpha}{K}+1+i-1} \to \frac{m_{<i\,k}}{i}$ where $m_{<i\,k} = \sum_{j<i} z_{jk} > 0$ is the number of objects prior to i with feature k. Repeating the argument for the first object, object i will also have $\text{Poisson}(\frac{\alpha}{i})$ new features not belonging to previous objects. Note that even though the total number of available features is unbounded, the actual number K^+ of used features is always finite (and in fact is distributed as $\text{Poisson}(\alpha \sum_{i=1}^{N} \frac{1}{i})$).

The above generative process can be understood using the metaphor of an Indian buffet restaurant. Customers (objects) come into the restaurant one at a time, and can sample an infinite number of dishes (features) at the buffet counter. Each customer will try each dish that previous customers have tried with probabilities proportional to how popular each dish is; in addition the customer will try a number of new dishes that others have not tried before.

To complete the model, let θ_k be parameters associated with feature k and x_i be an observation associated with object i. Let

$$\theta_k \sim H \qquad \text{independently } \forall k \qquad (2a)$$

$$x_i \sim F(z_{i,:}, \theta_:) \qquad \text{independently } \forall i \qquad (2b)$$

where H is the prior over parameters, $F(z_{i,:}, \theta_:)$ is the data distribution given the features $z_{i,:} = \{z_{ik}\}_{k=1}^{\infty}$ corresponding to object i and feature parameters $\theta_: = \{\theta_k\}_{k=1}^{\infty}$. We assume that $F(z_{i,:}, \theta_:)$ depends only on the parameters of the present features.

2.1 GIBBS SAMPLER

The above generative process for the IBP can be used directly in a Gibbs sampler for posterior inference of Z and θ given data $\mathbf{x} = \{x_i\}$ [6]. The representation consists of the number K^+ of used (active) features, the matrix $Z_{1:N,1:K^+}$ of occurrences among the K^+ active features, and their parameters $\theta_{1:K^+}$. The superscript $^+$ denotes active features. The sampler iterates through $i = 1, \ldots, N$, for each object i it updates the feature occurrences for the currently used features, then considers adding new features to model the data x_i.

For the already used features $k = 1, \ldots, K^+$, the conditional probability of $z_{ik} = 1$ given other variables is just

$$p(z_{ik} = 1|\text{rest}) \propto \frac{m_{\neg i\,k}}{N} f(x_i|z_{i,\neg k}, z_{ik} = 1, \theta_{1:K^+}) \quad (3)$$

where $m_{\neg i\,k} = \sum_{j \neq i} z_{jk}$. The fraction is the conditional prior of $z_{ik} = 1$, obtained by using exchangeability among the customers and taking customer i to be the last customer to enter the restaurant; the second term $f(\cdot|\cdot)$ is the data likelihood of x_i if $z_{ik} = 1$. It is possible to integrate $\theta_{1:K^+}$ out from the likelihood term if H is conjugate to F. In fact it is important for H to be conjugate to F when we consider the probabilities of new features being introduced, because all possible parameters for these new features have to be taken into account. If L_i is the number of new features introduced, we have

$$p(L_i|\text{rest}) \propto \frac{(\frac{\alpha}{N})^{L_i} e^{-\frac{\alpha}{N}}}{L_i!} \times \int f(x_i|z_{i,1:K^+}, z^\circ_{i,1:L_i} = 1, \theta_{1:K^+}, \theta^\circ_{1:L_i})\, dh(\theta^\circ_{1:L_i}) \quad (4)$$

where $z^\circ_{i,1:L_i}$ are occurrences for the new features and $\theta^\circ_{1:L_i}$ are their parameters, the superscript $^\circ$ denoting currently unused (inactive) features. The fraction comes from the probability of introducing L_i new features under $\text{Poisson}(\frac{\alpha}{N})$ while the second term is the data likelihood, with the parameters $\theta^\circ_{1:L_i}$ integrated out with respect to the prior density $h(\cdot)$.

The need to integrate out the parameters for new features is similar to the need to integrate out parameters for new clusters in the Dirichlet process (DP) mixture model case (see [13]). To perform this integration efficiently, conjugacy is important, but the requirement for conjugacy limits the applicability of the IBP in more elaborate settings. It is possible to devise samplers in the non-conjugate case analogous to those developed for DP mixture models [10, 5]. In the next section we develop a different representation of the IBP in terms of a stick-breaking construction, which leads us to an easy to implement slice sampler for the non-conjugate case.

3 STICK BREAKING CONSTRUCTION

In this section, we describe an alternative representation of the IBP where the feature probabilities are not integrated out, and a specific ordering is imposed on the features. We call this the stick-breaking construction for the IBP. We will see that the new construction bears strong relationships with the standard stick-breaking construction for CRPs, paving the way to novel generalizations of and inference techniques for the IBP.

