Flexible Martingale Priors for Deep Hierarchies

Jacob Steinhardt
Massachusetts Institute of Technology

Zoubin Ghahramani
University of Cambridge

Abstract

When building priors over trees for Bayesian hierarchical models, there is a tension between maintaining desirable theoretical properties such as infinite exchangeability and important practical properties such as the ability to increase the depth of the tree to accommodate new data. We resolve this tension by presenting a family of infinitely exchangeable priors over discrete tree structures that allows the depth of the tree to grow with the data, and then showing that our family contains all hierarchical models with certain mild symmetry properties. We also show that deep hierarchical models are in general intimately tied to a process called a *martingale*, and use Doob's martingale convergence theorem to demonstrate some unexpected properties of deep hierarchies.

1 Introduction

One of the most fundamental questions we face in machine learning is what structure we should use to interpret our data. Hierarchical modeling provides one answer to this question — by modeling data at multiple levels of abstraction, we can capture broad trends over the entire data set while also taking advantage of more specific patterns that only occur over small portions of the data. A hierarchical structure over a data set can thus provide a very powerful way of sharing statistical strength over different parts of the data. However, in most cases the hierarchical structure is not known in advance and must instead be learned. There are many heuristics for finding such structure, typically by iteratively merging together subtrees that are similar under some metric (Duda et al., 2000; Heller and Ghahramani, 2005; Blundell et al., 2010). From a statistical perspective, these approaches are troublesome — there is no principled way to add new data to the tree, and it is unclear how to compare two different trees over the same data set if they have different numbers of internal nodes. Such heuristics also limit the scope of the model — for instance, it is not clear how to deal with hierarchies over latent parameters or with missing data.

The Bayesian solution to these problems is to specify a probability distribution over tree structures. In this way a hierarchical model has two components — a *prior* over the possible tree structures, including where the data lie in the tree, and a *likelihood* that specifies a distribution over latent parameters, and how those parameters affect the data. The task then is to find a suitable prior for trees. There are four general proposals for such a prior — Kingman coalescents (Kingman, 1982; Pitman, 1999; Teh et al., 2007), Dirichlet diffusion trees (Neal, 2003; Knowles and Ghahramani, 2011), tree-structured stick breaking (Adams et al., 2010), and nested Chinese restaurant processes (Blei et al., 2010).

Kingman coalescents and Dirichlet diffusion trees are both inherently continuous models, with paths either splitting or merging according to some arrival process, and the data associated with a path corresponding to the final state of a diffusion process. In addition to being infinitely exchangeable, these models have the nice property that the complexity of the implied tree structure can grow to accommodate increasing amounts of data. Unfortunately, to use these models, one needs a time-indexed stochastic process (such as a Wiener process) to underlie the data. There is thus a distinction between discrete tree structures, where any likelihood may underlie the data, and continuous structures such as Dirichlet diffusion trees, where the likelihood must correspond to some continuous process. In some important cases, such as a hierarchical beta process (Thibaux and Jordan, 2007), no underlying continuous process is known to exist.

It is therefore important to also consider inherently discrete distributions over trees. This is the approach

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

of tree-structured stick breaking (TSSB) as well as the nested Chinese restaurant process (nCRP). In both cases, the tree is fixed to be countably deep, with every node having countably many children; the interesting structure emerges in the locations of the data.

In TSSB a stick breaking procedure is used to assign a probability distribution over nodes: the root node is given a constant portion of the probability mass (drawn from a beta distribution), and the rest of the mass is partitioned among subtrees of the root using a Dirichlet process (Teh et al., 2004). The mass in each subtree is then recursively divided in the same way. Finally, data are distributed throughout the tree according to the resulting probability distribution. While this model is infinitely exchangeable, the depth of the tree is fixed by the prior — all data lie with high probability at some finite collection of depths that does not increase with the size of the data. This is an important way in which the complexity of the tree is unable to grow to accommodate the data.

Kingman coalescents, Dirichlet diffusion trees, and TSSB all separate out the prior over trees from the likelihood for the latent parameters and the data. The nCRP departs from this pattern. It associates each data point with a path down the tree; there is then an implicit tree structure based on where the different paths branch. Because the paths are infinitely long, care must be taken in choosing the likelihood to make sure that the model is well-defined. In (Blei et al., 2010), the likelihood is obtained by using a Dirichlet process (DP) to form a mixture over distributions given at each of the nodes in the path. This likelihood has the important property that the mass that the DP places on a tail of the path decays to zero; otherwise, the resulting mixture distribution would not be well-defined.

The nCRP makes the elegant decision to associate data with paths rather than nodes. By doing so, the depth of the tree can grow to accommodate new data. The nCRP is therefore the only prior over trees that is fully Bayesian, infinitely exchangeable, grows to accommodate new data, and can handle inherently discrete processes. However, these properties come at a cost. Because of the convergence issues arising from the infinite paths, it is unclear how to construct a conditional distribution for a data point given its path, except by a model similar to (Blei et al., 2010), which in many cases does not accurately represent prior beliefs about the data.

The main contribution of this paper is to give a general approach for constructing likelihoods for the nCRP or any similar path-based model. Our construction is universal for all path-based models (Theorem 2.3), and works by taking limits of latent parameters along paths down the tree, and using Doob's martingale convergence theorem (Lamb, 1973) to show that the limits exist with probability 1. We use this fact to construct a fully Bayesian hierarchical prior for both Dirichlet processes (Teh et al., 2004) and beta processes (Thibaux and Jordan, 2007). To show that inference is tractable in our model, we implement it for a hierarchical beta process (HBP).

It turns out that many existing hierarchical models already mimic our construction, except with finite rather than infinite trees. A second contribution of our paper is to use Doob's theorem to analyze the asymptotic properties of these models as the hierarchies grow deeper. This analysis yields some surprising results about HBPs and HDPs (hierarchical Dirichlet processes).

The rest of the paper is organized as follows. In Section 2, we describe our construction, introduce Doob's theorem, and use it to analyze several examples of deep hierarchical models, as well as to show that our proposed construction is both well-defined and universal. In Section 3, we derive the asymptotic depths of the nCRP and TSSB as a function of the data size and hyperparameters. Finally, in Section 4, we construct an infinitely deep HBP and show how to perform inference in this model.

2 Model Description and Properties

In this section we present a general construction for hierarchical models which associate data with paths in the tree. For concreteness, we will use the nCRP as the prior over tree structures. We start with an informal description of the elements of our model, then formally state our model and show that it is well-defined. First, though, we need a bit of notation. Given a tree $\mathcal{T}$ and a vertex $v \in \mathcal{T}$, let $p(v)$ denote the parent of v and $\mathcal{A}(v)$ denote the ancestors of v. Also, let $\operatorname{Root}(\mathcal{T})$ denote the root of $\mathcal{T}$, $\operatorname{Subtree}(v)$ denote the subtree of $\mathcal{T}$ rooted at v, and $\operatorname{Depth}(v)$ denote the depth of v (with $\operatorname{Depth}(\operatorname{Root}(\mathcal{T})) = 0$).

2.1 Model Overview

We imagine that an infinite tree $\mathcal{T}$ underlies our data. Eventually, each datum will be associated with an infinite path down the tree, and be defined in terms of a limiting process of the latent parameters. We ignore this aspect of the model for now, and merely assume that at each node v in the tree there is an associated latent parameter θ_v. Moreover, in order to even say that the tree underlies the data, we should assume that θ_v depends only on its ancestors $\mathcal{A}(v)$; more formally, we assume that $p\left(\{\theta_v\}_{v\in\mathcal{T}}\right)$ factors as $\prod_{v\in\mathcal{T}} p(\theta_v \mid \theta_{\mathcal{A}(v)})$. Note that p does not depend on v; this reflects the philosophy that the latent parameters, and not just the data itself, should satisfy an exchangeability property. If a model factors in this way, and furthermore the prior over a data point depends only on the parameters along its path, then we call it **completely exchangeable**.

By replacing θ_v with $\theta_{v\cup\mathcal{A}(v)}$ (i.e. by concatenating all the parameters on the path from $\mathrm{Root}(\mathcal{T})$ to v), we can always obtain a model where θ_v depends only on its parent. In other words, the density can be assumed to factor as

$$p\left(\{\theta_v\}_{v\in\mathcal{T}}\right) = \prod_{v\in\mathcal{T}} p(\theta_v \mid \theta_{p(v)}). \tag{1}$$

We will therefore focus on this class of models for the remainder of the discussion, keeping in mind that we lose no generality in doing so.

It is often the case in models of the form given in (1) that some key quantity $f(\theta_v)$ is preserved in expectation as we walk down the tree – more formally,

$$\mathbb{E}[f(\theta_v) \mid \theta_{p(v)}] = f(\theta_{p(v)}). \tag{2}$$

For instance, in a hierarchical Dirichlet process, θ_v is a probability distribution over the space of possible data, and $\theta_v \mid \theta_{p(v)} \sim \mathrm{DP}(c\theta_{p(v)})$ for some concentration parameter c, where $\mathrm{DP}(\mu)$ is a Dirichlet process with base measure μ. In this case, $\mathbb{E}[\theta_v \mid \theta_{p(v)}] = \theta_{p(v)}$; that is, we can take $f(\theta) = \theta$.

