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. $\square$

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

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

$\square$

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

$$
\begin{aligned}
F(\alpha) &= \int_{\infty}^{\alpha} \int_0^1 x^{\tilde{\alpha}-1}(1-x)^{\beta-1} \log(x) dx d\tilde{\alpha} \\
&= \int_0^1 (1-x)^{\beta-1} \int_{\infty}^{\alpha} x^{\tilde{\alpha}-1} \log(x) d\tilde{\alpha} dx \\
&= \int_0^1 (1-x)^{\beta-1} \left(x^{\tilde{\alpha}-1} \Big|_{\infty}^{\alpha} \right) dx \\
&= \int_0^1 (1-x)^{\beta-1} x^{\alpha-1} \\
&= \mathrm{Beta}(\alpha, \beta)
\end{aligned}
$$

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*

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

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

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

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

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

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

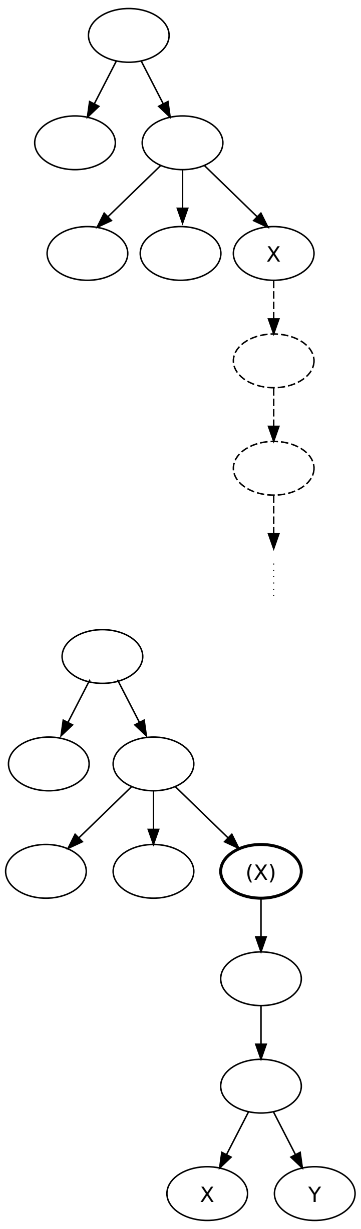

Figure 4: Illustration of how the tree is represented and modified by the inference algorithm. Top: X is a datum and thus corresponds to one of the leaves in the tree $\mathcal{T}'$. In the original tree $\mathcal{T}$, X corresponds to the infinite path represented by the dashed nodes. However, since 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.