Hierarchical Tree-Structured Stick-Breaking Priors

Hiroshi Noji
National Institute of Informatics, Tokyo
Department of Informatics
Graduate University for Advanced Studies
noji@nii.ac.jp

Daichi Mochihashi
The Insutite of Statistical Mathematics, Tokyo
Department of Statistical Science
Graduate University for Advanced Studies
daichi@ism.ac.jp

Yusuke Miyao
National Institute of Informatics, Tokyo
Department of Informatics
Graduate University for Advanced Studies
yusuke@nii.ac.jp

Abstract

We present a Bayesian nonparametric prior called the hierarchical tree-structured stick-breaking prior, which is a natural extension to the tree-structured prior of Adams et al. (2010). The discrete measure drawn from this prior has correlations with each other as in the hierarchical Dirichlet process, while the atoms in each measure have also positive correlations hierarchically. Using this prior, we construct an HMM in which each state is organized in a latent taxonomy on the tree, which grows in a nonparametric fashion.

1 Introduction

In the world, many data are composed of categories that exhibit hierarchical tree structure, i.e., taxonomy: all categories that are descendants of another category are subtypes of the latter category, which is a supertype of the former categories. Because we generally cannot observe these structures, we wish to recover the latent hierarchy of categories and the members comprising each category.

Nonparametric Bayesian methods are attractive to handle the potentially infinite size of the latent tree. People have enjoyed finding topic hierarchy [1, 2] or modeling human category learning [3] using nonparametric priors over the tree. Tree-structured stick breaking prior [1] is an extension to the Dirichlet process where each atom of the multinomial π drawn from this prior corresponds to a node of the tree. An interesting property of this prior is that the probabilities of two events $x_1, x_2 \sim \pi$ have correlations by a random walk behavior similar to the Dirichlet diffusion trees [4]. Unfortunately, this prior cannot be used for models in which multiple distributions that share a common tree structure are required, e.g., document modeling where a topic distribution for a document is defined on a tree, which is shared across multiple documents[1] or sequence modeling like HMM, in which each state is organized in a latent tree.

We propose the hierarchical tree-structured stick-breaking prior: by using a draw from the tree-structured stick-breaking as a base measure, we draw multiple discrete distributions on a tree, which shares the structure with the base measure. This process is similar to the hierarchical Dirichlet process [7], but we can exploit the tree structure to encode rich hierarchical information between distributions as well as learning the tree structure itself. For an application, we demonstrate an HMM on the potentially infinite tree, which recovers the latent taxonomy-like structure of hidden

[1] [2] can only model each topic distribution along a path on the tree. Recently some models that remove this limitation have been proposed [5, 6]. We discuss the relationships between these models in a later section.

states. HMMs are prevalent in many unsupervised learning tasks. In natural language processing, Bayesian HMMs have been popular for unsupervised part-of-speech induction [8, 9], but current systems ignore the inherent hierarchies of parts-of-speech. In contrast, our model can potentially find, for example, a group of nouns where pronouns and proper nouns are closely on a sub-tree.

2 Tree-Structured Stick-Breaking Prior

Adams et al. extend Sethuraman's stick-breaking construction of the Dirichlet process [10] to a tree-structured process, which generates a distribution over tree-structured infinite partitions on a unit interval. This process can be explained by two types of stick breakings, ν-break and ψ-break: ν-break determines the weight for staying at a node as $\nu_\epsilon \sim \mathrm{Beta}(1, \alpha)$, while ψ-break determines weights for selecting a node to descend as $\psi_\epsilon \sim \mathrm{Beta}(1, \gamma)$. We use subscript ϵ to denote a node, which specifies the path to that node; ϵi is an i-th child node of ϵ. ϕ is the root node. The actual construction of partition $\boldsymbol{\pi} = (\pi_\phi, \pi_1, \pi_{1\cdot 1}, \cdots, \pi_{1\cdot 2}, \cdots, \pi_2, \cdots)$ is as follows:

$$\pi_\epsilon = \nu_\epsilon \varphi_\epsilon \prod_{\epsilon' \prec \epsilon} \varphi_{\epsilon'}(1 - \nu_{\epsilon'}), \qquad \varphi_{\epsilon i} = \psi_{\epsilon i} \prod_{j=1}^{i-1} (1 - \psi_{\epsilon j}), \qquad \pi_\phi = \nu_\phi, \tag{1}$$

where $\{\epsilon' : \epsilon' \prec \epsilon\}$ is a set of ancestor nodes of ϵ. Note that we slightly simplify notations from the original paper [1]. We summarize this process as $\boldsymbol{\pi} \sim \mathrm{TSSB}(\alpha, \gamma)$.