If f satisfies (2), then f is said to be a **martingale**. The martingale property will be important in the sequel. It turns out that data living infinitely deep in the tree will have a well-defined distribution if and only if they depend on a countable collection of L^1-bounded martingales.

2.2 Formal Description

We now formally define our model. We have a tree $\mathcal{T}$ of countable depth, such that every node v has a countable collection of children $\mathcal{C}(v)$. For each $v \in \mathcal{T}$ we have parameters θ_v (a latent parameter governing data in that subtree) and π_v (a probability distribution over $\mathcal{C}(v)$). For each datum X, we have an associated path $\{v_n(X)\}_{n=0}^{\infty}$ such that $v_0(X) = \mathrm{Root}(\mathcal{T})$ and the parent of $v_{n+1}(X)$ is $v_n(X)$. The hyperparameters of our model are a positive real number γ and conditional distributions G and H, together with a function f that is a martingale with respect to G.

The generative process for our model is as follows. For each v:

1. $\pi_v \sim \mathrm{DP}(\mathcal{C}(v), \gamma)$
2. $\theta_v \mid \theta_{p(v)} \sim G(\theta_{p(v)})$

For each X:

1. $v_0(X) = \mathrm{Root}(\mathcal{T})$
2. $v_{n+1}(X) \mid v_n(X), \pi_{v_n(X)} \sim \mathrm{Multinomial}(\pi_{v_n(X)})$
3. $X \mid \{v_n(X)\}_{n=0}^{\infty} \sim H\left(\lim_{n\to\infty} f(\theta_{v_n(X)})\right)$

Thus a datum X is obtained by first sampling a path down the tree $\mathcal{T}$ (using the distributions $\{\pi_v\}_{v\in\mathcal{T}}$ to choose which edge to follow at each point), then taking a limit of latent parameters along that path, and finally sampling X from a distribution indexed by that limit. (The skeptical reader may wonder whether the limit in the last step exists. This is established later, in Theorem 2.2.)

In the sequel, we will omit the dependence of v_n on X when it is clear from context. We will also say that $X \in \mathrm{Subtree}(v)$ if $v_n(X) = v$ for some n.

2.3 Doob's Theorem

The potential problem with the procedure specified above is that $\lim_{n\to\infty} f\left(\theta_{v_n(X)}\right)$ need not exist. This is resolved by the following theorem (Lamb, 1973):

Theorem 2.1 (Doob's martingale convergence theorem). *Let $\{\theta_n\}_{n=0}^{\infty}$ be a Markov chain over a space Θ and let $f : \Theta \to \mathbb{R}$. Suppose that $\mathbb{E}[f(\theta_{n+1}) \mid \theta_n] = f(\theta_n)$ for each n. Furthermore, suppose that $\sup_n \mathbb{E}[|f(\theta_n)|] < \infty$. Then $\lim_{n\to\infty} f(\theta_n)$ exists with probability* 1.

Before exploring the consequences of Theorem 2.1 for the model proposed in Section 2.2, we go over some examples.

Example 1: Suppose that $\theta_0 \sim \mathrm{Beta}(1,1)$ and that $\theta_{n+1} \mid \theta_n \sim \mathrm{Beta}(c\theta_n, c(1-\theta_n))$ for $n \geq 0$. If $f(\theta) = \theta$, then $\mathbb{E}[f(\theta_{n+1}) \mid \theta_n] = \mathbb{E}[\theta_{n+1} \mid \theta_n] = \mathbb{E}[\mathrm{Beta}(c\theta_n, c(1-\theta_n))] = \theta_n$.[1] Furthermore, $0 \leq \theta_n \leq 1$, so $\sup_n \mathbb{E}[|f(\theta_n)|] \leq 1 < \infty$. Consequently, $\lim_{n\to\infty} \theta_n$ exists with probability 1.

Since θ_n converges, $\sum_{n=0}^{\infty}(\theta_{n+1} - \theta_n)^2 < \infty$, hence the variance of $\theta_{n+1} \mid \theta_n$ must converge to 0 in the limit. The variance of $\mathrm{Beta}(c\theta, c(1-\theta))$ is $\frac{\theta(1-\theta)}{c+1}$, so we can therefore conclude that $\lim_{n\to\infty} \frac{\theta_n(1-\theta_n)}{c+1} = 0$, hence $\lim_{n\to\infty} \theta_n \in \{0,1\}$ with probability 1. Figure 1 illustrates this behavior.

Note that this has interesting consequences for a hierarchical beta process, since it implies that parameters

[1] We abuse notation and use $\mathrm{Beta}(\alpha, \beta)$ to refer to a random variable whose distribution is $\mathrm{Beta}(\alpha, \beta)$.

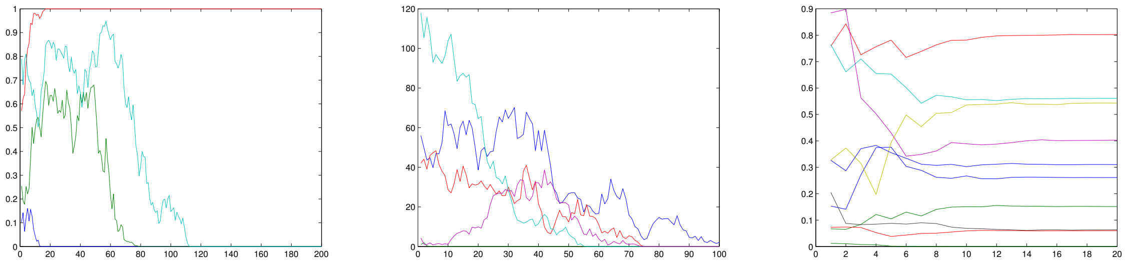

Figure 1: Examples of Doob's martingale convergence theorem in action. Left: sequences of beta random variables from Example 1, with $c = 50$. Center: sequences of gamma random variables from Example 2, with $\lambda = 50$. Right: $\frac{\alpha}{\alpha+\beta}$ from Example 3.

deep in the tree will necessarily be close to 0 or 1 with high probability. A similar story holds for hierarchical Dirichlet processes and hierarchical gamma processes (Thibaux, 2008).

Example 2: Suppose that $c_0 \sim \text{Gamma}(1, \lambda)$ and that $c_{n+1} \mid c_n \sim \text{Gamma}(c_n, 1)$. Then $\mathbb{E}[c_{n+1} \mid c_n] = c_n$. Since $c_{n+1} \geq 0$, we also have $\mathbb{E}[|c_{n+1}| \mid c_n] = c_n$. Consequently, $\sup_n \mathbb{E}[|c_n|] = \sup_n \mathbb{E}[c_0] = \lambda < \infty$. Thus $\lim_{n\to\infty} c_n$ exists with probability 1. Since the variance of $\text{Gamma}(c_n, 1)$ is c_n, we see that $\lim_{n\to\infty} c_n = 0$ with probability 1. The behavior of the sequence c_n is also illustrated in Figure 1.

Example 3: We now give an example of a martingale where f is not the identity function. Let $\alpha_0 \sim \text{Gamma}(1, 1)$, $\beta_0 \sim \text{Gamma}(1, 1)$, $d_n \mid \alpha_n \sim \text{Gamma}(\alpha_n, 1)$, $e_n \mid \beta_n \sim \text{Gamma}(\beta_n, 1)$, and $\alpha_{n+1} \mid \alpha_n \sim \alpha_n + d_n, \beta_{n+1} = \beta_n + e_n$. Note that the sequences α_n and β_n are certainly not martingales. Indeed, since $\mathbb{E}[\alpha_{n+1} \mid \alpha_n] = \alpha_n + \mathbb{E}[\text{Gamma}(\alpha_n, 1)] = 2\alpha_n$, and similarly for β_{n+1}, the sequences $\{\alpha_n\}$ and $\{\beta_n\}$ both increase exponentially in expectation. However, if we let $f(\alpha_n, \beta_n) = \frac{\alpha_n}{\alpha_n+\beta_n}$, then (see Appendix B) $\mathbb{E}[f(\alpha_{n+1}, \beta_{n+1}) \mid \alpha_n, \beta_n] = \frac{\alpha_n}{\alpha_n+\beta_n}$. Therefore, $\lim\limits_{n\to\infty} \frac{\alpha_n}{\alpha_n+\beta_n}$ exists with probability 1. This is again illustrated in Figure 1.

