Spectral Methods for the Hierarchical Dirichlet Process

Hsiao-Yu Fish Tung, Chao-Yuan Wu, Manzil Zaheer, Alexander J. Smola
Machine Learning Department
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
{htung, chaoyuaw, manzilz}@cs.cmu.edu, alex@smola.org

Abstract

The Hierarchical Dirichlet Process (HDP) is a versatile, albeit computationally expensive tool for statistical modeling of mixture models. In this paper, we introduce a spectral algorithm. We show that it is both computationally and statistically efficient. In particular, we derive the lower-order moments of the HDP and give reconstruction guarantees. Moreover, we show that hierarchical spectral method is able to generate a better results regarding likelihood performance.

1 Introduction

HDP mixture models are useful in modeling problems involving grouped data, where each observation within a group is drawn from a mixture model and it is desirable to share mixture components across all the groups. A natural application with this property is topic modeling for documents, possibly supplemented by an ontology. Inference in HDPs is typically carried out by using Markov Chain Monte Carlo methods [1, 2] or stochastic variational algorithms [3]. Unfortunately these methods are very costly when the model becomes complicated and has several layers.

To speed up parameter inference under HDP model, the present paper extends prior work on spectral methods from Dirichlet Processes [4], Hidden Markov Models [5], the Indian Buffet Process [6] to yet another model, the HDP. Spectral method is shown to provide superior speed in parameter inference under several models and their corresponding applications. The main idea of spectral learning is to use method of moments to infer topics. The algorithm is simple and is easy to implement. Besides, we will show later that the computation time is not affected by the number of layers.

Different from previous spectral learning work, our work provides the following adaptation:

- We introduce *hierarchical* decomposition tools. While we apply them to the *Hierarchical* Dirichlet Process, they are valuable beyond that, e.g. for hierarchical structures in general, such as the Nested Chinese Restaurant Franchise [2] or hierarchical clustering [7].
- The convergence analysis illustrates the interplay between sample size and data in hierarchies and can be applied on any arbitrary hierarchical tree structure.
- The algorithm is able to produce better result then LDA without increasing the time consumption.

[8] shows that a two-layer version of the algorithm already outperform sampling-based algorithms significantly both in terms of perplexity and speed. Since time consumption of our algorithm does not increase with the number of layers, the algorithm enjoys significant improvement in time over HDP sampler. In summary, the present work contributes to completing the tool set of spectral methods. This is an important goal to ensure that entire models can be translated wholly into spectral algorithms, rather than just parts.

2 Hierarchical Dirichlet Process

The HDP [1] uses a Dirichlet Process (DP) [9, 10] G_j for each group j of data to handle uncertainty in number of mixture components. At the same time, in order to share mixture components and clusters across groups, each of these DPs is drawn from a global DP G_0. More formally, we have the following statistical description of a L-level HDP.

Trees Denote by $\mathcal{T} = (V, E)$ a tree of depth L. For any vertex $\mathbf{i} \in \mathcal{T}$ we use $p(\mathbf{i}) \in V$, $c(\mathbf{i}) \subset V$ and $l(\mathbf{i}) \in \{0, 1..., L-1\}$ to denote the parent, the set of children and level of the vertex respectively. When needed, we enumerate the vertices of $\mathcal{T}$ in dictionary order. For instance, the root node is denoted by $\mathbf{i} = (0)$, whereas $\mathbf{i} = (0, 4, 2)$ is the node obtained by picking the fourth child of the root node and then the second child thereof respectively. Finally, we have sets of observations $X_\mathbf{i}$ associated with the vertices $\mathbf{i}$ (in some cases only the leaf nodes may contain observations). This yields

$$\begin{aligned} G_0 &\sim \mathrm{DP}(H, \gamma_0) & \theta_{\mathbf{i}j} &\sim G_\mathbf{i} \\ G_\mathbf{i} &\sim \mathrm{DP}\left(G_{p(\mathbf{i})}, \gamma_{l(\mathbf{i})}\right) & x_{\mathbf{i}j} &\sim \mathrm{Categorical}(\theta_{\mathbf{i}j}) \end{aligned}$$

Here $\gamma_\mathbf{i}$ denotes the concentration parameter at vertex $\mathbf{i}$ and H is the base distribution which governs the a priori distribution over data items. Figure 1 illustrates the full model.

