Tree-Structured Stick-Breaking Processes for Hierarchical Modeling

Ryan Prescott Adams[*]
Department of Computer Science
University of Toronto
rpa@cs.toronto.edu

Zoubin Ghahramani
Department of Engineering
University of Cambridge
zoubin@eng.cam.ac.uk

Michael I. Jordan
Electrical Engineering and Computer Science
University of California, Berkeley
jordan@cs.berkeley.edu

Many data are naturally modeled by hierarchies, but often the hierarchy itself is unobserved. In this situation, it is appealing to construct a nonparametric prior on tree-structured partitions of data. Several such models have been proposed, but these typically have the property that the data live only at the bottom of the tree. This modeling assumption does not fit well with many data we would expect to model with hierarchies. For example, in topic modeling, *cars* might be a natural ancestor of *Hondas*, but we nevertheless expect to find some documents that are about cars generally and not about a specific brand.

To remedy this shortcoming, we propose a distribution over tree-structured measures, based on the stick-breaking approach to the Dirichlet process. Our construction provides an intuitive and flexible model for hierarchical data, while maintaining tractability for inference.

The Dirichlet process (DP) is a powerful and popular tool for specifying distributions of random measures. In particular, the stick-breaking construction by Sethuraman [1] has become a widely-used method for representing the DP. If Θ is the space of interest, and we wish to draw a random measure G from a Dirichlet process with base measure αH, we can do so via two sequences of independent random variables:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \qquad\qquad \theta_i \,|\, H \sim H$$

$$\pi_i = \nu_i \prod_{i'=1}^{i-1} (1 - \nu_{i'}) \qquad\qquad \nu_i \,|\, \alpha \sim \text{Beta}(1, \alpha),$$

where $\pi_1 = \nu_1$. This elegant method of constructing a Dirichlet process has proven so useful that the marginal distribution over the partition of the unit interval (the sequence $\pi_1, \pi_2, \ldots$) has its own name, the $\text{GEM}(\alpha)$ distribution, and there have been several generalisations of this aspect alone.

The key to Sethuraman's stick-breaking construction is a closure property of Dirichlet-distributed random vectors. Let $\boldsymbol{h}$, $\boldsymbol{s}$ and $\boldsymbol{s}'$ be vectors in the J-dimensional probability simplex where $\boldsymbol{s}$ has a Dirichlet distribution with parameter $\alpha\boldsymbol{h}$. Let j be a draw from the categorical distribution defined by $\boldsymbol{h}$, independent of $\boldsymbol{s}$. Using j, construct a J-dimensional binary vector $\boldsymbol{e}_j$ that has a one in the jth component and is zero everywhere else. Under these conditions, if $\nu \sim \text{Beta}(1, \alpha)$ independently of j and $\boldsymbol{s}$, then

$$\boldsymbol{s}' = \nu \boldsymbol{e}_j + (1 - \nu)\boldsymbol{s} \tag{1}$$

has a Dirichlet distribution with parameter $\alpha\boldsymbol{h}$. This property is closely tied to the conjugacy of the Dirichlet distribution to the multinomial.
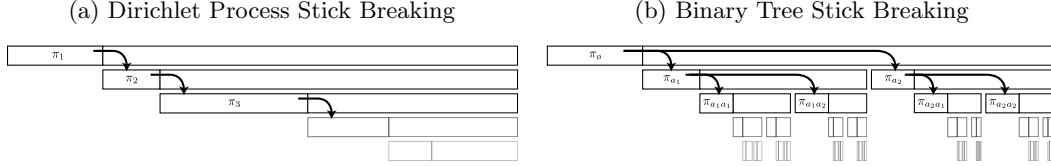
---

[*]http://www.cs.toronto.edu/~rpa/

Figure 1: The left figure shows the stick-breaking approach of Sethuraman. The right figure shows a binary version ($M=2$) of the tree-structured stick breaking presented here.

If we view the beta distribution in the above recursion as a special case of the Dirichlet distribution, then we immediately see that this invariance property can be generalised. Rather than $\nu \sim \text{Beta}(1, \alpha)$, we introduce $\boldsymbol{\eta}$, which takes values in the $M+1$-dimensional probability simplex and is distributed according to a Dirichlet distribution with parameter vector $[1, \alpha/M, \alpha/M, \cdots]$. Similarly, rather than a single Dirichlet random vector $\boldsymbol{s}$, we now introduce M of them, $\{\boldsymbol{s}_m\}_{m=1}^M$, $\boldsymbol{s}_m \sim \text{Dir}(\alpha\boldsymbol{h})$. It is easy to see that the invariance property of Equation 1 can be rewritten for arbitrary $M \in \mathbb{N}_1$ as