Example 4: Doob's theorem provides guarantees on the convergence of real-valued sequences satisfying the martingale condition. But there are many cases when we care about more than just a single real number. For instance, in a hierarchical Dirichlet process, we might care about a sequence $\{\mu_n\}_{n=0}^{\infty}$ where $\mu_{n+1} \mid \mu_n \sim \text{DP}(\mu_n)$. Fortunately, we can still use Doob's theorem; since the output of a Dirichlet process consists of countably many atoms, we only need to worry about $\mu_n(\{p\})$ for the countably many points p that are atoms of μ_1. Since $\mathbb{E}[\mu_{n+1} \mid \mu_n] = \mu_n$, we also have $\mathbb{E}[\mu_{n+1}(\{p\}) \mid \mu_n] = \mu_n(\{p\})$, hence $\lim\limits_{n\to\infty} \mu_n(\{p\})$ exists almost surely for each p. Since there are only countably many such p, we then have that $\lim\limits_{n\to\infty} \mu_n(\{p\})$ exists for all p almost surely. Also, by logic similar to example 1, each $\mu_n(\{p\})$ must converge to either 0 or 1, implying that the measure μ_n converges to a single atom in the infinite limit.[2]

Example 5: We finally go over an example of a martingale that does *not* converge. Let $x_0 = 0$ and let $x_{n+1} \mid x_n \sim \mathcal{N}(x_n, 1)$. In other words, x_{n+1} is equal to x_n perturbed by Gaussian noise with variance 1. Then $\mathbb{E}[x_{n+1} \mid x_n] = x_n$, so the sequence $\{x_n\}_{n=0}^{\infty}$ is a martingale. However, $\mathbb{E}[|x_n|] = \Theta(\sqrt{n})$, so $\sup_n \mathbb{E}[|x_n|] = \infty$. As a result, Theorem 2.1 does not apply, and indeed, the sequence $\{x_n\}$ clearly does not have a limit.

2.4 Constraints on the Likelihood

We hinted in Section 2.3 that Doob's theorem would give us conditions under which the process in Section 2.2 leads to a well-defined generative distribution. We now formalize this.

Theorem 2.2. *Let $\theta_v \mid \theta_{p(v)} \sim G(\theta_{p(v)})$, and suppose that $\mathbb{E}[f(\theta_v) \mid \theta_{p(v)}] = f(\theta_{p(v)})$. Further suppose that f is an at most countable product $\{f_k\}_{k=1}^{\infty}$ of real-valued functions, and that each f_k satisfies $\sup_n \mathbb{E}[|f_k(\theta_{v_n(X)})|] < \infty$. Then $\lim\limits_{n\to\infty} f(\theta_{v_n(X)})$ exists with probability 1.*

Proof. By Doob's theorem, $\lim_{n\to\infty} f_k(\theta_{v_n})$ exists almost surely for each k individually. Since there are only countably many f_k, and the intersection of

[2] This actually requires a bit more of an argument than before, as the μ_n could converge in distribution but not almost surely; for instance we could have $\mu_n = \delta_{p_n}$ for a countable sequence of distinct points p_n, in which case $\lim_{n\to\infty} \mu_n(\{p\})$ would be identically zero for all p, but $\lim_{n\to\infty} \mu_n$ would not converge almost surely to any probability distribution. However, we will ignore these issues for this example.

a countable collection of almost-sure events is still almost-sure, the theorem follows. □

We thus end up with two constraints on the likelihood that we need in order to use our model — the martingale condition, and the boundedness of $\mathbb{E}[|f_k(\theta)|]$. Intuitively, we can think of a martingale sequence as revealing gradually more information about a random variable until it is completely determined. From this perspective, the parameter θ_v captures information that is true across all of Subtree(v), with the parameters at descendants containing more precise information about their specific subtrees. However, Example 5 shows that this intuition is not perfect, which is why we need the boundedness condition as well.

We note that the conditions of Theorem 2.2 hold for any countable-dimensional martingale that is bounded either above or below (see Example 2 of Section 2.3). In particular, letting $f(\theta) = \theta$, they hold for hierarchical Dirichlet processes ($G(\theta) = \mathrm{DP}(c\theta)$), hierarchical beta processes ($G(\theta) = \mathrm{BP}(\theta, c)$), and hierarchical gamma processes ($G(\theta) = \mathrm{GammaP}(\theta)$), since these are all non-negative and depend on only a countable collection of atoms.

We next give a converse to Theorem 2.2, proved in Appendix A, showing that the martingale and boundedness conditions are both necessary, and thus the construction in Subsection 2.2 is universal for completely exchangeable models.

Theorem 2.3. *Consider any completely exchangeable model where the data lie in a Polish space $\mathbb{X}$. Then there exists latent parameters $\theta_v \in \Theta$, a function $f : \Theta \to [0,1]^{\mathbb{N}}$, and distributions G and H such that $\theta_v \mid \theta_{p(v)} \sim G(\theta_v)$, $\mathbb{E}[f(\theta_v) \mid \theta_{p(v)}] = f(\theta_{p(v)})$, and $X \mid \{v_n(X), \theta_{v_n(X)}\}_{n=0}^{\infty} \sim H\left(\lim_{n\to\infty} f(\theta_{v_n(X)})\right)$.*

A Polish space is a completely metrizable separable space. This is the most generable space for which a suitable notion of conditional probability exists. Therefore, all spaces of interest in statistics are Polish.

Consider the hierarchical latent Dirichlet allocation model of (Blei et al., 2010), where each node v in the nCRP has a distribution μ_v over words, there is a global distribution π over levels of the tree, and each word in a document X with path $\{v_n(X)\}_{n=0}^{\infty}$ is drawn from the mixture distribution $\sum_{n=0}^{\infty} \pi_n \mu_{v_n(X)}$. Furthermore, $\mu_v \mid \mu_{p(v)} \sim \mathrm{DP}(c\theta_v)$. We can recover this model with our construction by having parameters μ_v and θ_v at each node, where μ_v is as defined above and the conditional distribution for θ_v at depth d is deterministic and given by

$$\theta_v \mid \theta_{p(v)}, \mu_{p(v)}, \pi = \frac{(\sum_{i=0}^{d-1} \pi_i)\theta_{p(v)} + \pi_l \phi_v}{\sum_{i=0}^{d} \pi_i}.$$

The distribution for a word in X is then given by the limiting value of θ_v.

3 Depth of the nCRP and TSSB

The key property of an nCRP that makes it desirable over tree-structured stick breaking is the depth of the resulting tree. Note that in an nCRP, every datum is associated with an infinite path, and thus lies infinitely deep in the tree. However, we can talk about the *effective depth* of a data point as the smallest depth at which that point is the unique datum in its subtree.

Proposition 3.1. *The effective depth of a data point under* nCRP(γ) *is* $\Theta\left(\frac{\log(N)}{\xi+\psi(1+\gamma)}\right)$ *with high probability, where $\xi = 0.5772\ldots$ is the Euler-Mascheroni constant and ψ is the digamma function.*

To prove Proposition 3.1, we first need a basic lemma about Dirichlet processes:

Lemma 3.2. *The posterior distribution of $\pi_{v_n}(v_{n+1}) \mid X \in$* Subtree$(v_{n+1})$ *is equal to* Beta$(1,\gamma)$*. In other words, the mass assigned to a child conditioned on a single datum having already been assigned to that child is distributed as* Beta$(1,\gamma)$*.*

Proof. Note that $DP(\gamma)$ can be obtained by drawing a sample from $DP(\gamma U)$, where U is uniform on $[0,1]$, and assigning the masses of the atoms to the children of v. Therefore, if we let $\mu \sim \mathrm{DP}(\gamma U)$ and $q \sim \mathrm{Multinomial}(\mu)$, then the posterior distribution of $\pi_{v_n}(v_{n+1}) \mid X \in \mathrm{Subtree}(v_{n+1})$ is equivalent to the distribution of $\mu(\{p\}) \mid q = p$. By conjugacy, $\mu \mid q = p \sim \mathrm{DP}(\delta_p + \gamma U)$. Then, by the defining property of a Dirichlet process, $(\mu(\{p\}), \mu([0,1]\backslash\{p\})) \sim \mathrm{Dirichlet}(1,\gamma)$, hence $\mu(\{p\}) \sim \mathrm{Beta}(1,\gamma)$. □

Now we are ready to prove Proposition 3.1.