As explained earlier, the distributions $G_\mathbf{i}$ have a stick breaking representation sharing common atoms:

$$G_\mathbf{i} = \sum_{v=1}^{\infty} \pi_{\mathbf{i}v} \delta_{\phi_v} \text{ with } \phi_v \sim H. \tag{1}$$

3 Moments of the HDP

To construct spectral method for HDP, a crucial step is to derive the orthogonally decomposable tensors from the moments. We follow the notations in [11] and define $\otimes$ to be the outer product, $v^{\otimes p} := v \otimes \cdots \otimes v$ to be a p-th order tensor powers and $A := \sum_{i=1}^{k} v_i^{\otimes p}$ to be a rank-p tensor. Moreover, in order to neatly describe tensor multiplication in our algorithm, for a rank-p tensor, we define the multilinear tensorial reduction $T(A, V^1, V^2 \dots V^p) \in \mathbb{R}^{m_1 \times \cdots \times m_p}$ as

$$T(A, V^1, V^2 \dots V^p)_{i_i, i_2 \dots i_p} := \sum_{j_1, j_2 \dots j_p \in [n]} A_{j_1, j_2 \dots j_p} V^1_{j_1, i_1} V^2_{j_2, i_2} \cdots V^p_{j_p, i_p}.$$

Since the order of the data is exchangable, we introduce a symmetrization operator $\mathfrak{S}_k$ as

$$\mathfrak{S}_k[A] = \frac{1}{k!} \sum_{\pi \in S(k)} A_\pi, \tag{2}$$

where $S(k)$ denotes the symmetric group and π denotes the permutations.

The first-order moment is equivalent to the weighted sum of latent topics using topic distribution under node $\mathbf{i}$, so it is simply the weighted combination of Φ where the weight vector is π_0, i.e,

$$M_1 := \mathbf{E}[x] = \mathbf{E}[\Phi h_{\mathbf{i}j}] = \mathbf{E}[\Phi \pi_\mathbf{i}] = \Phi \pi_0. \tag{3}$$

The last equation uses the fact that, for $\pi \sim \mathrm{Dirichlet}(\gamma_0 \pi_0)$, $\mathbf{E}[\pi] = \pi_0$. Besides, for such variable π, using the definition of Dirichlet distribution, we have $\mathbf{E}[[\pi^2]_{ii}] = \frac{\gamma_0}{\gamma_0+1} \pi_{0i}(\pi_{0i} + 1)$ and $\mathbf{E}[[\pi^2]_{ij}] = \frac{\gamma_0}{\gamma_0+1} \pi_{0i} \pi_{0j}$. The second-order moment thus becomes

$$M_2 := \mathbf{E}[x_1 \otimes x_2] = \mathbf{E}\left[\Phi h_{\mathbf{i}1} h_{\mathbf{i}2}^T \Phi^T\right] = \Phi \mathbf{E}\left[\pi_\mathbf{i} \pi_\mathbf{i}^T\right] \Phi^T = \Phi E \Phi^T, \tag{4}$$

where $[E]_{ii} = \frac{1}{\gamma_0+1} \pi_{0i}(\gamma_0 \pi_{0i} + 1)$ and $[E]_{ij} = \frac{\gamma_0}{\gamma_0+1} \pi_{0i} \pi_{0j}$. Matrix E can be decompose as the summation of a diagonal matrix and a symmetric matrix, $\pi_0 \otimes \pi_0$. By replacing E with these two matrices, the second-order moment can be re-written as

$$M_2 = \Phi E \Phi^T = \Phi \left(\frac{\gamma_0}{\gamma_0 + 1} \pi_0 \otimes \pi_0 + \frac{1}{\gamma_0 + 1} \mathrm{diag}(\pi_0) \right) \Phi^T, \tag{5}$$

where $\Phi\pi_0$ in the first term can be further replaced with M_1. Thus, we define the second term as the second-order tensor, which is a rank-k matrix,

$$S_2 := M_2 - \frac{\gamma_0}{\gamma_0+1} M_1 \otimes M_1 = \Phi\left(\frac{1}{\gamma_0+1}\mathrm{diag}(\pi_0)\right)\Phi^T = \sum_{i=1}^{K} \frac{\pi_{0i}}{\gamma_0+1}\phi_i \otimes \phi_i. \tag{6}$$

The third-order tensor is defined in the form of $S_3 := \sum_{i=1}^{K} C_6 \cdot \pi_{0i} \cdot \phi_i \otimes \phi_i \otimes \phi_i$, and can be derived using M_1, M_2 and M_3 by applying the same technique of decomposing matrix or tensor into the summation of symmetric tensors and diagonal tensor. The derivation details for a multi-layer HDP tensor is provided in the Appendix. Furthermore, Lemma 2 in the Appendix shows the generalized tensors for HDP with different number of layers. Using Lemma 2, we found that the coefficient and moment for different hierarchical tree can be derived recursively using a bottom-up approach, i.e., coefficient for $k-$layer HDP can be derived using the coefficient of $(k-1)$-layer HDP and moments at a node $\mathbf{i}$ can be derived using the moments calculating under its children, $c(\mathbf{i})$. The recursive rule is provided in Lemma 2.