$$\boldsymbol{s}' = \eta_1 \boldsymbol{e}_j + \sum_{m=1}^M \eta_{m+1} \boldsymbol{s}_m \tag{2}$$

and $\boldsymbol{s}'$ again has a Dirichlet distribution with parameter $\alpha\boldsymbol{h}$. This can be viewed as a two-stage stick-breaking procedure. First, the standard $\text{Beta}(1, \alpha)$ break is performed, and then the remaining portion of the stick is broken into M pieces using a Dirichlet distribution with uniform parameters α/M. The process then recurses into the second-stage pieces.

In the Sethuraman approach, the graph structure over the partitions is naturally viewed as a chain. If we reconstruct the Dirichlet process with this expanded recursion, however, we have a stick-breaking that gives a tree topology of branching factor M. The marginal distribution over partitions arising from this process is still GEM.

To describe this more general topology, we develop some additional notation. Let $\mathcal{A} = \{a_1, a_2, \cdots, a_M\}$ be an alphabet of size M. A string composed from this alphabet is denoted $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_k)$, where $\epsilon_i \in \mathcal{A}$. We denote the set of all length-k strings as $\mathcal{A}^k$. The set $\mathcal{A}^0$ has only the null string, which we denote as $\boldsymbol{\epsilon} = \emptyset$. We define $\mathcal{A}^\star$ to be the union of all such sets $\mathcal{A}^k$ for $k \in \mathbb{N}_0$. We are particularly concerned with strings that are *prefixes* of other strings. If $\boldsymbol{\epsilon}'$ can be constructed by appending additional characters to $\boldsymbol{\epsilon}$, then we identify this as $\boldsymbol{\epsilon}' \succ \boldsymbol{\epsilon}$. Note that the null string $\emptyset$ is a prefix for all other strings. We also use the notation $\boldsymbol{\epsilon}a_m$ to indicate the string resulting from appending a_m to $\boldsymbol{\epsilon}$.

We have borrowed this notation from the closely-related Pólya tree (PT) construction [2]. These strings allow us to index nodes in the tree and thereby assign them partitions from the stick-breaking procedure. This representation also allows us to see that we have constructed a distribution on probability measures over strings of finite length, i.e., a way to generate random measures on $\mathcal{A}^\star$. Appending a new letter to the string corresponds to descending into the next level of the tree, while termination corresponds to reaching the atom specified by the completed string. With this representation, our stick-breaking method is given by

$$\nu_{\boldsymbol{\epsilon}} \,|\, \alpha \sim \text{Beta}(1, \alpha) \qquad\qquad \{\varphi_{\boldsymbol{\epsilon}a_1}, \varphi_{\boldsymbol{\epsilon}a_2}, \cdots, \varphi_{\boldsymbol{\epsilon}a_M}\} \,|\, \alpha, M \sim \text{Dir}(\alpha/M)$$

$$\pi_{\boldsymbol{\epsilon}a_m} = \nu_{\boldsymbol{\epsilon}a_m} \varphi_{\boldsymbol{\epsilon}a_m} \prod_{\boldsymbol{\epsilon}' \prec \boldsymbol{\epsilon}} \varphi_{\boldsymbol{\epsilon}'}(1 - \nu_{\boldsymbol{\epsilon}'}) \qquad\qquad \varphi_\emptyset = 1.$$