Proof of Proposition 3.1. Let X be a data point. The probability that $\mathrm{Depth}(X) \leq d$ is the probability that none of the other $N-1$ data points lie in Subtree$(v_d(X))$, which is equal to $\left(1 - \prod_{i=0}^{d-1} \pi_{v_i}(v_{i+1}(X))\right)^{N-1}$. But $\prod_{i=0}^{d-1} \pi_{v_i}(v_{i+1}) = e^{\sum_{i=0}^{d-1} \log \pi_{v_i}(v_{i+1})}$. The $\log \pi_{v_i}(v_{i+1})$ are independent, and by Lemma 3.2 they are $\log \mathrm{Beta}(1,\gamma)$-distributed. Since $\log \mathrm{Beta}(1,\gamma)$ has finite variance, it follows by Chebyshev's inequality that $\sum_{i=0}^{d-1} \log \pi_{v_i}(v_{i+1}) = d\mathbb{E}[\log \mathrm{Beta}(1,\gamma)] + O(\sqrt{d})$ with high probability. One

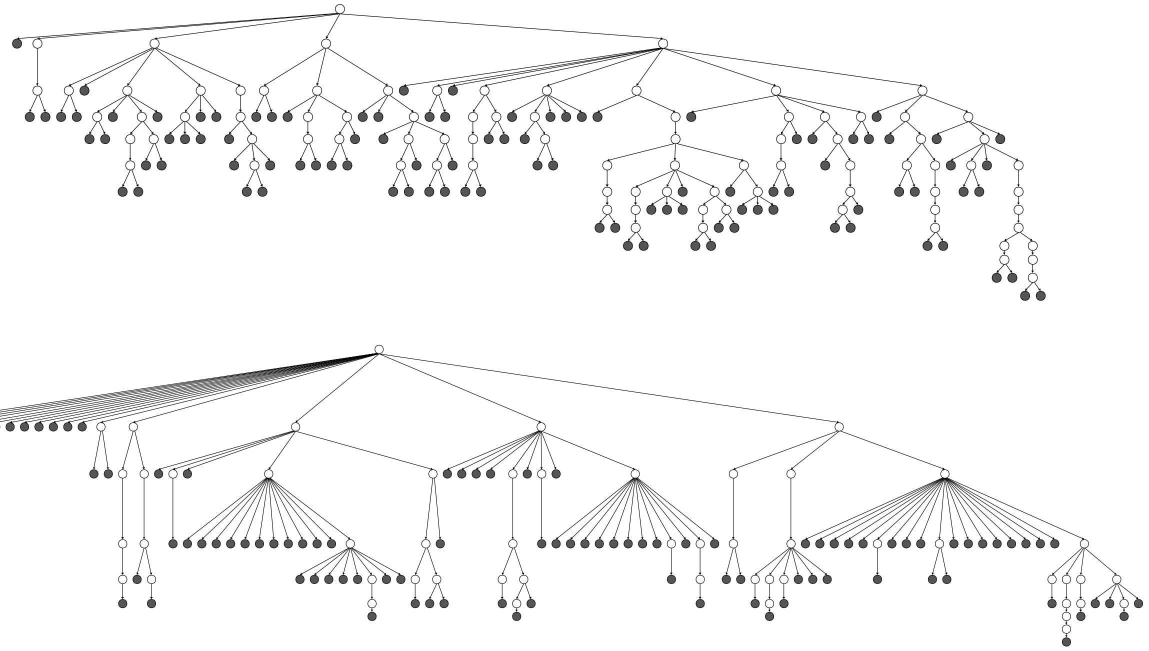

Figure 2: Trees drawn from the prior of the nCRP (top) and TSSB (bottom) models with $N = 100$ data points. In both cases we used a hyper-parameter of $\gamma = 1$. For TSSB, we further set $\alpha = 10$ and $\lambda = \frac{1}{2}$ (these are parameters that do not exist in the nCRP). Note that the tree generated by TSSB is very wide and shallow. A larger value of α would fix this for $N = 100$, but increasing N would cause the problem to re-appear.

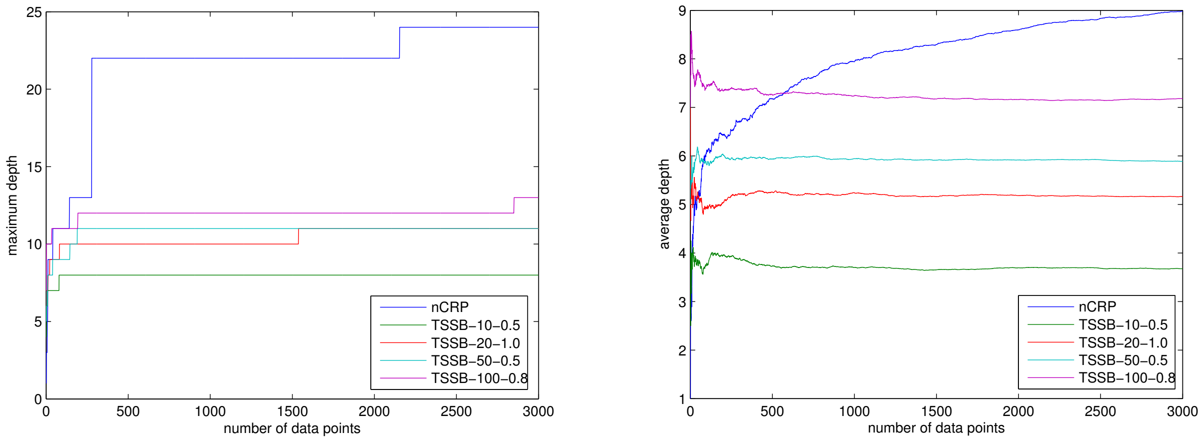

Figure 3: Tree depth versus number of data points. We drew a single tree from the prior for the nCRP as well as for tree-structured stick-breaking, and computed both the maximum and average depth as more data was added to the tree. The above plots show that the depth of the nCRP increases with the amount of data, whereas the depth of TSSB quickly converges to a constant. The different curves for the TSSB model correspond to different settings of the hyperparameters α and λ.

can show (see Appendix B) that $\mathbb{E}[\log \mathrm{Beta}(1,\gamma)] = \psi(1) - \psi(1+\gamma) = -\xi - \psi(1+\gamma)$. Then

$$\mathbb{P}[\mathrm{Depth}(X) \leq d] = \left(1 - e^{-d(\xi+\psi(1+\gamma))+O(\sqrt{d})}\right)^{N-1}.$$

If we let $d = \alpha\frac{\log(N)}{\xi+\psi(1+\gamma)}$, then the right-hand-side above becomes $\left(1 - N^{-\alpha+O\left(\frac{1}{\sqrt{\log(N)}}\right)}\right)^{N-1}$, which decays quickly from 1 to 0 as α passes the threshold value of 1. It follows that $\alpha = \Theta(1)$ with high probability, so $d = \Theta\left(\frac{\log(N)}{\xi+\psi(1+\gamma)}\right)$ with high probability, which completes the proposition. □

In contrast to Proposition 3.1, the depth distribution of a datum is constant in the TSSB model. As a review of TSSB, the probability that a path stops at a node v at depth d is drawn from $\mathrm{Beta}(1, \alpha\lambda^d)$, and otherwise the path continues to one of the children of v based on a $\mathrm{DP}(\gamma)$ draw. If a path stops at v, then it creates a new leaf as a child of v at which to place the data point. Because the stopping criterion for a path is independent of the other data points, the depth of a path under the prior is independent of the amount of data, and the following proposition gives its dependence on the hyperparameters:

Proposition 3.3. *The depth of a data point under* $\mathrm{TSSB}(\gamma,\alpha,\lambda)$ *is equal to* $\Theta\left(\min\left(\frac{\log(\alpha)}{\log(1/\lambda)}, \frac{\log(1+\alpha(\lambda^{-1}-1))}{\log(1/\lambda)}\right)\right)$ *with high probability.*

Note that the former term dominates except when λ is close to 1.

Proof. Since the probability of a path stopping at any given depth j is $\mathrm{Beta}(1,\alpha\lambda^j)$-distributed, the probability that a node lies at a depth of at least d is equal to $\prod_{j=0}^{d-1}\left(1 - \frac{1}{1+\alpha\lambda^j}\right)$. It suffices to show that d is unlikely to be much greater than $\log(\alpha)/\log(1/\lambda)$, and that if d is less than $\log(\alpha)/\log(1/\lambda)$ then it is close to $\log(1+\alpha(\lambda^{-1}-1))/\log(1/\lambda)$ with high probability. So, first suppose that $d \gg \log(\alpha)/\log(1/\lambda)$. Then a constant fraction of the terms in the product are less than $\frac{1}{2}$, and thus the product is exponentially small. Now suppose instead that $d < \log(\alpha)/\log(1/\lambda)$. Then $1 - \frac{1}{1+\alpha\lambda^j}$ is within a constant factor of $e^{-\frac{1}{\alpha\lambda^j}}$. Thus the product is equal to $e^{-\Theta\left(\frac{1}{\alpha}\sum_{j=0}^{d}\lambda^{-j}\right)}$, or $e^{-\Theta\left(\frac{1}{\alpha}\frac{\lambda^{-d}-1}{\lambda^{-1}-1}\right)}$. We thus want to find the range when $\frac{1}{\alpha}\frac{\lambda^{-d}-1}{\lambda^{-1}-1}$ is equal to $\Theta(1)$, which is when $d = \Theta\left(\frac{\log(1+\alpha(\lambda^{-1}-1))}{\log(1/\lambda)}\right)$. □

The constant depth of the TSSB leads to overly wide and shallow trees. This is illustrated in Figures 2 and 3, where we show samples from the prior over tree structures for both the nCRP and TSSB.