Lemma 1 (Symmetric Tensors of HDP) *Given a L-level HDP, with hyperparameters $\gamma_{\mathbf{i}}$, the symmetric Tensors for a node $\mathbf{i}$ at layer l can be expressed as:*

$$S_1^{\mathbf{i}} := M_1^{\mathbf{i}} = T(\pi_{\mathbf{i}}, \Phi),\ S_2^{\mathbf{i}} := M_2^{\mathbf{i}} - C_2^l \cdot S_1^{\mathbf{i}} {S_1^{\mathbf{i}}}^T = T(C_3^l \cdot \mathrm{diag}\left(\pi_{\mathbf{i}}\right), \Phi, \Phi),$$
$$S_3^{\mathbf{i}} := M_3^{\mathbf{i}} - C_4^l \cdot S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}} - C_5^l \cdot \mathfrak{S}_3\left[S_2^{\mathbf{i}} \otimes M_1^{\mathbf{i}}\right] = T(C_6^l \cdot \mathrm{diag}\left(\pi_{\mathbf{i}}\right), \Phi, \Phi, \Phi),$$

where

$$C_2^{(l)} = \frac{\gamma_{l+1}}{\gamma_{l+1}+1} C_2^{(l+1)},\ C_3^{(l)} = C_3^{(l+1)} + \frac{C_2^{(l)}}{\gamma_{l+1}},\ C_4^{(l)} = \frac{{\gamma_{l+1}}^2}{(\gamma_{l+1}+1)(\gamma_{l+1}+2)} C_4^{(l+1)}$$
$$C_5^{(l)} = \frac{\gamma_{l+1}}{(\gamma_{l+1}+1)} \frac{C_3^{(l+1)}}{C_3^{(l)}} C_5^{(l+1)} + \frac{1}{\gamma_{l+1} C_3^{(l)}} C_4^{(l)},\ C_6^{(l)} = C_6^{(l+1)} + 3 \cdot \frac{C_5^{(l)} C_3^{(l)}}{\gamma_{l+1}} - \frac{C_4^{(l)}}{\gamma_{l+1}^2}$$

Here $M_r^{\mathbf{i}}$ denotes the k-th moment at node $\mathbf{i}$

$$M_r^{\mathbf{i}} := \frac{1}{|c(\mathbf{i})|} \sum_{\mathbf{j} \in c(\mathbf{i})} M_r^{\mathbf{j}} \tag{7}$$

starting with $M_r^{\mathbf{i}} = \mathbb{E}[\bigotimes_{s=1}^{r} x_{\mathbf{i}s}]$ whenever $\mathbf{i}$ represents an leaf node. In other words, $M_r^{\mathbf{i}}$ is obtained by averaging the associated moments over all of its children evenly.

4 Spectral Algorithm for HDP

Here we provide a simple method for estimating number of topics, K. Next, we introduce a way to estimate the moment at the leaf nodes. The estimated moments are used to estimate the diagonalized tensors. Last, applying general tensor decomposition technique gives us the estimated topic vectors.

Inferring mixture number In contrast to the CRF where the number of mixture components k is settled by means of repeated sampling, we use an approach that directly infers k from data itself. The concatenation of all the first-order moments spans the space of Φ with high probability. Thus, the number of linearly independent mixtures k, is close to the rank of $\tilde{M}_1$. While direct calculation of the rank of $\tilde{M}_1$ is expensive, one can estimate k by the following procedure: draw a random matrix $R \in \mathbb{R}^{n_l \times k'}$ for some $k' \geq k$, and examine the eigenvalues of $\tilde{M}_1' = \left(\tilde{M}_1 R\right)^T \left(\tilde{M}_1 R\right) \in \mathbb{R}^{k' \times k'}$. We estimate the rank of $\tilde{M}_1$ to be the point where abrupt decrease occurs on the magnitude of eigenvalues.

Moment estimation An L-level HDP could be viewed as a L-level tree, where each node represents a DP. Then the estimated moments for the whole model can be calculated recursively by (7). The estimated moments at the leaf node $\mathbf{i}$ are defined as:

$$\hat{M}_r^{\mathbf{i}} := \varphi_r\left(\mathbf{x}_{\mathbf{i}}\right) \text{ for leaf } \text{ where } \varphi_r(\mathbf{x}_{\mathbf{i}}) := \frac{(m_{\mathbf{i}} - r)!}{m_{\mathbf{i}}!} \sum_{j_1, j_2} x_{\mathbf{i}j_1} \otimes x_{\mathbf{i}j_2} \cdots \otimes x_{\mathbf{i}j_r}$$

and $m_{\mathbf{i}}$ is the number of words in the observation $x_{\mathbf{i}}$. Here (x_1, x_2) and (x_1, x_2, x_3) denote the ordered tuples in $\mathbf{x}$, with x_i encoded as a binary vector, i.e. $x_i = e_j$ iff the i-th data is j.