3.1 DERIVATION

Let $\mu_{(1)} > \mu_{(2)} > \ldots > \mu_{(K)}$ be a decreasing ordering of $\mu_{1:K} = \{\mu_1, \ldots, \mu_K\}$, where each μ_l is $\text{Beta}(\frac{\alpha}{K}, 1)$. We will show that in the limit $K \to \infty$ the $\mu_{(k)}$'s obey the following law, which we shall refer to as the *stick-breaking construction* for the IBP,

$$\nu_{(k)} \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1) \qquad \mu_{(k)} = \nu_{(k)}\mu_{(k-1)} = \prod_{l=1}^{k} \nu_{(l)} \quad (5)$$

We start by considering $\mu_{(1)}$. For finite K it is

$$\mu_{(1)} = \max_{l=1,\ldots,K} \mu_l \tag{6}$$

where each μ_l is $\mathrm{Beta}(\frac{\alpha}{K}, 1)$ and has density:

$$p(\mu_l) = \tfrac{\alpha}{K}\mu_l^{\frac{\alpha}{K}-1}\mathbb{I}(0 \le \mu_l \le 1) \tag{7}$$

where $\mathbb{I}(A)$ is the indicator function for a condition (measurable set) A: $\mathbb{I}(A) = 1$ if A is true, and 0 otherwise. The cumulative distribution function (cdf) for μ_l is then:

$$\begin{aligned} F(\mu_l) &= \int_{-\infty}^{\mu_l} \tfrac{\alpha}{K} t^{\frac{\alpha}{K}-1}\mathbb{I}(0 \le t \le 1)\,dt \\ &= \mu_l^{\frac{\alpha}{K}}\mathbb{I}(0 \le \mu_l \le 1) + \mathbb{I}(1 < \mu_l) \end{aligned} \tag{8}$$

Since the μ_l's are independent, the cdf of $\mu_{(1)}$ is just the product of the cdfs of each μ_l, so

$$\begin{aligned} F(\mu_{(1)}) &= \left(\mu_{(1)}^{\frac{\alpha}{K}}\mathbb{I}(0 \le \mu_{(1)} \le 1) + \mathbb{I}(1 < \mu_{(1)} < \infty)\right)^K \\ &= \mu_{(1)}^{\alpha}\mathbb{I}(0 \le \mu_{(1)} \le 1) + \mathbb{I}(1 < \mu_{(1)}) \end{aligned} \tag{9}$$

Differentiating, we see that the density of $\mu_{(1)}$ is

$$p(\mu_{(1)}) = \alpha\mu_{(1)}^{\alpha-1}\mathbb{I}(0 \le \mu_{(1)} \le 1) \tag{10}$$

and therefore $\mu_{(1)} \sim \mathrm{Beta}(\alpha, 1)$.

We now derive the densities for subsequent $\mu_{(k)}$'s. For each $k \ge 1$ let l_k be such that $\mu_{l_k} = \mu_{(k)}$ and let $\mathbf{L}_k = \{1, \ldots, K\}\backslash\{l_1, \ldots, l_k\}$. Since $\mu_{(1:k)} = \{\mu_{(1)}, \ldots, \mu_{(k)}\}$ are the k largest values among $\mu_{1:K}$, we have

$$\mu_l \le \min_{k' \le k} \mu_{(k')} = \mu_{(k)} \tag{11}$$

for each $l \in \mathbf{L}_k$. Restricting the range of μ_l to $[0, \mu_{(k)}]$, the cdf becomes

$$\begin{aligned} F(\mu_l|\mu_{(1:k)}) &= \frac{\int_0^{\mu_l} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\,dt}{\int_0^{\mu_{(k)}} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\,dt} \\ =&\mu_{(k)}^{-\frac{\alpha}{K}}\mu_l^{\frac{\alpha}{K}}\mathbb{I}(0 \le \mu_l \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_l) \end{aligned} \tag{12}$$

Now $\mu_{(k+1)} = \max_{l \in \mathbf{L}_k} \mu_l$ with each μ_l independent given $\mu_{(1:k)}$. The cdf of $\mu_{(k+1)}$ is again the product of the cdfs of μ_l over $l \in \mathbf{L}_k$,

$$\begin{aligned} &F(\mu_{(k+1)}|\mu_{(1:k)}) \qquad (13) \\ =&\mu_{(k)}^{-\frac{K-k}{K}\alpha}\mu_{(k+1)}^{\frac{K-k}{K}\alpha}\mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_{(k+1)}) \\ \to&\mu_{(k)}^{-\alpha}\mu_{(k+1)}^{\alpha}\mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_{(k+1)}) \end{aligned}$$

as $K \to \infty$. Differentiating, the density of $\mu_{(k+1)}$ is,

$$\begin{aligned} &p(\mu_{(k+1)}|\mu_{(1:k)}) \\ =&\alpha\mu_{(k)}^{-\alpha}\mu_{(k+1)}^{\alpha-1}\mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)}) \end{aligned} \tag{14}$$

Notice that the $\mu_{(k)}$'s have a Markov structure, with $\mu_{(k+1)}$ conditionally independent of $\mu_{(1:k-1)}$ given $\mu_{(k)}$.