The TSSB model uses two extra hyperparameters (α and λ) that do not occur in the nCRP. By setting α to N and λ to $e^{-\xi-\psi(1+\gamma)}$, it is possible to approximate the depth distribution of the nCRP with tree-structured stick breaking. However, the models are still qualitatively different. While TSSB can mimic the marginal depth distribution as measured from the root of the tree, it cannot mimic the depth distribution as measured from an arbitrary subtree. Separately, setting α to N makes the prior data-dependent, which is problematic in itself.

4 Implementation for a Hierarchical Beta Process

We now show how our construction applies in the case of a hierarchical beta process (Thibaux and Jordan, 2007). Recall that an HBP is a model that generates an exchangeable sequence $\{X_n\}_{n=1}^{\infty}$, where each X_n is a finite collection of binary features. Typically a beta process is used when the feature set is not known a priori and is potentially infinite (Griffiths and Ghahramani, 2011). For simplicity, however, we will assume that the feature set is both finite and known in advance, so that each X_n can be represented as a binary vector of some length L. We place a tree structure over the X_n using an nCRP prior. For each internal node v, we have a latent parameter $\theta_v \in [0,1]^L$, and our likelihood is given by:

1. $\theta_{\mathrm{Root}(\mathcal{T}),l} = 0.5$ for all $l \in \{1,\ldots,L\}$
2. $\theta_{v,l} \mid \theta_{p(v),l} \sim \mathrm{Beta}(c\theta_{p(v),l}, c(1-\theta_{p(v),l}))$
3. $X_l \mid \{v_n(X), \theta_{v_n(X)}\}_{n=0}^{\infty} = \lim_{n\to\infty} \theta_{v_n(X),l}$

As shown in Example 1, $\lim_{n\to\infty}\theta_{v_n(X),l}$ lies in $\{0,1\}$ almost surely, so X lies in $\{0,1\}^L$.

Due to space constraints, we cannot give a full account of how to perform inference in this model. Our goal in the remainder of this section will be to give a high-level overview, showing in particular how to tractably deal with the infinitely long paths created by the nCRP. A more detailed description of inference is given in Appendices C and D.

Representing the tree The first issue is how to represent the tree. The prior specifies infinitely long paths for each datum, which is problematic for computation. We deal with this using Lemma 4.1, which implies that if a subtree contains only a single datum, then we can analytically marginalize out *all* of the parameters of that subtree:

Lemma 4.1. *The marginal distribution of* $X \mid (X \in \mathrm{Subtree}(v), \theta_{p(v)})$ *is equal to* $\mathrm{Bernoulli}(\theta_{p(v)})$. *Furthermore,* $X \mid (X \in \mathrm{Subtree}(v), \theta_{p(v)})$ *is independent of* Y *for any* $Y \notin \mathrm{Subtree}(v)$.

We thus represent $\mathcal{T}$ by a truncated tree $\mathcal{T}'$ as follows: each internal node v of $\mathcal{T}'$ corresponds to a node of $\mathcal{T}$ with a non-empty subtree. Each leaf w of $\mathcal{T}'$ corresponds to a data point X, which implicitly represents an entire subtree of $\mathcal{T}$ that has been marginalized out using Lemma 4.1. Subtrees with no data are omitted altogether in $\mathcal{T}'$. As more data is added to a tree, a new datum Y might end up taking a path through X. In this case, X is replaced with an internal node that then branches into new leafs containing X and Y (if the paths of X and Y share many vertices, then many new internal nodes will be created).

Incremental Gibbs Sampling Our specific approach to inference is incremental Gibbs sampling, although other MCMC variants could be used as well. There are three types of MCMC moves that we consider: adding a data point, removing a data point, and resampling the latent parameters. We outline each below.

Adding a data point We can add a new data point Y to $\mathcal{T}'$ either by making it the child of an already existing internal node v, or by expanding an external node w. It is straightforward to calculate the likelihood in the first case, as it is the probability that a datum would take the path to v under the nCRP prior, times the probability of creating a new table under the CRP at node v, times the probability of generating Y from $\mathrm{Subtree}(v)$ (which is given by Lemma 4.1). Expanding an external node is more complicated, as we need to create new internal nodes and sample their parameters conditioned on X and Y. We also need to compute the conditional distribution over how deep the paths of X and Y first branch. Both of these calculations can be made, and are given in Appendix C.

Removing a data point We remove the data point and delete any nodes that now have zero data points in their subtree. It is also now possible that an internal node could have a single datum as its child and nothing else, in which case that node should be collapsed.

Resampling the parameters An algorithm for resampling the latent parameters of an HBP was first proposed in (Thibaux and Jordan, 2007). Unfortunately, this algorithm is not suited to sampling deep hierarchies due to general numerical issues with hierarchical Beta processes. The numerical issues occur when we are resampling the parameters of a node and one of the values of the children is very close to 0 or 1. If a child parameter is very close to 0, for instance, it actually matters for the likelihood whether the parameter is equal to 10^{-10} or 10^{-50} (or even 10^{-1000}). Since we cannot actually distinguish between these numbers with floating point arithmetic, this introduces innacuracies in the posterior that push all of the parameters closer to 0.5. To deal with this problem, we assume that we cannot distinguish between numbers that are less than some distance ϵ from 0 or 1. If we see such a number, we treat it as having a censored value (so it appears as $\mathbb{P}[\theta < \epsilon]$ or $\mathbb{P}[\theta > 1 - \epsilon]$ in the likelihood). We then obtain a log-concave conditional density, for which efficient sampling algorithms exist (Leydold, 2003).

Scalability If there are N data points, each with L features, and the tree has depth D, then the time it takes to add a data point is $O(NL)$, the time it takes to remove a data point is $O(D)$, and the time it takes to resample a single set of parameters is (amortized) $O(L)$. The dominating operation is adding a node, so to make a Gibbs update for all data points will take total time $O(N^2L)$.

Implementation To demonstrate inference in our model, we created a data set of 53 stick figures determined by the presence or absence of a set of 29 lines, which we treated as binary vectors in $\{0,1\}^{29}$. We then ran incremental Gibbs sampling for 100 iterations with hyperparameters of $\gamma = 1.0$, $c = 20.0$. The output of the final sample is given in the supplementary material.

5 Conclusion

We have presented an exchangeable prior over discrete hierarchies that can flexibly increase its depth to accommodate new data, and shown that our prior is universal for completely exchangeable models. We have also implemented this prior for a hierarchical beta process. Along the way, we identified a common model property — the martingale property — that has interesting and unexpected consequences in deep hierarchies.

This paper has focused on a general theoretical characterization of infinitely exchangeable distributions over trees based on the Doob martingale convergence theorem, on elucidating properties of deep hierarchical beta processes as an example of such models, and on defining an efficient inference algorithm for such models, which was demonstrated on a small binary data set. A full experimental evaluation of nonparametric Bayesian models for hierarchies is outside the scope of this paper but clearly of interest.

References

Ryan P. Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. *Advances in Neural Information Processing Systems*, 23, 2010.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2), Jan 2010.

Charles Blundell, Yee Whye Teh, and Katherine A. Heller. Bayesian rose trees. *Uncertainty in Artificial Intelligence*, 2010.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley Interscience, 2nd edition, 2000.

Thomas L. Griffiths and Zoubin Ghahramani. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, Apr 2011.

Katherine Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. *International Conference on Machine Learning*, 22, 2005.

John Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

David Knowles and Zoubin Ghahramani. Pitman-Yor diffusion trees. *Uncertainty in Artificial Intelligence*, 27, 2011.

Charles W. Lamb. A short proof of the martingale convergence theorem. *Proceedings of the American Mathematical Society*, 38(1), Mar 1973.

Josef Leydold. Short universal generators via generalized ratio-of-uniforms method. *Mathematics of Computation*, 72(243):1453–1471, Mar 2003.

Radford M. Neal. Density modeling and clustering using dirichlet diffusion trees. In *Bayesian Statistics 7*, pages 619–629, 2003.

Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, 1999.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. Technical Report 653, 2004.

Yee Whye Teh, Hal Daume III, and Dan M. Roy. Bayesian agglomerative clustering with coalescents. *Advances in Neural Information Processing Systems*, 2007.

Romain Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, University of California, Berkeley, 2008.

Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the Indian buffet process. *AISTATS*, 2007.

A Converse to Doob's Theorem

Theorem 2.3. *Consider any completely exchangeable model where the data lie in a Polish space $\mathbb{X}$. Then there exists latent parameters $\theta_v \in \Theta$, a function $f : \Theta \to [0,1]^{\mathbb{N}}$, and distributions G and H such that $\theta_v \mid \theta_{p(v)} \sim G(\theta_v)$, $\mathbb{E}[f(\theta_v) \mid \theta_{p(v)}] = f(\theta_{p(v)})$, and $X \mid \{v_n(X), \theta_{v_n(X)}\}_{n=0}^{\infty} \sim H\left(\lim_{n\to\infty} f(\theta_{v_n(X)})\right)$.*

Recall that a Polish space is a completely metrizable separable space.