What is the benefit of endowing hierarchy to the Dirichlet process in this manner? Adams et al. emphasizes the importance of hierarchical emissions: we can define an emission distribution G_ϵ at node ϵ using that of parent node ϵ' by e.g., the hierarchical Dirichlet process $G_\epsilon \sim \mathrm{DP}(\eta, G_{\epsilon'})$. This is interesting, but an additional benefit that have not explicitly touched on [1] is that two atoms π_{ϵ_1} and π_{ϵ_2} have positive correlations if the two nodes are located closely on a sub-tree because each π_ϵ is determined by a succession of binary branching processes. This property becomes more interesting when we consider generating multiple distributions on a tree. For example, consider a topic model where each document is modeled with a tree-structured topic distribution and each node corresponds to a topic. This model can capture correlations among topics, e.g., one document may have larger weight on a sub-tree, which broadly covers the sports topic, while another document may have larger weight on the sub-tree of politics. However, the tree-structured stick-breaking cannot be used for these purposes, because each draw from this prior creates a different tree structure, so correspondences of topics will not be preserved among documents.

3 Hierarchical Tree-Structured Stick-Breaking Prior

The problem of the previous section motivates us to endow another type of hierarchy to the tree-structured stick-breaking, which is similar to the construction between the Dirichlet process and the hierarchical Dirichlet process (HDP) [7]. In the following, we use superscripts with parentheses to distinguish different distributions. Our goal is to define a set of distributions $\{\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \cdots\}$, which share the same tree structure. For familiality, we give the model exposition with the topic model as an application. We extend the idea to construct an HMM on a tree in section 4.

We describe the generative process of each $\boldsymbol{\pi}^{(k)}$ as follows:

$$\boldsymbol{\pi} \sim \mathrm{TSSB}(\alpha, \gamma), \qquad \boldsymbol{\pi}^{(k)} \sim \mathrm{HTSSB}(a, b, \boldsymbol{\pi}). \tag{2}$$

The central component of this process is HTSSB, shorthand for the hierarchical tree-structured stick breaking. Informally, HTSSB receives a distribution over a tree $\boldsymbol{\pi}$ as a prior, and then returns a distribution over the same tree with different weights on each node, which enables applications like a topic model in the previous section ($\boldsymbol{\pi}^{(k)}$ is a topic distribution of document k).

HTSSB also constructs distribution $\boldsymbol{\pi}^{(k)}$ by two types of stick-breakings: ν-break samples $\nu_\epsilon^{(k)}$ for the stopping decision, and ψ-break samples $\psi_\epsilon^{(k)}$ for child branching. Using these variables, the mechanism of the construction of each $\boldsymbol{\pi}^{(k)}$ is the same as TSSB:

$$\pi_\epsilon^{(k)} = \nu_\epsilon^{(k)} \varphi_\epsilon^{(k)} \prod_{\epsilon' \prec \epsilon} \varphi_{\epsilon'}^{(k)} (1 - \nu_{\epsilon'}^{(k)}), \qquad \varphi_{\epsilon i}^{(k)} = \psi_{\epsilon i}^{(k)} \prod_{j=1}^{i-1} (1 - \psi_{\epsilon j}^{(k)}), \qquad \pi_\phi^{(k)} = \nu_\phi^{(k)}, \tag{3}$$

Figure 1 shows how $\boldsymbol{\pi}^{(k)}$ is defined on a unit interval.

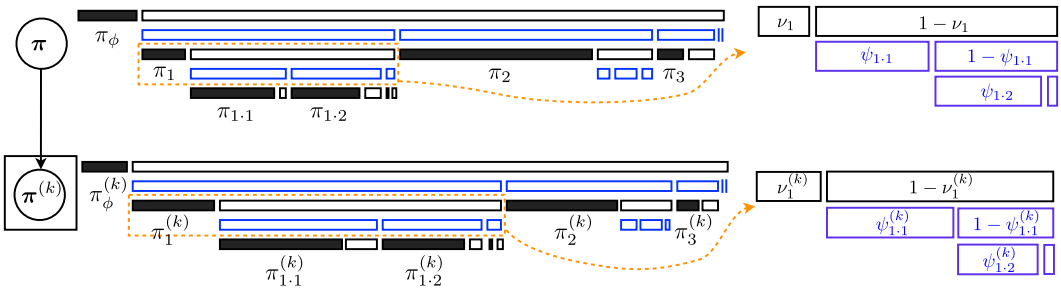


Figure 1: The construction of $\boldsymbol{\pi}^{(k)}$ given a prior $\boldsymbol{\pi}$. Each $\boldsymbol{\pi}^{(k)}$ is defined on the same tree as $\boldsymbol{\pi}$, but has a different weight on each node. Rows 1, 3 and 5 show ν-breaks while rows 2 and 4 show ψ-breaks. A symbol below a stick (e.g., $\pi_{1\cdot1}$) is the probability of stopping at that node, while a symbol inside a stick (e.g, $\psi_{1\cdot1}$ in the right part) shows the ratio of the left stick to the right stick.