Finally, instead of working with the variables $\mu_{(k)}$ directly, we introduce a new set of variables $\nu_{(k)} = \frac{\mu_{(k)}}{\mu_{(k-1)}}$ with range $[0, 1]$. Using a change of variables, the density of $\nu_{(k)}$ is derived to be,

$$p(\nu_{(k)}|\mu_{(1:k-1)}) = \alpha\nu_{(k)}^{\alpha-1}\mathbb{I}(0 \le \nu_{(k)} \le 1) \tag{15}$$

Thus $\nu_{(k)}$ are independent from $\mu_{(1:k-1)}$ and are simply $\mathrm{Beta}(\alpha, 1)$ distributed. Expanding $\mu_{(k)} = \nu_{(k)}\mu_{(k-1)} = \prod_{l=1}^{k} \nu_{(l)}$, we obtain the stick-breaking construction (5).

The construction (5) can be understood metaphorically as follows. We start with a stick of length 1. At iteration $k = 1, 2, \ldots$, we break off a piece at a point $\nu_{(k)}$ relative to the current length of the stick. We record the length $\mu_{(k)}$ of the stick we just broke off, and recurse on this piece, discarding the other piece of stick.

3.2 RELATION TO DP

In iteration k of the construction (5), after breaking the stick in two we always recurse on the stick whose length we denote by $\mu_{(k)}$. Let $\pi_{(k)}$ be the length of the other discarded stick. We have,

$$\pi_{(k)} = (1 - \nu_{(k)})\mu_{(k-1)} = (1 - \nu_{(k)})\prod_{l=1}^{k-1}\nu_{(l)} \tag{16}$$

Making a change of variables $v_{(k)} = 1 - \nu_{(k)}$,

$$v_{(k)} \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha) \qquad \pi_{(k)} = v_{(k)}\prod_{l=1}^{k-1}(1 - v_{(l)}) \tag{17}$$

thus $\pi_{(1:\infty)}$ are the resulting stick lengths in a standard stick-breaking construction for DPs [15, 7].

In both constructions the final weights of interest are the lengths of the sticks. In DPs, the weights $\pi_{(k)}$ are the lengths of sticks discarded, while in IBPs, the weights $\mu_{(k)}$ are the lengths of sticks we have left. This difference leads to the different properties of the weights: for DPs, the stick lengths sum to a length of 1 and are not decreasing, while in IBPs the stick lengths need not sum to 1 but are decreasing. Both stick-breaking constructions are shown in Figure 1. In both the weights decrease exponentially quickly in expectation.

The direct correspondence to stick-breaking in DPs implies that a range of techniques for and extensions to the DP can be adapted for the IBP. For example, we can generalize the IBP by replacing the $\mathrm{Beta}(\alpha, 1)$ distribution on $\nu_{(k)}$'s with other distributions. One possibility is a Pitman-Yor [14] extension of the IBP, defined as

$$\nu_{(k)} \sim \mathrm{Beta}(\alpha + kd, 1 - d) \qquad \mu_{(k)} = \prod_{l=1}^{k}\nu_{(l)} \tag{18}$$

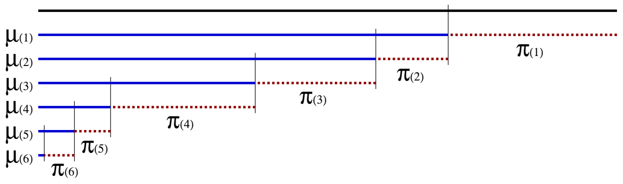

Figure 1: Stick-breaking construction for the DP and IBP. The black stick at top has length 1. At each iteration the vertical black line represents the break point. The brown dotted stick on the right is the weight obtained for the DP, while the blue stick on the left is the weight obtained for the IBP.

where $d \in [0,1)$ and $\alpha > -d$. The Pitman-Yor IBP weights decrease in expectation as a $O(k^{-\frac{1}{d}})$ power-law, and this may be a better fit for some naturally occurring data which have a larger number of features with significant but small weights [4].