Proof of Theorem 2.3. Our strategy will be to find a countable collection of bounded statistics that uniquely determine any probability distribution over $\mathbb{X}$, then augment the original latent variables at each node v with this collection. We will then show that these statistics form a martingale, and that their limit determines the conditional distribution of X given the latent parameters on its path.

First, we show that there exist a countable collection $\mathcal{C}$ of measurable subsets S of $\mathbb{X}$ such that knowing $\mathbb{P}_p[X \in S]$ for all $S \in \mathcal{C}$ completely determines any probability distribution p over $\mathbb{X}$. Indeed, if $\mathbb{X}$ is Polish then the space $\mathbb{D}$ of probability measures on $\mathbb{X}$ is also Polish in the topology generated by sets of the form $U_{S,a,b} := \{p \mid a < \mathbb{P}_p[X \in S] < b\}$. In particular, since $\mathbb{D}$ is a separable metric space, it is second-countable. Let B be any countable base, and note that every member U_0 of B is second countable and hence Lindelöf, so that we can find a countable collection of the $U_{S,a,b}$ that exactly covers U_0. Unioning over all the U_0 in B gives us a countable basis B' consisting of sets of the form $U_{S,a,b}$. We then claim that $\mathcal{C} := \{S \mid U_{S,a,b} \in B'\}$ is the desired collection of measurable sets. Indeed, suppose that p and q are two distributions in $\mathbb{D}$. Since $\mathbb{D}$ is Hausdorff, there exists some $U_{S,a,b} \in B'$ such that $p \in U_{S,a,b}$ and $q \notin U_{S,a,b}$, which in particular implies that $\mathbb{P}_p[X \in S] \neq \mathbb{P}_q[X \in S]$. Taking the converse, if $\mathbb{P}_p[X \in S] = \mathbb{P}_q[X \in S]$ for all $S \in \mathcal{C}$, then $p = q$, and hence knowing $\mathbb{P}_p[X \in S]$ for all $S \in \mathcal{C}$ completely determines p.

Now let ϕ_v be equal to the countable tuple $(\mathbb{P}[X \in S \mid X \in \text{Subtree}(v)])_{S \in \mathcal{C}}$, and let ψ_v be the original latent parameter at v in $\mathcal{T}$. By the Markov property, ψ_v determines ϕ_v, so if we let $\theta_v = (\phi_v, \psi_v)$, then θ_v is statistically equivalent to the original latent parameter ψ_v. Since by assumption there exists a fixed conditional distribution G_0 for $\psi_v \mid \psi_{p(v)}$, there also exists a fixed conditional distribution G for $\theta_v \mid \theta_{p(v)}$. On the other hand, if we let $f(\theta_v) = \phi_v$, then f is clearly bounded (since all its coordinates are probabilities and thus lie in $[0,1]$), and is a martingale since $\mathbb{E}[\mathbb{P}[X \in S \mid X \in \text{Subtree}(v)] \mid \theta_{p(v)}] = \mathbb{P}[X \in S \mid X \in \text{Subtree}(p(v))]$. Finally, let $H(\theta_v)$ be the unique distribution defined by ϕ_v. To finish the proof, we need to show that $H\left(\lim_{n\to\infty} \theta_{v_n(X)}\right)$ is the distribution of $X \mid \{v_n(X), \psi_{v_n(X)}\}_{n=0}^{\infty}$. In other words, we need to show that $\mathbb{P}[X \in S \mid \{v_n(X), \psi_{v_n(X)}\}]$ is equal to $\lim_{n\to\infty} \mathbb{P}[X \in S \mid v_n(X), \theta_{v_n(X)}]$ for all $S \in \mathcal{C}$. This follows directly from Levy's zero-one law, which states that if F_∞ is the minimal σ-algebra generated by a filtration $F_0, F_1, \ldots$ of a probability space, then $\lim_{k\to\infty} \mathbb{E}[Z \mid F_k] = \mathbb{E}[Z \mid F_\infty]$ almost surely for any random variable Z (in our case Z is the indicator for the event that $X \in S$). So the θ_v are indeed the desired set of latent variables, and the proof is complete. □

B Statistics of Beta and Gamma Functions

Lemma B.1. *Let $d_n \sim \text{Gamma}(\alpha_n, 1)$, $e_n \sim \text{Gamma}(\beta_n, 1)$, $\alpha_{n+1} = \alpha_n + d_n$, and $\beta_{n+1} = \beta_n + e_n$. Then $\mathbb{E}\left[\frac{\alpha_{n+1}}{\alpha_{n+1}+\beta_{n+1}}\right] = \frac{\alpha_n}{\alpha_n+\beta_n}$.*

Proof. We first note that if d and e are independent and distributed as $\text{Gamma}(\alpha, 1)$ and $\text{Gamma}(\beta, 1)$, then the conditional distribution of d given that $d+e = s$ is equal to $s\,\text{Beta}(\alpha, \beta)$ (the proof is a straightforward calculation of probability densities). Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\alpha_{n+1}}{\alpha_{n+1}+\beta_{n+1}}\right] \\
&= \mathbb{E}_{d_n,e_n}\left[\frac{\alpha_n + d_n}{\alpha_n + \beta_n + d_n + e_n}\right] \\
&= \mathbb{E}_s\left[\mathbb{E}_{d_n}\left[\frac{\alpha_n + d_n}{\alpha_n + \beta_n + s} \mid d_n + e_n = s\right]\right] \\
&= \mathbb{E}_s\left[\mathbb{E}_{d_n}\left[\frac{\alpha_n + d_n}{\alpha_n + \beta_n + s} \mid d_n \sim s\,\text{Beta}(\alpha_n, \beta_n)\right]\right] \\
&= \mathbb{E}_s\left[\frac{\alpha_n + s\frac{\alpha_n}{\alpha_n+\beta_n}}{\alpha_n + \beta_n + s}\right] \\
&= \mathbb{E}_s\left[\frac{\alpha_n}{\alpha_n + \beta_n}\right] \\
&= \frac{\alpha_n}{\alpha_n + \beta_n}.
\end{aligned}
$$

□

Lemma B.2. *If $X \sim \text{Beta}(\alpha, \beta)$, then $\mathbb{E}[\log(X)] = \psi(\alpha) - \psi(\alpha + \beta)$, where ψ is the digamma function defined by $\psi(x) = \frac{d}{dx}\log\text{Gamma}(x)$.*

Proof. Let $F(\alpha) = \int_\infty^\alpha \left(\int_0^1 x^{\tilde{\alpha}-1}(1-x)^{\beta-1}\log(x)dx\right) d\tilde{\alpha}$. Then by the fundamental theorem of calculus, $\frac{dF}{d\alpha} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\log(x)dx = \text{Beta}(\alpha, \beta)\mathbb{E}[\log(X)]$. We claim that $F(\alpha) = \text{Beta}(\alpha, \beta)$. Indeed, we have

$$
\begin{aligned}
F(\alpha) &= \int_{\infty}^{\alpha} \int_{0}^{1} x^{\tilde{\alpha}-1}(1-x)^{\beta-1}\log(x)dxd\tilde{\alpha} \\
&= \int_{0}^{1} (1-x)^{\beta-1} \int_{\infty}^{\alpha} x^{\tilde{\alpha}-1}\log(x)d\tilde{\alpha}dx \\
&= \int_{0}^{1} (1-x)^{\beta-1} \left(x^{\tilde{\alpha}-1}\big|_{\infty}^{\alpha} \right) dx \\
&= \int_{0}^{1} (1-x)^{\beta-1} x^{\alpha-1} \\
&= \mathrm{Beta}(\alpha,\beta)
\end{aligned}
$$

Then it follows that

$$
\begin{aligned}
\mathbb{E}[\log(X)] &= \frac{\frac{d}{d\alpha}\mathrm{Beta}(\alpha,\beta)}{\mathrm{Beta}(\alpha,\beta)} \\
&= \frac{d}{d\alpha}\log\mathrm{Beta}(\alpha,\beta) \\
&= \frac{d}{d\alpha}\left(\log\mathrm{Gamma}(\alpha) - \log\mathrm{Gamma}(\alpha+\beta)\right) \\
&= \psi(\alpha) - \psi(\alpha+\beta),
\end{aligned}
$$

which proves the lemma. □

C Properties of Hierarchical Beta Processes

In this section, we prove Lemma 4.1, and make some additional calculations regarding the hierarchical beta process model that will be useful for inference. We deal with inference itself in the next section. We let X denote a data point, X_l denote the lth coordinate of X, and θ_n denote the parameter at the node at depth n in the path corresponding to X. We also let $\theta_{n,l}$ denote the lth coordinate of θ_n.