A graphical representation of this stick-breaking procedure and the resulting topology for $M = 2$ is shown in Figure 1. Often when constructing Bayesian models, we would prefer not to have to make an arbitrary choice of dimensionality, such as is enforced by the choice of the branching factor M. As in other nonparametric Bayesian models, we take the limit as M goes to infinity to obtain a Dirichlet process model for $\{\varphi_{\boldsymbol{\epsilon}a_1}, \varphi_{\boldsymbol{\epsilon}a_2}, \cdots\}$. This also gives us a way to allow greater flexibility in the model when we do not require the partitions to have a GEM marginal distribution, by using a different concentration parameter γ for the branching process than for the stick break on $\nu_{\boldsymbol{\epsilon}}$, which continues to use α. This more general nonparametric construction can be written
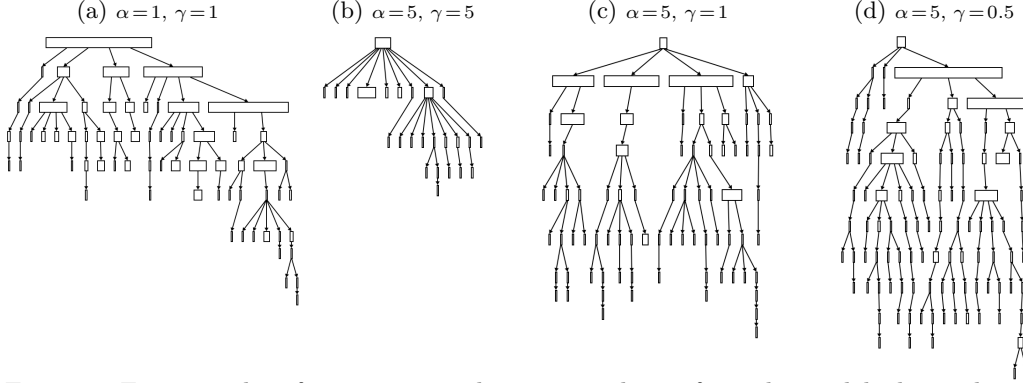
Figure 2: Four samples of tree-structured partitions drawn from the model, shown down to a minimum width of 0.005, with four different settings of the parameters. The $\alpha = \gamma = 5$ tree appears to have less mass because it has many small (unrendered) partitions.

$$\nu_{\boldsymbol{\epsilon}} \,|\, \alpha \sim \mathrm{Beta}(1, \alpha) \qquad\qquad \varphi_{\boldsymbol{\epsilon} a_m} = \psi_{\boldsymbol{\epsilon} a_m} \prod_{m'=1}^{m-1} (1 - \psi_{\boldsymbol{\epsilon} a_{m'}})$$

$$\psi_{\boldsymbol{\epsilon} a_m} \,|\, \gamma \sim \mathrm{Beta}(1, \gamma) \qquad\qquad \pi_{\boldsymbol{\epsilon} a_m} = \nu_{\boldsymbol{\epsilon} a_m} \varphi_{\boldsymbol{\epsilon} a_m} \prod_{\boldsymbol{\epsilon'} \prec \boldsymbol{\epsilon}} \varphi_{\boldsymbol{\epsilon'}} (1 - \nu_{\boldsymbol{\epsilon'}}),$$

where $\varphi_{\emptyset} = 1$ as before. This construction results in a distribution on trees with unbounded branching and unbounded depth. Figure 2 shows several draws from this process.

Even though it is very flexible, this prior distribution on measures on $\mathcal{A}^{\star}$ is conjugate to multinomial observations. Given an observed string $\boldsymbol{\epsilon}$, we can easily update the posterior of each relevant beta distribution. This gives the entire construction a *path reinforcing* property on the tree. Similarly, due to this conjugacy and the fact that we have defined this process in terms of random measures, data resulting from path reinforcing random walks down the tree are infinitely exchangeable.

We have so far only addressed a distribution on partitions and their topology and not discussed the corresponding draws from the base measure. In general, using draws from a single shared base measure will not take advantage of the tree structure of the partitions. Instead, a natural approach is to associate with each node a draw from a base measure that depends on its ancestors. There are a variety of different ways one might link data to tree-structured model parameters, but we believe one of these is of particular elegance and interest: the hierarchical Dirichlet process (HDP) [3]. To apply the HDP to this model, each child has a Dirichlet process whose base measure $G_{\boldsymbol{\epsilon} a_m}$ is its parent's Dirichlet process with concentration parameter β, up to a global base measure H, i.e.,

$$G_{\boldsymbol{\epsilon} a_m} \,|\, \beta, G_{\boldsymbol{\epsilon}} \sim \mathrm{DP}(\beta, G_{\boldsymbol{\epsilon}}) \qquad\qquad G_{\emptyset} \,|\, \beta, H \sim \mathrm{DP}(\beta, H).$$

Several other authors have introduced tree-structured random measures based on the Dirichlet process, but these all have different properties (and, indeed, different objectives) than the construction we have presented here. Like the Pólya tree, the Dirichlet diffusion tree (DFT) [4] uses a binary branching process to yield Dirichlet-based random measures that support continuous densities. In both of these approaches, however, the data appear at the bottom of the infinitely-deep tree and do not assign atoms of mass to internal nodes. While the Pólya tree can be easily generalised to trees of arbitrary branching factor, generalising the DFT to this case is more awkward.

[1] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
[2] R. Daniel Mauldin, William D. Sudderth, and S. C. Williams. Pólya trees and random distributions. *The Annals of Statistics*, 20(3):1203–1221, September 1992.
[3] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[4] Radford M. Neal. Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian Statistics 7*, pages 619–629, 2003.