An example technique for the DP which we could adapt to the IBP is to truncate the stick-breaking construction after a certain number of break points and to perform inference in the reduced space. [7] gave a bound for the error introduced by the truncation in the DP case which can be used here as well. Let K^* be the truncation level. We set $\mu_{(k)} = 0$ for each $k > K^*$, while the joint density of $\mu_{(1:K^*)}$ is,

$$p(\mu_{(1:K^*)}) = \prod_{k=1}^{K^*} p(\mu_{(k)}|\mu_{(k-1)}) \qquad (19)$$

$$=\alpha^{K^*}\mu_{(K^*)}^{\alpha}\prod_{k=1}^{K^*}\mu_{(k)}^{-1}\mathbb{I}(0 \le \mu_{(K^*)} \le \cdots \le \mu_{(1)} \le 1)$$

The conditional distribution of Z given $\mu_{(1:K^*)}$ is simply[1]

$$p(Z|\mu_{(1:K^*)}) = \prod_{i=1}^{N}\prod_{k=1}^{K^*}\mu_{(k)}^{z_{ik}}(1-\mu_{(k)})^{1-z_{ik}} \qquad (20)$$

with $z_{ik} = 0$ for $k > K^*$. Gibbs sampling in this representation is straightforward, the only point to note being that adaptive rejection sampling (ARS) [3] should be used to sample each $\mu_{(k)}$ given other variables (see next section).

4 SLICE SAMPLER

Gibbs sampling in the truncated stick-breaking construction is simple to implement, however the predetermined truncation level seems to be an arbitrary and unnecessary approximation. In this section, we propose a non-approximate scheme based on slice sampling, which can be seen as adaptively choosing the truncation level at each iteration. Slice sampling is an auxiliary variable method that samples from a distribution by sampling uniformly from the region under its density function [12]. This turns the problem of sampling from an arbitrary distribution to sampling from uniform distributions. Slice sampling has been successfully applied to DP mixture models [8], and our application to the IBP follows a similar thread.

In detail, we introduce an auxiliary slice variable,

$$s|Z,\mu_{(1:\infty)} \sim \mathrm{Uniform}[0,\mu^*] \qquad (21)$$

where μ^* is a function of $\mu_{(1:\infty)}$ and Z, and is chosen to be the length of the stick for the last active feature,

$$\mu^* = \min\left\{1, \min_{k:\,\exists i, z_{ik}=1} \mu_{(k)}\right\}. \qquad (22)$$

The joint distribution of Z and the auxiliary variable s is

$$p(s,\mu_{(1:\infty)},Z) = p(Z,\mu_{(1:\infty)})\,p(s|Z,\mu_{(1:\infty)}) \qquad (23)$$

where $p(s|Z,\mu_{(1:\infty)}) = \frac{1}{\mu^*}\mathbb{I}(0 \le s \le \mu^*)$. Clearly, integrating out s preserves the original distribution over $\mu_{(1:\infty)}$ and Z, while conditioned on Z and $\mu_{(1:\infty)}$, s is simply drawn from (21). Given s, the distribution of Z becomes:

$$p(Z|\mathbf{x},s,\mu_{(1:\infty)}) \propto p(Z|\mathbf{x},\mu_{(1:\infty)})\tfrac{1}{\mu^*}\mathbb{I}(0 \le s \le \mu^*) \qquad (24)$$

which forces all columns k of Z for which $\mu_{(k)} < s$ to be zero. Let K^* be the maximal feature index with $\mu_{(K^*)} > s$. Thus $z_{ik} = 0$ for all $k > K^*$, and we need only consider updating those features $k \le K^*$. Notice that K^* serves as a truncation level insofar as it limits the computational costs to a finite amount without approximation.

Let $K^\dagger$ be an index such that all active features have index $k < K^\dagger$ (note that $K^\dagger$ itself would be an inactive feature). The computational representation for the slice sampler consists of the slice variables and the first $K^\dagger$ features: $\langle s, K^*, K^\dagger, Z_{1:N,1:K^\dagger}, \mu_{(1:K^\dagger)}, \theta_{1:K^\dagger}\rangle$. The slice sampler proceeds by updating all variables in turn.

Update s. The slice variable is drawn from (21). If the new value of s makes $K^* \ge K^\dagger$ (equivalently, $s < \mu_{(K^\dagger)}$), then we need to pad our representation with inactive features until $K^* < K^\dagger$. In the appendix we show that the stick lengths $\mu_{(k)}$ for new features k can be drawn iteratively from the following distribution:

$$p(\mu_{(k)}|\mu_{(k-1)}, z_{:,>k}=0) \propto \exp(\alpha \textstyle\sum_{i=1}^{N}\frac{1}{i}(1-\mu_{(k)})^i)$$
$$\mu_{(k)}^{\alpha-1}(1-\mu_{(k)})^N\mathbb{I}(0 \le \mu_{(k)} \le \mu_{(k-1)}) \qquad (25)$$

We used ARS to draw samples from (25) since it is log-concave in $\log \mu_{(k)}$. The columns for these new features are initialized to $z_{:,k} = 0$ and their parameters drawn from their prior $\theta_k \sim H$.

[1] Note that we are making a slight abuse of notation by using Z both to denote the original IBP matrix with arbitrarily ordered columns, and the equivalent matrix with the columns reordered to decreasing μ's. Similarly for the feature parameters θ's.

