Deep Nested Hierarchical Dirichlet Processes

Lavanya Sita Tekumalla
Indian Institue of Science
lavanya.iisc@gmail.com

Priyanka Agrawal
IBM Research India
priyanka.svnit@gmail.com

Indrajit Bhattacharya
TCS Innovation Labs, India
b.indrajit@tcs.com

1 Introduction

The focus of this paper is on deep Bayesian non-parametric models and, more specifically, on admixture models of this kind. In admixture or mixed membership models [5], each data item is modeled as a mixture over components, and such models have significantly richer representational power compared to mixture models and also wide applicability. The Hierarchical Dirichlet Process (HDP) [5] extends the notion of finite admixture models, famously captured by Latent Dirichlet Allocation (LDA) [4], to infinite mixture components. The HDP does this by coupling Dirichlet Processes (DP) [2], such that a draw from a DP serves as the base distribution for another DP. In this paper, we explore how such layers of admixtures can be nested arbitrarily deeply. While the Bayesian framework serves as a natural protection against overfitting, we would like to provide as much flexibility to a model within this framework. Layers of admixtures, instead of mixtures, provides such flexibility. The Author-Topic Model [9] explores such a two level model in the finite setting, where documents are mixtures over 'topics' and topics are mixtures over 'authors'. We explore such models in the infinite setting, which are also arbitrarily deep - where the representation consists of layers of entities, and each entity is a mixture over entities at the previous layer.

We build upon the idea of the nested Dirichlet Process [8], which is a two level 'nesting' of Dirichlet processes - where one Dirichlet Process is the base distribution of another Dirichlet Process. We extend this idea to show that layers of admixtures can be created by nesting Hierarchical Dirichlet Processes in a similar way. Such a nesting can then be made arbitrarily deep - hence the name deep Nested Hierarchical Dirichlet Processes.

We note that the nested CRP [3] and its extension (also interestingly called the nested HDP) [7], while being deep non-parametric models, are not deep admixture models in our sense. In these, 'topic' distributions are structured as an arbitrarily deep and arbitrarily wide tree, where a topic at any layer has exactly one topic from the previous layer as its 'parent'. Our model is significantly more flexible by allowing entities at each layer to have a distribution over all entities at the previous layer.

We further explore relations between such an infinite admixture model and finite counterparts, and observe that the deep nested HDP arises as infinite limits of two different constructions of deep finite admixture models. We then study sampling based inference algorithms for the deep nHDP based on an equivalent Chinese Restaurant Process representation, which we call the deep nested Chinese Restaurant Franchise. We extend two different sampling algorithms that have been proposed for the HDP - a direct and an indirect sampling scheme based on table indexes. While these two algorithms are known to perform similarly for the single-layer HDP, we show that, when extended to multiple layers, the direct sampling scheme scales linearly with number of layers, while the indirect sampling scheme has complexity growing exponentially with number of layers.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

We demonstrate this difference in complexity of the two sampling algorithms using experiments on text corpora, while also showing that deep non-parametric admixture models show better generalization performance than their shallow counterparts.

In an earlier conference paper [1], we had introduced the notion of nesting between HDPs to create 2-layer author topic model. In this paper, we have extended the definition for deep nesting, formally studied the relationship with finite admixture models, extended the inference algorithms for arbitrary number of layers and identified how they differ in complexity beyond a single layer.

2 Model

We define a multi-layer Nested Hierarchical Dirichlet Processes which has an HDP at every layer. Let L denote the number of layers of nesting, indexed by $l \in \{0, ..., L-1\}$. The base distribution of HDP at layer l is the HDP at layer $l-1$ and we term this as *nesting* of two HDPs.

Stick-breaking Construction: For creating an admixture at layer $l \in \{0, ..., L-1\}$ we follow the HDP process. We create an infinite set of entities from previous layer $l-1$ sampled from base distribution B^l: $\{\phi^l_k\}_{k=1}^{\infty} \sim B^l$ and a global entity distribution for this layer over these: $G^l_B = \sum_{k=1}^{\infty} \beta^l_k \delta_{\phi^l_k}$, where the weights are drawn from a stick-breaking distribution, $\beta^l \sim GEM(\gamma^l)$. An infinite set of entities at layer l are defined over layer $l-1$ entities, where entity r in layer l is defined by its own distinct local entity distribution: $G^l_r = \sum_{k=1}^{\infty} \pi^l_{rk} \delta_{\phi^l_k}$. The weights are drawn from a DP with the global popularities of layer $l-1$ entities as the base distribution: $\pi^l_r \sim DP(\alpha^l, \beta^l)$.