We then apply Excess Correlation Analysis (ECA) to infer hidden topics, Φ. Dimensionality reduction and whitening is then performed on $\hat{S}_r^0$ to make the eigenvectors of it orthogonal and to project to a lower dimensional space. Finally, we decompose the tensor and transform the eigenvectors back into the original space to obtain the reconstructed $\hat{\Phi}$. The complete algorithm is described in Algorithm 1 in the Appendix with details given. Besides, concentration of measure for these estimated quantities are given in Section F in the Appendix.

5 Experiments

An attractive application of HDP is topic modelling in a corpus where in documents are grouped naturally. We use Enron email corpus [12] and Multi-Domain Sentiment Dataset [13] to validate our algorithm. After the usual cleaning steps (stop word removal, numbers, infrequent words), our training dataset for Enron consisted of $167,851$ emails sent with $10,000$ vocabulary size and average 91 words in each email. Among these, $126,697$ emails are sent internally within Enron and 41154 are from external sources. In order to show that the topics are able to cover topics from external and internal sources and are not biased toward the larger group, we have 537 internal emails and $4,63$ external email in our test data. To evaluate the computational efficiency of the spectral algorithms (using fast count sketch tensor decomposition(FC) [8], robust tensor method (RB) and alternating least square (ALS)), we compare the CPU time and per-word likelihood among these approaches.

Table 1: Results on Enron dataset with different tree structures and different solvers: spectral HDP using fast count sketch method (sketch length is set to 10), alternating least square (ALS) and robust tensor power method (RB).

Tree	K		sHDP (FC)	sHDP (ALS)	sHDP (RB)
Enron 2-layer	50	like./time	8.09/**67**	7.86/119	7.86/2641
	100	like./time	8.16/**104**	7.82/668	7.82/5841
Enron 3-layer	50	like./time	7.93/**68**	7.78/121	**7.77**/2710
	100	like./time	8.18/**101**	7.69/852	**7.68**/5782

We further compare spectral method under balanced/unbalanced tree structure of data on Multi-Domain Sentiment Dataset. The dataset contains reviews from Amazon reviews that fall into four categories: books, DVD, electronics and kitchen. We generate 2 training datasets where one has balanced number of reviews under each categories (1900 reviews for each category) and the other has highly unbalanced number of examples at the leaf node (1900/1500/700/300 reviews for the four categories), while the test dataset consisted of 100 reviews for each categories. The result in Table 2 show that spectral algorithm with multi-layers structure will perform even better than with flat model when the tree structure is unbalanced.

Table 2: Results on Sentimental dataset. Train data 1 is selected so that the there are balanced numbers of reviews under each category. Train data 2 is selected to have highly unbalanced child number at the leaf nodes.

Tree	train data 1	K=50	K=100	train data 2	K=50	k= 100
Sentiment 2-layer	like./time	**7.9**/38	7.99/151	like./time	8.23/36	8.14/147
Sentiment 3-layer	like./time	7.92/37	**7.96**/150	like./time	**8.17**/38	**8.07**.148

The results of the experiments throw light on two key points. First, leveraging the information in the form of hierarchical structure of documents, instead of blinding grouping the documents into a single-layer model like LDA, will result in better performance (i.e. higher log-likelihood) under different settings. The tree structure is able to eliminate the pernicious effects caused by unbalanced data. For example, a 2-layer model like LDA considers every email to be equally important, and so for a topic relating to external world it will perform worse, as most of the emails are exchanged within the company and they are unlikely to possess topics related to the external emails. Second, although spectral method cannot obtain a solution that has higher performance in perplexity, it can be used as a tool for picking up a nice initial point.

References

[1] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 576, pp. 1566–1581, 2006.

[2] A. Ahmed, L. Hong, and A. Smola, "Nested chinese restaurant franchise processes: Applications to user tracking and document modeling," in *ICML*, Hyderabad, India, 2013.

[3] C. Wang, J. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," in *Proc. Intl. Conference on Artificial Intelligence and Statistics*, 2011.

[4] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu, "Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation," *CoRR*, vol. abs/1204.6703, 2012.

[5] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. J. Smola, "Hilbert space embeddings of hidden markov models," in *International Conference on Machine Learning*, 2010.

[6] H. Tung and A. J. Smola, "Spectral methods for indian buffet process inference," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 1484–1492. [Online]. Available: http://papers.nips.cc/paper/5516-spectral-methods-for-indian-buffet-process-inference

[7] R. Adams, Z. Ghahramani, and M. Jordan, "Tree-structured stick breaking for hierarchical data," in *Neural Information Processing Systems*, 2010, pp. 19–27.