Update Z. Given s, we only need to update z_{ik} for each i and $k \leq K^*$. The conditional probabilities are:

$$p(z_{ik} = 1|\text{rest}) \propto \frac{\mu_{(k)}}{\mu^*} f(x_i|z_{i,\neg k}, z_{ik} = 1, \theta_{1:K^\dagger}) \quad (26)$$

The μ^* denominator is needed when different values of z_{ik} induces different values of μ^* by changing the index of the last active feature.

Update θ_k. For each $k = 1, \ldots, K^\dagger$, the conditional probability of θ_k is:

$$p(\theta_k|\text{rest}) \propto h(\theta_k) \prod_{i=1}^{N} f(x_i|z_{i,1:K^\dagger}, \theta_{\neg k}, \theta_k) \quad (27)$$

Update $\mu_{(k)}$. For $k = 1, \ldots, K^\dagger - 1$, combining (19) and (20), the conditional probability of $\mu_{(k)}$ is

$$\begin{aligned} p(\mu_{(k)}|\text{rest}) \propto & \mu_{(k)}^{m_{\cdot k}-1}(1 - \mu_{(k)})^{N-m_{\cdot k}} \\ & \mathbb{I}(\mu_{(k+1)} \leq \mu_{(k)} \leq \mu_{(k-1)}) \end{aligned} \quad (28)$$

where $m_{\cdot k} = \sum_{i=1}^{N} z_{ik}$. For $k = K^\dagger$, in addition to taking into account the probability of features $K^\dagger$ is inactive, we also have to take into account the probability that all columns of Z beyond $K^\dagger$ are inactive as well. The appendix shows that the resulting conditional probability of $\mu_{(K^\dagger)}$ is given by (25) with $k = K^\dagger$. We draw from both (28) and (25) using ARS.

5 CHANGE OF REPRESENTATIONS

Both the stick-breaking construction and the standard IBP representation are different representations of the same nonparametric object. In this section we consider updates which change from one representation to the other. More precisely, given a posterior sample in the stick-breaking representation we wish to construct a posterior sample in the IBP representation and vice versa. Such changes of representation allow us to make use of efficient MCMC moves in both representations, e.g. interlacing split-merge moves in IBP representation [10] with the slice sampler in stick-breaking representation. Furthermore, since both stick lengths and the ordering of features are integrated out in the IBP representation, we can efficiently update both in the stick-breaking representation by changing to the IBP representation and back.

We appeal to the infinite limit formulation of both representations to derive the appropriate procedures. In particular, we note that the IBP is obtained by ignoring the ordering on features and integrating out the weights $\mu_{(1:K)}$ in an arbitrarily large finite model, while the stick-breaking construction is obtained by enforcing an ordering with decreasing weights. Thus, given a sample in either representations, our approach is to construct a corresponding sample in an arbitrarily large finite model, then to either ignore the ordering and weights (to get IBP) or to enforce the decreasing weight ordering (to get stick-breaking).

Changing from stick-breaking to the standard IBP representation is easy. We simply drop the stick lengths as well as the inactive features, leaving us with the K^+ active feature columns along with the corresponding parameters. To change from IBP back to the stick-breaking representation, we have to draw both the stick lengths and order the features in decreasing stick lengths, introducing inactive features into the representation if required. We may index the K^+ active features in the IBP representation as $k = 1, \ldots, K^+$ in the finite model. Let $Z_{1:N,1:K^+}$ be the feature occurrence matrix. Suppose that we have $K \gg K_+$ features in the finite model. For the active features, the posterior for the lengths are simply

$$\begin{aligned} \mu_k^+|z_{:,k} \sim & \text{Beta}(\frac{\alpha}{K} + m_{\cdot k}, 1 + N - m_{\cdot k}) \\ & \rightarrow \text{Beta}(m_{\cdot k}, 1 + N - m_{\cdot k}) \end{aligned} \quad (29)$$

as $K \rightarrow \infty$. For the rest of the $K - K^+$ inactive features, it is sufficient to consider only those inactive features with stick lengths larger than $\min_k \mu_k^+$. Thus we consider a decreasing ordering $\mu_{(1)}^\circ > \mu_{(2)}^\circ > \cdots$ on these lengths. (25) gives their densities in the $K \rightarrow \infty$ limit and ARS can be used to draw $\mu_{(1:K^\circ)}^\circ$ until $\mu_{(K^\circ)}^\circ < \min_k \mu_k^+$. Finally, the stick-breaking representation is obtained by reordering $\mu_{1:K^+}^+, \mu_{(1:K^\circ)}^\circ$ in decreasing order, with the feature columns and parameters taking on the same ordering (columns and parameters corresponding to inactive features are set to 0 and drawn from their prior respectively), giving us $K^+ + K^\circ$ features in the stick-breaking representation.

5.1 SEMI-ORDERED STICK-BREAKING

In deriving the change of representations from the IBP to the stick-breaking representation, we made use of an intermediate representation whereby the active features are unordered, while the inactive ones have an ordering of decreasing stick lengths. It is in fact possible to directly work with this representation, which we shall call semi-ordered stick-breaking.