We can recognize this overall construction for layer l entities as a draw from an HDP, which we name as H^l: $G^l_r \sim$ HDP$(\alpha^l, \gamma^l, B^l) \equiv H^l$. In our nested structure, H^l is the base distribution for next layer: $B^{l+1} = H^l$. Note that this can equivalently be represented as $B^{l+1} =$ HDP$(\alpha^l, \gamma^l, B^l) =$ HDP$(\alpha^l, \gamma^l,$HDP$(\alpha^{l-1}, \gamma^{l-1}, B^{l-1}))$. We define this nested structure as the nested HDP (**nHDP**) and write - $B^{l+1} =$ **nHDP**$(l+1, \{\alpha^l, \gamma^l\}, ..., \{\alpha^0, \gamma^0\}, \bar{H})$

The details of the generative process are as follows. First the distributions for the 'entities' at different layers are sampled starting with the 'topics' at layer $l = 0$.

prior for topics $\quad B^0 = \bar{H} = Dir(\gamma^0)$

layer $l \leq L$

$$\phi^l_k \sim B^l$$

global weight for l-entity $\quad \beta^l \sim GEM(\gamma^l)$

$$G^l_B = \sum_{k=1,...} \beta^l_k \delta_{\phi^l_k} \sim D^l = DP(\gamma^l, B^l)$$

l+1 entity's wt for l-entity $\quad \pi^l_r \sim DP(\alpha^l, \beta^l)$

$$G^l_r = \sum_k \pi^l_{rk} \delta_{\phi^l_k} \sim H^l = \ HDP(\alpha^l, \gamma^l, B^l)$$

coupling between layers $\quad B^{l+1} = H^l \ =$ nHDP$(l+1, \{\alpha^l, \gamma^l\}, \cdots, \{\alpha^0, \gamma^0\}, \bar{H})$

There is an equivalent *indirect representation* of the local topic preferences using topic samples (also called tables) [10]. Let $\{\psi^l_{jt}\}_{t=1}^{\infty}$ denote the samples/tables for the j^{th} entity at layer l drawn from the topic distributions at the previous layer: $\phi^l_{jt} \sim G^{l-1}_B$, and $\{k^l_{jt}\}_{t=1}^{\infty}$ denote the set of indexes of the layer $l-1$ topics corresponding to each topic sample. Their corresponding weights $\{\bar{\pi}^l_{jt}\}_{t=1}^{\infty}$ are drawn from a stick-breaking distribution $GEM(\alpha^l)$, $\pi^l_j \sim GEM(\alpha^l)$.

l+1 entity's wt for l-table $\quad \bar{\pi}^l_r \sim GEM(\alpha^l)$

$$G^l_r = \sum_t \bar{\pi}^l_{rt} \delta_{\psi^l_{rt}} \sim H^l = HDP(\alpha^l, \gamma^l, B^l)$$

Finally, the data items are generated by sampling entities at each layer l as $\theta^l_{ji} \sim \theta^{l+1}_{ji} \in \{\phi^l_1 \ldots\}$. The data item (word) is sampled at the final layer as $x_{ji} \sim \theta^0_{ji}$. Note that in the document modeling use-case, the grouping of words with respect to the level L-entity is observed and termed as a *document*.

Restaurant Process Representation: Just as the HDP has a restaurant process interpretation - the Chinese Restaurant Franchise (CRF)[10] - for the deep nested HDP we show an equivalent interpretation in terms of multiple layers of nested CRFs, corresponding to the multiple layers of HDP. In the restaurant interpretation, each group r represents a restaurant at each layer l, the entity samples ψ^l_{rt} are called tables, and the sampled entities ϕ^l_k are called dishes, which get served at tables.

For a given entity layer $l \in \{0, ..., L-1\}$, assume K entities $\{\phi^l_1, ..., \phi^l_K\}$ have been drawn from base distribution B^l. Let $(\psi^l_{r1}, ..., \psi^l_{r,t-1})$ be sequence of entity samples / tables drawn from G^l_B for r^{th} entity of this layer and an indirect representation of G^l_r is constructed using entity samples as above. Due to the nested structure, the predictive distribution for the draw of layer l entity for word ji denoted by θ^l_{ji} is additionally conditioned on the corresponding draw at layer $l+1$ i.e. θ^{l+1}_{ji}. Given that $\theta^{l+1}_{ji} = G^l_r, \theta^l_{ji} \sim \theta^{l+1}_{ji}$ is drawn by integrating out the group level distribution G^l_r.

$$\theta^l_{ji}|\theta^{l+1}_{ji} = G^l_r, \theta^l_{11}, \ldots, \theta^l_{1N_1}, \theta^l_{21}, \ldots, \theta^l_{j,i-1}, \alpha^l, G^l_B \sim \sum_{t=1}^{m^l_{r.}} \frac{n^l_{rt}}{i-1+\alpha^l} \delta_{\psi^l_{rt}} + \frac{\alpha^l}{n^l_{r..} + \alpha^l} G^l_B \quad (1)$$

where n^l_{rt} is the number of times any of $\{\theta^l_{11}, \ldots, \theta^l_{1N_1}, \theta^l_{21}, \ldots, \theta^l_{j,i-1}\}$ got assigned to the t^{th} atom ψ^l_{rt} of restaurant r and $m^l_{rk} = \sum_t \delta(\psi^l_{rt}, \phi^l_k)$ is the number of entity samples in r^{th} restaurant assigned to previous layer entity ϕ^l_k.