[8] A. S. Yining Wang, Hsiao-Yu Tung and A. Anandkumar, "Fast and guaranteed tensor decomposition via sketching," *NIPS*, 2015.

[9] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[10] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[11] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *arXiv preprint arXiv:1210.7559*, 2012.

[12] B. Klimt and Y. Yang, "Introducing the enron corpus." in *CEAS*, 2004.

[13] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Association for Computational Linguistics*, Prague, Czech Republic.

[14] A. Anandkumar, R. Ge, and M. Janzamin, "Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates," *CoRR*, vol. abs/1402.5180, 2014. [Online]. Available: http://arxiv.org/abs/1402.5180

[15] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[16] C. McDiarmid, "On the method of bounded differences," in *Survey in Combinatorics*. Cambridge University Press, 1989, pp. 148–188.

A HDP structure

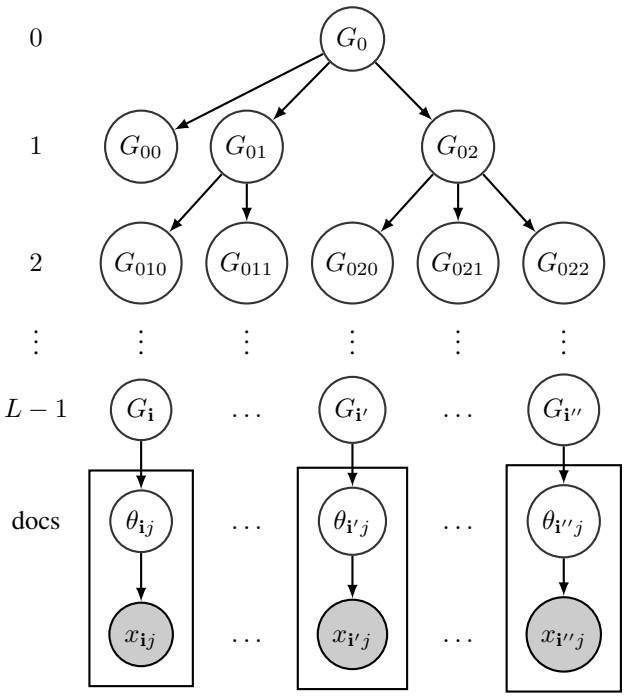

Figure 1: Hierarchical Dirichlet Process with observations at the leaf nodes.

B Generalized HDP Tensor and Proof of Symmetric Tensors

Symmetric Tensors for the HDP
We begin our analysis by deriving the moments for a three layer HDP. This allows us to provide detail without being hampered by cumbersome notation. After that, we analyze the general expansion.

B.1 Multiple Layer HDP

Lemma 2 (Symmetric Tensors of HDP) *For an L-level HDP, with hyperparameters,* $\gamma_1, \gamma_2, \ldots \gamma_{L-1}$ *we have*

$$S_1 := M_1 = T(\pi_0, \Phi),$$
$$S_2 := M_2 - C_2 \cdot S_1 S_1^T = T(C_3 \cdot \mathrm{diag}\left(\pi_0\right), \Phi, \Phi),$$
$$S_3 := M_3 - C_4 \cdot S_1 \otimes S_1 \otimes S_1 - C_5 \cdot \mathfrak{S}_3 \left[S_2 \otimes M_1\right] = T(C_6 \cdot \mathrm{diag}\left(\pi_0\right), \Phi, \Phi, \Phi),$$

The key difference is that here the coefficients C_i *are recursively defined since we need to take expectations all the way up to the root node. This yields*

$$C_2 = \frac{\prod_{i=1}^{L-1} \gamma_i}{\prod_{i=1}^{L-1} (\gamma_i + 1)}, \quad C_3 = \sum_{i=1}^{L-1} \frac{\prod_{j=1}^{i-1} \gamma_{L-j}}{\prod_{j=1}^{i} (\gamma_{L-j} + 1)};$$

$$C_4 = \frac{\prod_{i=1}^{L-1} {\gamma_i}^2}{\prod_{i=1}^{L-1} \left((\gamma_i + 1)(\gamma_i + 2)\right)}, \quad C_5 = \sum_{i=1}^{L-1} \frac{\prod_{j=1}^{i-1} \gamma_{L-j} \prod_{j=1}^{L-1} \gamma_j}{\prod_{j=1}^{i} (\gamma_{L-j} + 2) \prod_{j=1}^{L-1} (\gamma_j + 1)} / C_3$$

$$C_6 = 3\sum_{i=1}^{L-2}\sum_{j=i+1}^{L}\frac{\prod_{k=1}^{i-1}\gamma_{L-k}\prod_{k=1}^{j-1}\gamma_{L-k}}{\prod_{k=1}^{i}(\gamma_{L-k}+2)\prod_{k=1}^{j}(\gamma_{L-k}+1)} + 2\sum_{i=1}^{L-1}\frac{\prod_{j=1}^{i-1}\gamma_{L-j}{}^2}{\prod_{j=1}^{i}((\gamma_{L-j}+1)(\gamma_{L-j}+2))}$$

B.2 Three Layers

The three layer HDP is structurally similar to LDA. Its tensors are derived in [4]. We begin by considering a three model to gain intuition of how to obtain the general format of the tensors. The goal is to reconstruct the latent factors in Φ. In the case of topic modeling, the j-th column denotes the word distribution of the j-th topic.