The representation consists of K^+ active and unordered features, as well as an ordered sequence of inactive features. The stick lengths for the active features have conditional distributions:

$$\mu_k^+|z_{:,k} \sim \text{Beta}(m_{\cdot,k}, 1 + N - m_{\cdot,k}) \quad (30)$$

while for the inactive features we have a Markov property:

$$\begin{aligned} p(\mu_{(k)}^\circ|\mu_{(k-1)}^\circ, z_{:,>k} = 0) \propto & \exp(\textstyle\sum_{i=1}^{N} \frac{1}{i}(1 - \mu_{(k)}^\circ)^i)) \\ & (\mu_{(k)}^\circ)^{\alpha-1}(1 - \mu_{(k)}^\circ)^N \mathbb{I}(0 \leq \mu_{(k)}^\circ \leq \mu_{(k-1)}^\circ) \end{aligned} \quad (31)$$

5.2 SLICE SAMPLER

To use the semi-ordered stick-breaking construction as a representation for inference, we can again use the slice sampler to adaptively truncate the representation for inactive features. This gives an inference scheme which works in the non-conjugate case, is not approximate, has an adaptive truncation level, but without the restrictive ordering constraint of the stick-breaking construnction. The representation $\langle s, K^+, Z_{1:N,1:K^+}, \mu^+_{1:K^+}, \theta_{1:K^+}\rangle$ consists only of the K^+ active features and the slice variable s,

$$s \sim \text{Uniform}[0, \mu^*] \quad \mu^* = \min\left\{1, \min_{1\leq k\leq K^+} \mu^+_k\right\} \tag{32}$$

Once a slice value is drawn, we generate K° inactive features, with their stick lengths drawn from (31) until $\mu^\circ_{(K^\circ+1)} < s$. The associated feature columns $Z^\circ_{1:N,1:K^\circ}$ are initialized to 0 and the parameters $\theta^\circ_{1:K^\circ}$ drawn from their prior. Sampling for the feature entries and parameters for both the active and just generated inactive features proceed as before. Afterwards, we drop from the list of active features any that became inactive, while we add to the list any inactive feature that became active. Finally, the stick lengths for the new list of active features are drawn from their conditionals (30).

6 EXPERIMENT

In this section we compare the mixing performance of the two proposed slice samplers against Gibbs sampling. We chose a simple synthetic dataset so that we can be assured of convergence to the true posterior and that mixing times can be estimated reliably in a reasonable amount of computation time. We also chose to apply the three samplers on a conjugate model since Gibbs sampling requires conjugacy, although our implementation of the two slice samplers did not make use of this. In the next section we demonstrate the modelling performance of a non-conjugate model using the semi-ordered slice sampler on a dataset of MNIST handwritten digits.

We used the conjugate linear-Gaussian binary latent feature model for comparing the performances of the different samplers [6]. Each data point x_i is modelled using a spherical Gaussian with mean $z_{i,:}A$ and variance σ^2_X, where $z_{i,:}$ is the row vector of feature occurrences corresponding to x_i, and A is a matrix whose kth row forms the parameters for the kth feature. Entries of A are drawn i.i.d. from a zero mean Gaussian with variance σ^2_A. We generated $1, 2$ and 3 dimensional datasets from the model with data variance fixed at $\sigma^2_X = 1$, varying values of the strength parameter $\alpha = 1, 2$ and the latent feature variance $\sigma^2_A = 1, 2, 4, 8$. For each combination of parameters we produced five datasets with 100 data points, giving a total of 120 datasets. For all datasets, we fixed σ^2_X and σ^2_A to the generating values and learned the feature matrix Z and α.

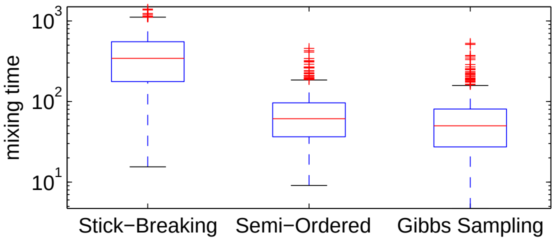

Figure 2: Autocorrelation times for K^+ for the slice sampler in decreasing stick lengths ordering, in semi-ordered stick-breaking representation, and for the Gibbs sampler.

For each dataset and each sampler, we repeated 5 runs of $15,000$ iterations. We used the autocorrelation coefficients of the number of represented features K^+ and α (with a maximum lag of 2500) as measures of mixing time. We found that mixing in K^+ is slower than in α for all datasets and report results only for K^+ here. We also found that in this regime the autocorrelation times do not vary with dimensionality or with σ^2_A. In Figure 2 we report the autocorrelation times of K^+ over all runs, all datasets, and all three samplers. As expected, the slice sampler using the decreasing stick lengths ordering was always slower than the semi-ordered one. Surprisingly, we found that the semi-ordered slice sampler was just as fast as the Gibbs sampler which fully exploits conjugacy. This is about as well as we would expect a more generally applicable non-conjugate sampler to perform.