The predictive distribution of t^{th} entity sample, after integrating out G^l_B is as follows:

$$\psi^l_{rt}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{rt-1}, \gamma, B^l \sim \sum_{k=1}^{K^l} \frac{m^l_{.k}}{m_{..} + \gamma^l} \delta_{\phi^l_k} + \frac{\gamma^l}{m^l_{..} + \gamma^l} B^l \quad (2)$$

where $m^l_{.k}$ is the number of entity samples across all restaurants in layer l assigned to k^{th} entity ϕ^l_k of previous layer.

Relation with deep finite admixture models: The HDP can be shown to arise as the infinite limit of two different finite admixture models [10]. We show that a similar relationship persists across the nested coupling between the deep nHDP and deep finite admixture models.

Consider a L-layer admixture model $G_j(L, \{K^l\})$ with K^l entities at layer l defined using direct sampling of entities.

Theorem 1 *For each $l \in \{0, \ldots, L\}$, as $K^l \to \infty$, the finite layer approaches an nHDP.*

$$\lim_{K^l \to \infty} G_j(L, \{K^l\}) = G^L_j \sim nHDP(L, \{\alpha^l, \gamma^l\}, \bar{H}) \quad (3)$$

A similar result holds for the indirect L-layer finite admixture model construction $G_j(L, \{K^l\}, \{T^l\})$ with K^l entities and T^l entity samples or tables at layer l.

Theorem 2 *For each $l \in \{0, \ldots, L\}$, as $K^l \to \infty$, and $T^l_r \to \infty, \forall r \in \{1, \ldots, K^l\}$, the generative process of multi-layer finite admixture model is equivalent to the nHDP.*

$$\lim_{K^l, T^l \to \infty} G_j(L, \{K^l\}, \{T^l\}) = G^L_j \sim nHDP(L, \{\alpha^l, \gamma^l\}, \bar{H}) \quad (4)$$

Deep Nested Non-parametric Flexible Models: We end with a discussion of an enhancement to our model, where each layer has the flexibility to be either an admixture or a mixture, while retaining its non-parametric nature. This may be useful in the presence of specific knowledge about the relations between entities in the domain. To achieve a mixture, instead of an admixture, at any layer, we replace an HDP-HDP nesting with a DP-HDP nesting (where an HDP has a DP as the base distribution) or a DP-DP nesting, depending on the nature of the subsequent layer. The instance of this model with a mixture at every layer is directly related to the nCRP [3]. The inference algorithms that we propose for the nested HDP in the next section can be modified in a reasonably straight-forward manner for these flexible variants.

3 Inference

We use Gibbs sampling for approximate inference. The conditional posterior for these variables can be derived from the nCRF conditionals. We propose two inference schemes, building upon similar schemes for the HDP:

Indirect Sampling: The conditional distributions from the nCRF scheme lend themselves to an inference algorithm. Hence, we sample at every level $l \in \{0, \dots, L\}$, the table assignments t^l_{ji}, the t^{th} level table for customer i of the j document, and dish assignments k^l_{rt} for tables t at restaurant r at level l. We refer to this as the *nCRF index sampling* inference scheme. However unlike the inference for a single level HDP, a naive approach of sampling all the above indices is intractable leading to an exponential complexity at each level due to the tight coupling between the variables.

Lemma: The complexity of the nCRF index sampling scheme is $O((T^L_{max})^{L+1}MN) + O(LK^L_{max}(T^L_{max})^{S^{max}})$ where S^{max} is intuitively the upper bound on the count of all data items (words) assigned to a single table at a specific restaurant at a specific level. , K^l_{max} is the maximum number of dishes at any level and T^L_{max} is the maximum number of tables in a single restaurant at any level, M the total number of documents and N_{max} the maximum number of words in any document.