Lemma 3 (Symmetric tensors of 3-layer HDP) *Given a 3-layer HDP with hyperparameters γ_1 and γ_2 at layers 1 and 2 respectively, the symmetric tensors are given by*

$$\begin{aligned}
S_1 &:= M_1 = T(\pi_0, \Phi),\\
S_2 &:= M_2 - C_2 S_1 S_1^T = M_2 - \frac{\gamma_2\gamma_1}{(\gamma_2+1)(\gamma_1+1)} S_1 S_1^T = T\left(C_3 \cdot \operatorname{diag}(\pi_0), \Phi, \Phi\right),\\
S_3 &:= M_3 - C_4 \cdot S_1 \otimes S_1 \otimes S_1 - C_5 \cdot \mathfrak{S}_3\left[S_2 \otimes M_1\right] = T\left(C_6 \cdot \operatorname{diag}(\pi_0), \Phi, \Phi, \Phi\right),
\end{aligned}$$

where

$$\begin{aligned}
C_3 &= \frac{\gamma_2+\gamma_1+1}{(\gamma_1+1)(\gamma_2+1)}\\
C_4 &= \frac{\gamma_1^2\gamma_2^2}{(\gamma_1+1)(\gamma_1+2)(\gamma_2+1)(\gamma_2+2)}\\
C_5 &= \frac{\gamma_2\gamma_1(\gamma_1+\gamma_2+2)}{(\gamma_1+2)(\gamma_2+2)(\gamma_1+\gamma_2+1)}\\
C_6 &= \frac{6\gamma_1+6\gamma_2+2\gamma_1^2+2\gamma_2^2+3\gamma_2\gamma_1+4}{(\gamma_1+1)(\gamma_1+2)(\gamma_2+1)(\gamma_2+2)}.
\end{aligned}$$

Proof Note that by definition of the Dirichlet Process, the means match that of the reference measure. This means that we can integrate over the hierarchy

$$\begin{aligned}
\mathbf{E}[x] &= \mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\Phi\pi_{\mathbf{i}}\right]\right] &(8)\\
&= \mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\Phi\pi_{p(i)}\right] &(9)\\
&= \Phi\pi_{p(p(\mathbf{i}))} = \Phi\pi_0 &(10)
\end{aligned}$$

Then deriving the first-order tensor is straightforward,

$$S_1 := M_1 = \mathbf{E}_x[x_1] = \Phi\pi_0 = T(\pi_0, \Phi) \tag{11}$$

Similarly, to derive the second order tensor, we first need the following terms: for $i \neq j$ we have

$$\begin{aligned}
&\mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\Phi_i\pi_{\mathbf{i}i}\pi_{\mathbf{i}j}^T\Phi_j^T\right]\right] &(12)\\
&= \mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\Phi_i\frac{\gamma_2\pi_{p(\mathbf{i})i}\pi_{p(\mathbf{i})j}^T}{\gamma_2+1}\Phi_j^T\right]\\
&= T\left(\frac{\gamma_2}{\gamma_2+1}\frac{\gamma_1\pi_{0i}\pi_{0j}}{\gamma_1+1}, \Phi_i, \Phi_j\right) &(13)
\end{aligned}$$

Likewise, when the indices match, we obtain

$$\begin{aligned}
&\mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\Phi_i\pi_{\mathbf{i}i}\pi_{\mathbf{i}i}^T\Phi_i^T\right]\right] &(14)\\
&= \mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\Phi_i\frac{(\gamma_2\pi_{p(\mathbf{i})i}+1)\pi_{p(\mathbf{i})i}}{(\gamma_2+1)}\Phi_i^T\right]
\end{aligned}$$

$$= T\left(\frac{\gamma_2}{(\gamma_2+1)}\frac{\gamma_1\pi_{0i}^2+\pi_{0i}}{(\gamma_1+1)}+\frac{1}{(\gamma_2+1)}\pi_{0i},\Phi_i,\Phi_i\right)$$