Lemma 4.1. *The marginal distribution of* $X \mid (X \in \mathrm{Subtree}(v), \theta_{p(v)})$ *is equal to* $\mathrm{Bernoulli}(\theta_{p(v)})$. *Furthermore,* $X \mid (X \in \mathrm{Subtree}(v), \theta_{p(v)})$ *is independent of* Y *for any* $Y \notin \mathrm{Subtree}(v)$.

Proof of Lemma 4.1. Since $X_l \in \{0,1\}$, we have $\mathbb{P}[X_l = 1 \mid \theta_{p(v)}] = \mathbb{E}[X_l \mid \theta_{p(v)}]$, hence $X_l \mid \theta_{p(v)} \sim \mathrm{Bernoulli}(\mathbb{E}[X_l \mid \theta_{p(v)}])$. But

$$
\begin{aligned}
\mathbb{E}[X_l \mid \theta_{p(v)}] &= \mathbb{E}\left[\mathrm{Bernoulli}\left(\lim_{n\to\infty}\theta_{n,l}(X)\right) \mid \theta_{p(v)}\right] \\
&= \mathrm{Bernoulli}\left(\mathbb{E}\left[\lim_{n\to\infty}\theta_{n,l}(X) \mid \theta_{p(v)}\right]\right) \\
&= \mathrm{Bernoulli}(\theta_{p(v),l}),
\end{aligned}
$$

where the last step uses the martingale property.[3] This proves that $X \mid (X \in \mathrm{Subtree}(v), \theta_{p(v)})$ is $\mathrm{Bernoulli}(\theta_{p(v)})$-distributed. The conditional independence property then follows from the fact that the joint distribution satisfies the Markov property for the tree $\mathcal{T}$. □

Our next lemma is useful for determining the probability that a new datum Y would be generated given that it lies in the subtree corresponding to an existing datum X.

Lemma C.1. *For any depth* $n \geq 0$, *and any* $m \geq n$, *we have*

$$
\mathbb{E}[\theta_{m,l} \mid \theta_n, X] = \left\{ \begin{array}{lcl} \left(\frac{c}{c+1}\right)^{m-n}\theta_{n,l} & : & X_l = 0 \\ 1 - \left(\frac{c}{c+1}\right)^{m-n}(1-\theta_{n,l}) & : & X_l = 1 \end{array} \right.
$$

Furthermore, if Y *is another datum and the least common ancestor of* X *and* Y *is at a depth* $d \geq n$, *then*

$$
\mathbb{P}[Y_l = 1 \mid \theta_n, X] = \left\{ \begin{array}{lcl} \left(\frac{c}{c+1}\right)^{d-n}\theta_{n,l} & : & X_l = 0 \\ 1 - \left(\frac{c}{c+1}\right)^{d-n}(1-\theta_{n,l}) & : & X_l = 1 \end{array} \right.
$$

Proof of Lemma C.1. By Lemma 4.1, $\mathbb{P}[X_l = 1 \mid \theta_i] = \theta_{i,l}$ for any i. Then, by the conjugacy of the Beta distribution, $\theta_{i+1,l} \mid \theta_i, X \sim \mathrm{Beta}(c\theta_{i,l} + 1 - X_l, c(1-\theta_{i,l}) + X_l)$. It follows that

$$
\mathbb{E}[\theta_{i+1,l} \mid \theta_i, X] = \left\{ \begin{array}{lcl} \left(\frac{c}{c+1}\right)\theta_{i,l} & : & X_l = 0 \\ 1 - \left(\frac{c}{c+1}\right)(1-\theta_{i,l}) & : & X_l = 1 \end{array} \right.
$$

Iteratively applying this relation yields the first part of the lemma. The second part of the lemma then follows by applying Lemma 4.1 to see that

$$
\begin{aligned}
\mathbb{P}[Y_l = 1 \mid \theta_n, X] &= \mathbb{E}[\mathbb{P}[Y_l = 1 \mid \theta_{d,l}] \mid \theta_n, X] \\
&= \mathbb{E}[\theta_{d,l} \mid \theta_n, X]
\end{aligned}
$$

and then applying the first part of the lemma. □

Lemma C.2. *As in Lemma C.1, let* d *be the depth of the least common ancestor of* X *and* Y. *Then, for any*

[3] In fact, we need something stronger, since the expectation of a limit does not necessarily equal the limit of the expectation, as can be seen in Example 2 of Section 2.3. However, if the random variables involved are uniformly integrable, then a stronger version of Theorem 2.1 implies that the limit of the expectation is indeed equal to the expectation of the limit. Since the $\theta_{n,l}$ are bounded, they are uniformly integrable.

$n < d$, we have the following relations:

$$
\begin{aligned}
&\theta_{n+1,l} \mid (\theta_n, X_l \neq Y_l) \sim \\
&\qquad \mathrm{Beta}(c\theta_{n,l} + 1, c(1-\theta_{n,l}) + 1) \\
&\theta_{n+1,l} \mid (\theta_n, X_l = Y_l = 0) \sim \\
&\qquad \frac{\omega_1}{\omega_1 + \omega_2} \mathrm{Beta}(c\theta_{n,l} + 2, c(1-\theta_{n,l})) \\
&\quad + \quad \frac{\omega_2}{\omega_1 + \omega_2} \mathrm{Beta}(c\theta_{n,l} + 1, c(1-\theta_{n,l}) + 1) \\
&\theta_{n+1,l} \mid (\theta_n, X_l = Y_l = 1) \sim \\
&\qquad \frac{\omega_3}{\omega_3 + \omega_4} \mathrm{Beta}(c\theta_{n,l}, c(1-\theta_{n,l}) + 2) \\
&\quad + \quad \frac{\omega_4}{\omega_3 + \omega_4} \mathrm{Beta}(c\theta_{n,l} + 1, c(1-\theta_{n,l}) + 1),
\end{aligned}
$$

where

$$
\begin{aligned}
\omega_1 &= c(1-\theta_{n,l}) + 1 \\
\omega_2 &= c\theta_{n,l}\left(1 - \left(\frac{c}{c+1}\right)^{d-n-1}\right) \\
\omega_3 &= c\theta_{n,l} + 1 \\
\omega_4 &= c(1-\theta_{n,l})\left(1 - \left(\frac{c}{c+1}\right)^{d-n-1}\right).
\end{aligned}
$$

Proof of Lemma C.2. We will prove the assertion when $X_l = 0$, since the argument when $X_l = 1$ is identical. For brevity, we will drop the subscript of l on θ, X, and Y. Also, we let $r := \left(\frac{c}{c+1}\right)^{d-n-1}$. Then by Bayes' rule, we have:

$$
\begin{aligned}
&p(\theta_{n+1} \mid \theta_n, X = 0, Y = 1) \\
&\ \propto p(Y = 1 \mid \theta_{n+1}, X = 0) p(X = 0 \mid \theta_{n+1}) p(\theta_{n+1} \mid \theta_n) \\
&\ \propto r\theta_{n+1} \times (1 - \theta_{n+1}) \times \mathrm{Beta}(\theta_{n+1}; c\theta_n, c(1-\theta_n)) \\
&\ \propto \mathrm{Beta}(\theta_n; c\theta_n + 1, c(1-\theta_n) + 1).
\end{aligned}
$$

Here we applied Lemma C.1 to compute $p(Y = 1 \mid \theta_{n+1}, X = 0)$, and we applied Lemma 4.1 to compute $p(X = 0 \mid \theta_{n+1})$.

We now turn to the case when $Y = 0$. Then, using Lemmas 4.1 and C.1 in the same way, we have

$$
\begin{aligned}
&p(\theta_{n+1} \mid \theta_n, X = 0, Y = 0) \\
&\ \propto p(Y = 0 \mid \theta_{n+1}, X = 0) p(X = 0 \mid \theta_{n+1}) p(\theta_{n+1} \mid \theta_n) \\
&\ \propto [1 - r\theta_{n+1}] \times (1 - \theta_{n+1}) \\
&\qquad \times \mathrm{Beta}(\theta_{n+1}; c\theta_n, c(1-\theta_n)) \\
&\ \propto [1 - r\theta_{n+1}] \\
&\qquad \times \mathrm{Beta}(\theta_{n+1}; c\theta_n, c(1-\theta_n) + 1) \\
&\ \propto [(1 - \theta_{n+1}) + (1 - r)\,\theta_{n+1}] \\
&\qquad \times \mathrm{Beta}(\theta_{n+1}; c\theta_n, c(1-\theta_n) + 1) \\
&\ \propto (c(1-\theta_n) + 1)\,\mathrm{Beta}(\theta_{n+1}; c\theta_n, c(1-\theta_n) + 2) \\
&\qquad + c\theta_n\,(1 - r)\,\mathrm{Beta}(\theta_{n+1}; c\theta_n + 1, c(1-\theta_n) + 1),
\end{aligned}
$$

where the extra terms in the last expression come from the fact that $\mathrm{Beta}(\cdot; c\theta_n, c(1-\theta_n)+2)$ and $\mathrm{Beta}(\cdot; c\theta_n + 1, c(1-\theta_n)+1)$ have different normalization constants. □

D Inference for Hierarchical Beta Processes

Adding a Data Point

When we add a data point Y, there are two cases to consider. First, we can add Y as a new child of an internal node v (this happens if the CRP at that node creates a new table), or we can add Y to the subtree represented by a leaf w containing a datum X. Let $Z_1(Y, v)$ denote the probability that a new node of $\mathcal{T}'$ is generated as a child of v and creates the datum Y, and let $Z_2(Y, w, k)$ denote the probability that a datum first branches from the path of X k levels below w, and that the resulting datum is Y.