7 DEMONSTRATION

In this section we apply the semi-ordered slice sampler to 1000 examples of handwritten images of 3's in the MNIST dataset. The model we worked with is a generalization of that in Section 6, where in addition to modelling feature occurrences, we also model per object features values [6]. In particular, let Y be a matrix of the same size as Z, with i.i.d. zero mean unit variance Gaussian entries. We model each x_i as

$$x_i|Z, Y, A, \sigma^2_X \sim \mathcal{N}((z_{i,:} \odot y_{i,:})A, \sigma^2_X \text{I}), \tag{33}$$

where $\odot$ is elementwise multiplication. Specification for the rest of the model is as in Section 6. We can integrate Y or A out while maintaining tractability, but not both.

The handwritten digit images are first preprocessed by projecting on to the first 64 PCA components, and the sampler ran for 10000 iterations. The trace plot of the log likelihood and the distribution of the number of active features are shown in Figure 3. The model succesfully finds latent features to reconstruct the images as shown in Figure 4. Some of the latent features found are shown in Figure 5. Most appear to model local shifts of small edge segments

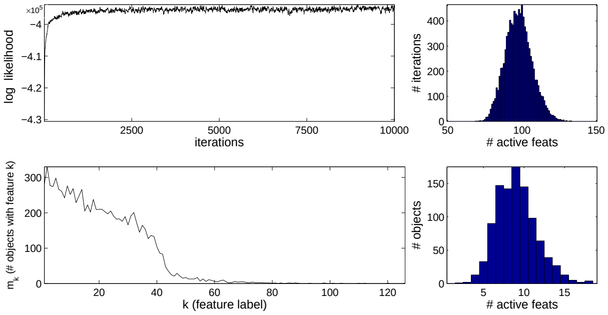

Figure 3: *Top-left*: the log likelihood trace plot. The sampler quickly finds a high likelihood region. *Top-right*: histogram of the number of active features over the 10000 iterations. *Bottom-left*: number of images sharing each feature during the last MCMC iteration. *Bottom-right*: histogram of the number of active features used by each input image. Note that about half of the features are used by only a few data points, and each data point is represented by a small subset of the active features.

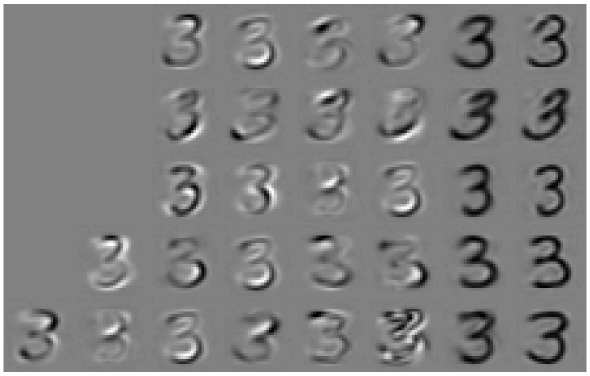

Figure 4: *Last column*: original digits. *Second last column*: reconstructed digits. *Other columns*: features used for reconstruction.

of the digits, and are reminiscent of the result of learning models with sparse priors (e.g. ICA) on such images [16].

8 DISCUSSION AND FUTURE WORK

We have derived novel stick-breaking representations of the Indian buffet process. Based on these representations new MCMC samplers are proposed that are easy to implement and work on more general models than Gibbs sampling. In experiments we showed that these samplers are just as efficient as Gibbs without using conjugacy.

[17] have recently showed that the IBP is a distribution on matrices induced by the Beta process with a constant strength parameter of 1. This relation to the Beta process is proving to be a fertile ground for interesting developments. A direct consequence of our stick-breaking construction is that a draw from such a Beta process has the form $A = \sum_{k=1}^{\infty} \mu_{(k)} \delta_{\theta_k}$ with $\mu_{(k)}$ drawn from (5) and θ_k drawn i.i.d. from the base measure H. This is a particularly simply case of a more general construction called the inverse Lévy measure [18, 9]. Generalizations to using other stick-breaking constructions automatically lead to generalizations of the Beta process, and we are currently exploring a number of possibilities, including the Pitman-Yor extension. Finally, the duality observed in Section 3.2 seems to be a hitherto unknown connection between the Beta process and the DP which we are currently trying to understand.

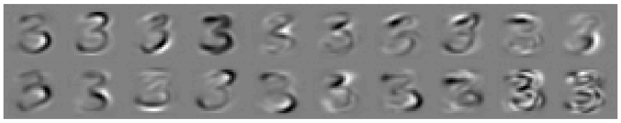

Figure 5: Features that are shared between many digits.