Then the moment M_2 could be written as

$$\begin{aligned}M_2 =&\mathbf{E}_x\left[x_1\otimes x_2\right]=\mathbf{E}\left[\mathbf{E}_x\left[x_1\otimes x_2|G_{\mathbf{i}}\right]\right]\\ =&\mathbf{E}\left[\mathbf{E}_x\left[x_1|G_{\mathbf{i}}\right]\otimes\mathbf{E}_x\left[x_2|G_{\mathbf{i}}\right]\right]\\ =&\Phi\mathbf{E}\left[\mathbf{E}\left[\pi_{\mathbf{i}}\pi_{\mathbf{i}}^T\right]\right]\Phi^T\\ =&\frac{\gamma_2\gamma_1}{(\gamma_2+1)(\gamma_1+1)}S_1\otimes S_1\\ &+T\left(\frac{\gamma_2+\gamma_1+1}{(\gamma_1+1)(\gamma_2+1)}\cdot\mathrm{diag}\left(\pi_0\right),\Phi,\Phi\right)\end{aligned}$$

The second-order symmetric tensor S_2 could then be obtained by defining

$$\begin{aligned}S_2 :=& M_2-\frac{\gamma_2\gamma_1}{(\gamma_2+1)(\gamma_1+1)}S_1\otimes S_1\\ =&T\left(\frac{\gamma_2+\gamma_1+1}{\gamma_1(\gamma_1+1)(\gamma_2+1)}\cdot\mathrm{diag}\left(G_0\right),\Phi,\Phi\right).\end{aligned}\tag{15}$$

Before deriving the third-order tensor, we derive $\mathbf{E}[(\Phi_i\pi_{\mathbf{i}i})\otimes(\Phi_j\pi_{\mathbf{i}j})\otimes(\Phi_j\pi_{\mathbf{i}j})]$ for the following three cases. First, for $i=j=k$, we have:

$$\begin{aligned}&\mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\left(\Phi_i\pi_{\mathbf{i}i}\right)^{\otimes 3}\right]\right]\\ &=\mathbf{E}\left[T(\frac{\pi_{p(\mathbf{i})i}\left(\gamma_2\pi_{p(\mathbf{i})i}+1\right)\left(\gamma_2\pi_{p(\mathbf{i})i}+2\right)}{(\gamma_2+1)(\gamma_2+2)},\Phi_i,\Phi_i,\Phi_i)\right]\\ &=T\Bigg(\frac{\gamma_2^2}{(\gamma_2+1)(\gamma_2+2)}\frac{\pi_{0i}(\gamma_1\pi_{0i}+1)(\gamma_1\pi_{0i}+2)}{(\gamma_1+1)(\gamma_1+2)}+\frac{3\gamma_2}{(\gamma_2+1)(\gamma_2+2)}\frac{\pi_{0i}(\gamma_1\pi_{0i}+1)}{(\gamma_1+1)}\\ &\qquad+\frac{2\pi_{0i}}{(\gamma_2+1)(\gamma_2+2)},\Phi_i,\Phi_i,\Phi_i\Bigg)\end{aligned}\tag{16}$$

Second, for $i=j\neq k$, we have:

$$\begin{aligned}&\mathbf{E}_{G_{p(\mathbf{i})}|G_{p(p(\mathbf{i}))},\gamma(p(\mathbf{i}))}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\left(\Phi_i\pi_{\mathbf{i}i}\right)^{\otimes 2}\otimes\left(\Phi_k\pi_{\mathbf{i}k}\right)\right]\right]\\ &=\mathbf{E}\left[T\left(\frac{\pi_{p(\mathbf{i})i}\left(\gamma_2\pi_{p(\mathbf{i})i}+1\right)\gamma_2\pi_{p(\mathbf{i})k}}{(\gamma_2+1)(\gamma_2+2)},\Phi_i,\Phi_i,\Phi_k\right)\right]\\ &=T\left(\frac{\gamma_2^2}{(\gamma_2+1)(\gamma_2+2)}\frac{\pi_{0i}(\gamma_1\pi_{0i}+1)\gamma_1\pi_{0k}}{(\gamma_1+1)(\gamma_1+2)}+\frac{\gamma_2}{(\gamma_2+1)(\gamma_2+2)}\frac{\gamma_1\pi_{0i}\pi_{0k}}{(\gamma_1+1)},\Phi_i,\Phi_i,\Phi_k\right)\end{aligned}\tag{17}$$