ν-break and ψ-break We endow a hierarchy between ν_ϵ and $\nu_\epsilon^{(k)}$ (ψ_ϵ and $\psi_\epsilon^{(k)}$) as follows:

$$\nu_\epsilon^{(k)} \sim \text{Beta}(a\nu_\epsilon, a(1-\nu_\epsilon)) \qquad \psi_\epsilon^{(k)} \sim \text{Beta}(b\psi_\epsilon, b(1-\psi_\epsilon)), \tag{4}$$

which can be interpreted as a two-dimensional Dirichlet and a and b are concentrations determining the strength of prior weights. A collapsed Gibbs sampler for this model can be constructed. We collapse all Beta variables and replace these with the Chinese restaurant processes whose restaurants have only two types of customers: customers selecting left sticks and customers selecting right sticks. In the topic model, we first sample topic z_{ji} of word w_i in document j from the posterior. We then add a customer selecting the left to the restaurant on $\nu_{z_{ji}}^{(k)}$, and add customers selecting the right to the restaurants on the path to z_{ji}. If a customer creates a new table, then we also add a customer to the corresponding restaurant on the base measure $\boldsymbol{\pi}$.

Relationships to other models Recently, some authors propose the models similar to ours for the tree-structured topic models. Although our construction is composed completely with Beta variables, the nested Chinese restaurant franchise (nCRF) [6] describes similar process based on the hierarchical Dirichlet processes: data at a node selects a child node with a Dirichlet process, whose base measure is the one on the same location of the global tree that defines $\boldsymbol{\pi}$. Their model can be represented in our formalism using the connection between the stick-breaking and the hierarchical Dirichlet process [7]. This is achieved by replacing the ψ-break of (4) with

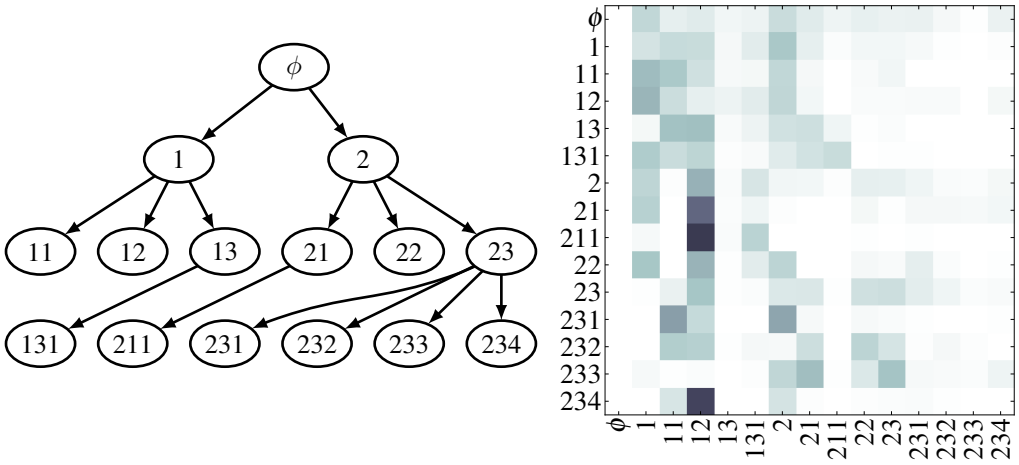
$$\psi_{\epsilon i}^{(k)} \sim \text{Beta}(a(1-\nu_\epsilon)\varphi_{\epsilon i}, a(1-\nu_\epsilon)(1-\textstyle\sum_{j=1}^{i-1}\varphi_{\epsilon j})). \tag{5}$$

Differently from our formulation, they do not distinguish between the ν-break and ψ-breaks at a node by assigning one child to the stop choice on every nodes. Note that this construction loses a hyper-parameter b in (4). More rigorous theoretical comparison between models is interesting but is beyond the scope of this paper. Another related model is the nested hierarchical Dirichlet process [5], which is almost the same as nCRF, but it ignores hierarchies between stop decisions (ν-breaks in our case) as in nCRF and ours.