As an aside, it is interesting to note the importance of feature ordering in the development of the IBP. To make the derivation rigorous, [6] had to carefully ignore the feature ordering by considering permutation-invariant equivalence classes before taking the infinite limit. In this paper, we derived the stick-breaking construction by imposing a feature ordering with decreasing feature weights.

To conclude, our development of a stick-breaking construction for the IBP has lead to interesting insights and connections, as well as practical algorithms such as the new slice samplers.

ACKNOWLEDGEMENTS

We thank the reviewers for insightful comments. YWT thanks the Lee Kuan Yew Endowment Fund for funding.

REFERENCES

[1] D. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin, 1985.

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

[3] W.R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[4] S. Goldwater, T.L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-

law generators. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

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

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

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

[8] M. Kalli and S. G. Walker. Slice sampling for the Dirichlet process mixture model. Poster presented at the Eighth Valencia International Meeting on Bayesian Statistics, 2006.

[9] P. Lévy. *Théorie de L'Addition des Variables Aléatoires*. Paris: Gauthier-Villars, 1937.

[10] E. Meeds, Z. Ghahramani, R. Neal, and S. T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems*, volume 19, to appear 2007.

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

[12] R. M. Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2003.

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

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

[15] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[16] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:1235–1260, Dec 2003.

[17] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. This volume, 2007.

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

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

APPENDIX

Recall from the construction of $\mu_{(k+1:\infty)}$ that it is simply the $K \to \infty$ limit of a decreasing ordering of $\mu_{\mathbf{L}_k}$. Since reordering does not affect the probabilities of z_{il}'s given the corresponding μ_l for each $l \in \mathbf{L}_k$,

$$\begin{aligned}&p(z_{:,>k} = 0|\mu_{(k)})\\ &= \lim_{K\to\infty} \int p(\mu_{\mathbf{L}_k}|\mu_{(k)})p(z_{:,\mathbf{L}_k} = 0|\mu_{\mathbf{L}_k})\, d\mu_{\mathbf{L}_k}\end{aligned}$$

Given $\mu_{(k)}$, μ_l's and z_{il}'s are conditionally i.i.d. across different l's, with cdf of μ_l as given in (12). Thus we have

$$= \lim_{K\to\infty} \left(\int_0^{\mu_{(k)}} (1-\mu)^N \tfrac{\alpha}{K} \mu_{(k)}^{-\frac{\alpha}{K}} \mu^{\frac{\alpha}{K}-1}\, d\mu \right)^{K-k} \tag{34}$$

Applying change of variables $\nu = \mu/\mu_{(k)}$ to the integral,

$$\begin{aligned}&\int_0^1 (1-\nu\mu_{(k)})^N \tfrac{\alpha}{K} \nu^{\frac{\alpha}{K}-1}\, d\nu\\ &= \int_0^1 (1-\nu+\nu(1-\mu_{(k)}))^N \tfrac{\alpha}{K} \nu^{\frac{\alpha}{K}-1}\, d\nu\\ &= \int_0^1 \sum_{i=0}^N \tbinom{N}{i} (1-\nu)^{N-i} (\nu(1-\mu_{(k)}))^i \tfrac{\alpha}{K} \nu^{\frac{\alpha}{K}-1}\, d\nu\\ &= \sum_{i=0}^N \tbinom{N}{i} (1-\mu_{(k)})^i \tfrac{\alpha}{K} \tfrac{\Gamma(\frac{\alpha}{K}+i)\Gamma(N-i+1)}{\Gamma(\frac{\alpha}{K}+i+N-i+1)}\\ &= \sum_{i=0}^N \tfrac{N!}{(N-i)!i!} (1-\mu_{(k)})^i \tfrac{\alpha}{K} \tfrac{\prod_{j=0}^{i-1}(\frac{\alpha}{K}+j)(N-i)!}{\prod_{j=0}^{N}(\frac{\alpha}{K}+j)}\\ &= \left(\tfrac{N!}{\prod_{j=1}^N \frac{\alpha}{K}+j} \right) \left(1 + \tfrac{\alpha}{K} \textstyle\sum_{i=1}^N \tfrac{\prod_{j=1}^{i-1} \frac{\alpha}{K}+j}{i!} (1-\mu_{(k)})^i \right)\end{aligned}$$

Finally, plugging the above into (34) and taking $K \to \infty$,

$$\begin{aligned}&p(z_{:,>k} = 0|\mu_{(k)})\\ &= \exp\big(-\alpha H_N + \alpha \textstyle\sum_{i=1}^N \tfrac{1}{i} (1-\mu_{(k)})^i \big)\end{aligned} \tag{35}$$

To obtain (25), we note that the conditional for $\mu_{(k)}$ is the posterior conditioned on both $z_{:,k} = 0$ and $z_{:,>k} = 0$. The prior given $\mu_{(k-1)}$ is (14), the probability of $z_{:,k} = 0$ is just $(1-\mu_{(k)})^N$, while the probability of $z_{:,>k} = 0$ is (35); multiplying all three gives (25).