Let the path to v be given by $v_0, v_1, \ldots, v_n$ with $v_n = v$, and let $\mathrm{Size}(u)$ denotes the number of data in $\mathrm{Subtree}(u)$. Also let θ denote the parameter at v. Then we can calculate $Z_1(Y, v)$ as the probability that a datum follows the path to v, times the probability that a child of v would be equal to Y.

$$
\begin{aligned}
&Z_1(Y, v) = \\
&\left(\frac{\gamma}{\gamma + \mathrm{Size}(v)} \prod_{i=0}^{n-1} \frac{\mathrm{Size}(v_{i+1})}{\mathrm{Size}(v_i) + \gamma}\right) \prod_l \theta_l^{Y_l} (1 - \theta_l)^{1 - Y_l}.
\end{aligned}
$$

Calculating $Z_2(Y, v, d)$ is a bit trickier. Let us adopt notation similar to before, except with θ denoting the parameter at $p(w)$ and $w_0, \ldots, w_n$ denoting the path to w. We can compute the probability that the path of a datum goes through w in the same way as before. Then we can use Lemma C.1 to compute the probability of Y given that X and Y first split into unique subtrees at exactly k levels deeper than w. Letting $r = \left(\frac{c}{c+1}\right)^k$, the joint probability is given by

$$
\begin{aligned}
&Z_2(Y, w, k) = \\
&\qquad \left(\frac{1}{\gamma + \mathrm{Size}(w)} \prod_{i=0}^{n-1} \frac{\mathrm{Size}(w_{i+1})}{\mathrm{Size}(w_i) + \gamma}\right) \left(\frac{1}{1 + \gamma}\right)^k \frac{\gamma}{1 + \gamma} \\
&\quad \times \prod_{l : X_l = 0} [r\theta_l]^{Y_l} [1 - r\theta_l]^{1 - Y_l} \\
&\quad \times \prod_{l : X_l = 1} [1 - r(1 - \theta_l)]^{Y_l} [r(1 - \theta_l)]^{1 - Y_l}.
\end{aligned}
$$

The function $Z_2(Y, w, k)$ is a product of log-concave factors in k, and is therefore itself log-concave. We can thus find a rejection sampler with a constant acceptance rate of at least 0.25 (Leydold, 2003), and

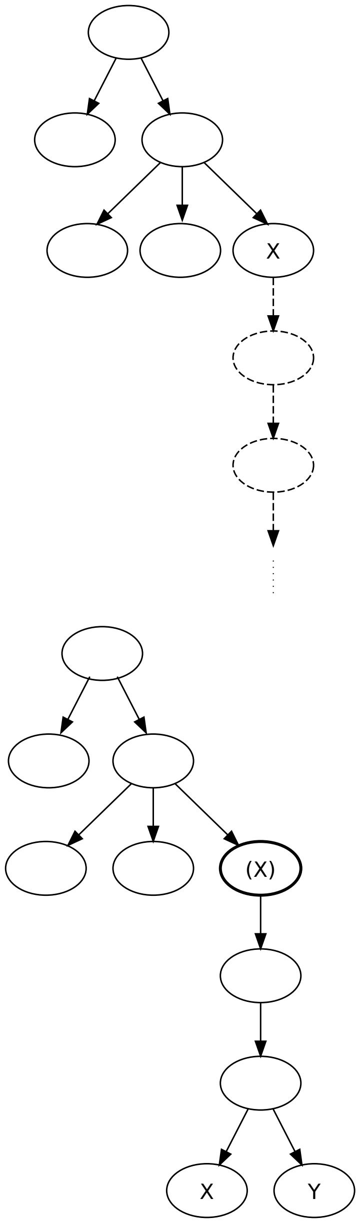

Figure 4: Illustration of how the tree is represented and modified by the inference algorithm. Top: X is a datum and thus corresponds to one of the leaves in the tree $\mathcal{T}'$. In the original tree $\mathcal{T}$, X corresponds to the infinite path represented by the dashed nodes. However, since no other data lie in that subtree, we ignore all of the dashed nodes when moving from $\mathcal{T}$ to $\mathcal{T}'$. Bottom: now a new datum Y is added to the same subtree as X. The paths of X and Y first diverge three levels below the old position of X. As a consequence, three new internal nodes needed to be created, and then X and Y are placed as the two children of the deepest of these nodes. If Y were to be removed from the tree, then these extra nodes would need to be removed and X would return to its old position.

compute the normalization constant $\hat{Z}_2(Y, w)$ of the enveloping function.

Now, to perform incremental Gibbs sampling, we add a data point to an internal node with probability proportional to $Z_1(Y, v)$, and we attempt to expand an external node with probability proportional to $\hat{Z}_2(Y, w)$. In the case that we try to expand an external node, we perform rejection sampling to determine what depth the two data points should branch at. If the sampler rejects, then we reject the Gibbs proposal, otherwise we insert the new data point at the given depth. We then need to sample all of the parameters at all of the newly created internal nodes, which can be done starting at the top and working iteratively towards the bottom using Lemma C.2.

Resampling Parameters

Resampling an internal parameter is straightforward in theory, since the conditional distribution over a parameter given its parent and children is log-concave (it is proportional to the product of several beta and Bernoulli densities). However, as noted before, there exist numerical issues when parameters are too close to either 0 or 1. We deal with this problem by assuming that we cannot distinguish between numbers that are less than some distance ϵ from 0 or 1. If we see such a number, we treat it as having a censored value (so it appears for instance as $\mathbb{P}[\theta < \epsilon]$ in the likelihood). A straightforward calculation shows that

$$\mathbb{P}[\theta_{v,l} < \epsilon \mid \theta_{p(v),l}] \approx \frac{\epsilon^{c\theta_{p(v),l}}}{c\theta_{p(v),l}},$$

and similarly

$$\mathbb{P}[\theta_{v,l} > 1 - \epsilon \mid \theta_{p(v),l}] \approx \frac{\epsilon^{c(1-\theta_{p(v),l})}}{c(1-\theta_{p(v),l})}.$$

With this strategy for dealing with the numerical issues, we now turn to the actual sampling algorithm.

The $\theta_{v,l}$ can be dealt with independently for different values of l, so we will restrict our attention to a fixed value of l. Suppose that θ is the parameter we want to sample, θ_0 is the value of its parent, $\theta_1, \ldots, \theta_m$ are the values of its children that are internal nodes, and $X_1, \ldots, X_p$ are the values of its children that are external nodes. Let $a = \sum_{j=1}^{p} X_j$ and $b = \sum_{j=1}^{p} 1 - X_j$. Then, letting $\mathrm{Beta}(\alpha, \beta)$ denote the normalization constant of a beta distribution, the likelihood for θ is given by

$$
\begin{aligned}
p(\theta \mid \theta_0, &\{\theta_i\}_{i=1}^m, \{X_j\}_{j=1}^p) \propto \\
&\theta^{c\theta_0+a-1}(1-\theta)^{c(1-\theta_0)+b-1} \\
\times &\prod_{i:\epsilon \leq \theta_i \leq 1-\epsilon} \frac{\theta_i^{c\theta-1}(1-\theta_i)^{c(1-\theta)-1}}{\mathrm{Beta}(c\theta, c(1-\theta))} \\
\times &\prod_{i:\theta_i < \epsilon} \frac{\epsilon^{c\theta}}{c\theta} \\
\times &\prod_{i:\theta_i > 1-\epsilon} \frac{\epsilon^{c(1-\theta)}}{c(1-\theta)}.
\end{aligned}
$$

One can check that this function is either (i) log-concave, (ii) has infinite density at $\theta = 0$, or (iii) has infinite density at $\theta = 1$. In the first case, we can sample from it efficiently (Leydold, 2003). In the second case, θ is very likely to be less than ϵ; since our sampler treats all numbers in the interval $[0, \epsilon)$ equivalently, we can arbitrarily set θ to 0. Similarly, in the third case, we can set θ to 1.

As a final note, we note that while this correction avoids the numerical issues of the sampler in (Thibaux, 2008), there is no longer any guarantee that the sampler converges to the true posterior distribution. While it might be somewhat desirable to obtain a characterization of the stationary distribution of this sampler, the real moral of the above is probably that the hierarchical beta process as it is currently formulated is not suitable for deep hierarchies. An interesting direction of future work would be to reformulate the HBP such that it is well-behaved even for infinitely deep hierarchies.