4 HMM on a Tree

We now construct the HMM on a tree using the presented hierarchical prior. In the model, each node corresponds to a latent state that is organized in a tree-structure to represent subtype-supertype relations between induced categories in an unsupervised fashion.

Our fundamental assumption is: two nodes close to each other (e.g., two children of a node) will have similar transition distributions. In the domain of language modeling where each state (node) corresponds to a word category (part-of-speech) and the transition distribution corresponds to the syntactic role of that category, this means that two categories close to each other have common syntactic properties, but also have some variations. For example, singular nouns and plural nouns both tend to precede verb categories, while only singular nouns tend to precede 3rd person present verbs. Cohen and Smith [11] encode these covariance information on manually-defined part-of-speech tags for the grammar induction problem with the shared logistic-normal priors. Our motivation is different. We are interested in finding the tree structure of word categories in a data-driven manner, not



(a) Sampled tree

(b) Sampled transition matrix

Figure 2: A draw from the generative model of the HMM on a tree. See the body for details of parameter settings. (b) is the transition matrix where an entry (i, j) represents the transition probability from node i to j. Darker color indicates higher probability.

capturing information on manually defined tags. The induced category tree is appealing not only from a scientific perspective, but also from an engineering one: at prediction of new data, we can represent the confidence of a state as the depth of the predicted node. We discuss the model with a simple HMM for a starting point, but the idea can be extended to other grammar induction problems, e.g., state splitting of manually defined tags on a syntactic tree [12, 13].

The generative process of this HMM is described as follows:

1. Sample $\boldsymbol{\pi} \sim \mathrm{TSSB}(\alpha, \beta)$ to define the global tree structure.
2. Sample the transition distribution at node ϵ as $\boldsymbol{\pi}^{(\epsilon)} \sim \mathrm{HTSSB}(a, b, \boldsymbol{\pi}^{(\epsilon')})$ by descending the tree in a breath-first manner (ϵ' is the parent node of ϵ).

This model encodes the assumption discussed above. To demonstrate this, we simulate the generative process above to draw node-to-node transition distributions. We set hyper-parameters as follows: $\alpha = 2.0, \beta = 0.25$ to encourage vertical growth of a tree, and $a = 3.0, b = 5.0$ so that the model prefers sharing of branching probabilities between parent-child distributions rather than sharing of stopping decisions. The structure of $\boldsymbol{\pi}$ is determined by truncating the sticks when the remaining length < 0.05. Figure 2(a) shows the sampled tree structure and Figure 2(b) illustrates the transition distributions on this tree. This transition matrix captures important characteristics of the model. When we inspect rows, a transition distribution of a node is similar to that of the parent node, e.g., rows of 13 and 131 both have higher probabilities to 11, 12, 2, 21 and 211. This property comes from the hierarchies of the model between distributions. We can observe another interesting feature by inspecting columns, for example, columns of 21, 22 and 23 look similar, which indicate, e.g., nodes which prefer transitions to 21 also tend to prefer transitions to 22 and 23. This property comes from the mechanism of tree-structured stick-breaking of each distribution and the sharing of internal branching probabilities.

5 Discussion

Finding latent hierarchies is fundamental for deeper understandings of the world. The presented nonparametric priors reveal the latent tree shared across groups of data, which we believe will be attractive for many data analysis. Although we only have touched the general properties of prior draws in this paper, our goal is, of course, to recover the latent tree by simulating the posterior. In general, the inference of tree structure is hard for extensive search spaces. To overcome this problem, we are currently exploring and implementing the efficient sampling algorithms.

References

[1] R.P. Adams, Z. Ghahramani, and Michael I Jordan. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, pages 19–27, 2010.

[2] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, (2).

[3] Kevin R. Canini and Thomas L. Griffiths. A nonparametric Bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

[4] Radford M. Neal. Density modeling and clustering using dirichlet diffusion trees. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 619–629, 2003.

[5] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested Hierarchical Dirichlet Processes. *arXiv.org*, stat.ML, October 2012.

[6] Amr Ahmed, Liangjie Hong, and Alexander Smola. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1426–1434, May 2013.

[7] Yee Whye Teh, Michael I Jordan, M.J. Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[8] Sharon Goldwater and Tom Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[9] Phil Blunsom and Trevor Cohn. A hierarchical pitman-yor process hmm for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 865–874, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

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

[11] Shay Cohen and Noah A. Smith. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[12] Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697.

[13] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. The infinite tree. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 272–279, Prague, Czech Republic, June 2007. Association for Computational Linguistics.