Direct Sampling: We propose an alternative *nCRF direct sampling scheme*, similar to the direct sampling scheme in [10] that samples the dishes z^l_{ji} at each level l for customer i from the j^{th} document conditioned on dish assignments at the remaining levels.

$$p(z^l_{ji} = p|\mathbf{z^l}_{-\mathbf{ji}}, z^{l+1}_{ji} = r, z^{l-1}_{ji} = q, \mathbf{z^l}_{-\mathbf{ji}}, \mathbf{m}, \beta, \mathbf{x}) \propto p(z^l_{ji} = p|\mathbf{z^l}_{-\mathbf{ji}}, z^{l+1}_{ji} = r)p(z^{l-1}_{ji} = q|\mathbf{z^{l-1}_{-ji}}, z^l_{ji} = p)$$

The first term is the predictive distribution of z^l_{ji} given the level $l+1$ dish assignment r, while the second term arises from the previous level dish assignment q that depends on the value of z^l_{ji}. Hence, $p(z^l_{ji} = p|\mathbf{z^l}_{-\mathbf{ji}}, z^{l+1}_{ji} = r)$ can be viewed as consisting of two terms. One from picking an existing table in restaurant r with dish assignment p and one from creating a new table in restaurant r at level l and assigning the dish p to it.

$p(z^l_{ji} = p|\mathbf{z^l}_{-\mathbf{ji}}, z^{l+1}_{ji} = r) \propto \frac{n^l_{rp}+\alpha^l\beta^l_p}{n^l_{r.}+\alpha^l}$ for an existing dish and $\frac{\alpha^l\beta^l_{new}}{n^l_{r.}+\alpha^l}$ for a New dish. Similarly, $p(z^{l-1}_{ji} = q|\mathbf{z^{l-1}_{-ji}}, z^l_{ji} = p) \propto \frac{n^{l-1}_{pq}+\alpha^{l-1}\beta^{l-1}_q}{n^{l-1}_{p.}+\alpha^{l-1}}$ for and Existing dish and $\frac{\alpha^{l-1}\beta^{l-1}_{new}}{n^{l-1}_{p.}+\alpha^{l-1}}$ for a new dish.

We sample β as $(\beta^l_1, \beta^l_2 \dots \beta^l_{K^l}, \beta_{new}) \sim Dir(m^l_{.1}, m^l_{.2} \dots m^l_{.K}, \gamma^l)$. We adapt the method from [6] for sampling the table counts m^l_{rk}.

Lemma: In each iteration, the complexity of direct sampling algorithm at layer l is $O(MN_{max}K^l_{max})$ where M is the number of documents, N_{max} is the maximum number of words in any document and K^l_{max} is current number of entities at layer l.

4 Experiments

Datasets: We use the following publicly available publication datasets for our experimental analysis. The *NIPS* dataset[1] is a collection of NIPS proceedings (volume 0-12). with 1,740 documents contributed by a total of 2,037 authors, with total 2,301,375 word tokens resulting in a vocabulary of 13,649 words.

Perplexity with Number of Layers: We observe that addition of non-parametric layers lead to better generalization performance over a finite model. Also a deeper model with more layers leads to better generalization performance

Comparing Direct and Indirect Sampling: We run the direct sampling and the Indirect sampling algorithm, on a two level nested non-parametric flexible model (DP-HDP model) and a comparison of run-time is shown in the figure. We observe that even for 2 layers, direct sampling scheme is significantly faster than the indirect sampling scheme.

[1] http://www.arbylon.net/resources.html

Model Model	Finite 2 level	nHDP 1 level	nHDP 2 level
Perplexity	2783	1775	1247

Table 1: Perplexity of Finite-2Level, nHDP-1Level and nHDP-2Level for NIPS

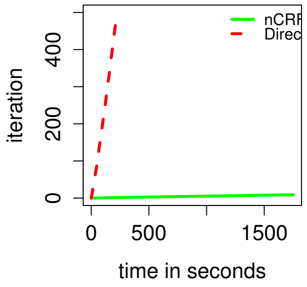

Table 2: Comparing run-time of two layer flexible nHDP (the DP-HDP) model

5 Conclusion

We have proposed nested Hierarchical Dirichlet Processes(nHDP) for deep multilevel non-parametric admixture modeling. We further explore relations between such a nested infinite admixture model and it's finite counterparts, and show that the deep nested HDP arises as infinite limit of deep finite admixture models. We have explored two techniques for posterior inference based on the Gibbs sampling and show that the direct sampling technique scales efficiently for arbitrarily deep models.

References

[1] P. Agrawal, L. Tekumalla, and I. Bhattacharya. Nested hierarchical dirichlet process for nonparametric entity-topic analysis. *ECML-PKDD*, 2013.

[2] C. Antoniak. Mixtures of Dirichlet Processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2(6):1152–1174, 1974.

[3] D. Blei, T. Griffiths, M. Jordan, and J. Tanenbaum. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *JACM*, 2010.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[5] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PANS*, 101-suppl(1), 2004.

[6] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A sticky hdp-hmm with application to speaker diarization. *Annals of Applied Stats.*, 5(2A):1020–1056, 2011.

[7] J. Paisley, C. Wang D. Blei, and M. Jordan. Nested hierarchical dirichlet processes. *Arxiv*, 2012.

[8] A. Rodriguez, D. Dunson, and A. Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[9] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *UAI*, 2004.

[10] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.