Third, for $i\neq j\neq k$, we have:

$$\begin{aligned}&\mathbf{E}\left[\mathbf{E}_{x|G_{p(\mathbf{i})},\gamma(\mathbf{i})}\left[\left(\Phi_i\pi_{\mathbf{i}i}\right)\otimes\left(\Phi_j\pi_{\mathbf{i}j}\right)\otimes\left(\Phi_k\pi_{\mathbf{i}k}\right)\right]\right]\\ &=\mathbf{E}\left[T\left(\frac{\gamma_2^2\pi_{p(\mathbf{i})i}\pi_{p(\mathbf{i})j}\pi_{p(\mathbf{i})k}}{(\gamma_2+1)(\gamma_2+2)},\Phi_i,\Phi_j,\Phi_k\right)\right]\\ &=T\left(\frac{\gamma_2^2}{(\gamma_2+1)(\gamma_2+2)}\frac{\gamma_1^2\pi_{0i}\pi_{0j}\pi_{0k}}{(\gamma_1+1)(\gamma_1+2)},\Phi_i,\Phi_j,\Phi_k\right).\end{aligned}\tag{18}$$

Defining

$$\begin{aligned}S_3 :=& M_3-C_4\cdot S_1\otimes S_1\otimes S_1-C_5\cdot\mathfrak{S}_3\left[S_2\otimes M_1\right]\\ =&T\left(C_6\cdot\mathrm{diag}\left(\pi_0\right),\Phi,\Phi,\Phi\right)\end{aligned}\tag{19}$$

we solve C_4, C_5, C_6 as follows.

Note that for $i \neq j \neq k$, $[S_3]_{ijk} = 0$ and $[\mathfrak{S}_3[S_2 \otimes M_1]]_{ijk} = 0$. Thus

$$\begin{aligned} C_4 &= \frac{[M_3]_{ijk}}{[S_1]_i [S_1]_j [S_1]_k} \\ &= \frac{\gamma_2^2}{(\gamma_2+1)(\gamma_2+2)} \frac{\gamma_1^2 \pi_{0i}\pi_{0j}\pi_{0k}}{(\gamma_1+1)(\gamma_1+2)} \cdot \frac{1}{\pi_{0i}\pi_{0j}\pi_{0k}} \\ &= \frac{\gamma_2^2 \gamma_1^2}{(\gamma_2+1)(\gamma_2+2)(\gamma_1+1)(\gamma_1+2)} \end{aligned} \tag{20}$$

Similarly, for $i = j \neq k$, $[S_3]_{ijk} = 0$. Thus

$$\begin{aligned} C_5 &= \frac{[M_3]_{iik} - C_4 [S_1]_i [S_1]_i [S_1]_k}{[S_2]_{ii} [M_1]_k} \\ &= \frac{\gamma_2^2 \gamma_1 \pi_{0i}\pi_{0k} + \gamma_2(\gamma_1+2)\gamma_1 \pi_{0i}\pi_{0k}}{(\gamma_2+1)(\gamma_2+2)(\gamma_1+1)(\gamma_1+2)} \cdot \frac{(\gamma_1+1)(\gamma_2+1)}{\pi_{0i}\pi_{0k}(\gamma_2+\gamma_1+1)} \\ &= \frac{\gamma_2 \gamma_1 (\gamma_1+\gamma_2+2)}{(\gamma_1+2)(\gamma_2+2)(\gamma_1+\gamma_2+1)} \end{aligned} \tag{21}$$

Finally,

$$\begin{aligned} C_6 &= \frac{[M_3]_{iii} - C_4 [S_1]_i [S_1]_i [S_1]_i - 3C_5 [S_2]_{ii} [M_1]_i}{\gamma_i} \\ &= \frac{2\gamma_2^2 \pi_{0i} + 3\gamma_2(\gamma_1+2)\pi_{0i} + 2(\gamma_1+1)(\gamma_1+2)\pi_{0i}}{(\gamma_2+1)(\gamma_2+2)(\gamma_1+1)(\gamma_1+2)\gamma_1 \pi_{0i}} \\ &= \frac{6\gamma_1 + 6\gamma_2 + 2\gamma_1^2 + 2\gamma_2^2 + 3\gamma_2\gamma_1 + 4}{\gamma_1(\gamma_1+1)(\gamma_1+2)(\gamma_2+1)(\gamma_2+2)} \end{aligned} \tag{22}$$

■

C Spectral Algorithm for HDP

Algorithm 1 Spectral Algorithm for HDP

Require: Observations $\mathbf{x}$

1: **Inferring mixture number**
Using all leaf nodes $\mathbf{i}_j$ of the HDP tree compute the rank k of $\tilde{M}_1 := \left[M_1^{\mathbf{i}_1}, M_1^{\mathbf{i}_2}, \cdots M_1^{\mathbf{i}_{n^{L-2}}}\right]$.

2: **Moment estimation**
Compute moment estimates $\hat{M}_r^0$ and tensors $\hat{S}_r^0$.

3: **Dimensionality reduction and whitening**
Find $W \in \mathbb{R}^{d \times k}$ such that $W^T \hat{S}_2^0 W = I_k$.

4: **Tensor decomposition**
Obtain eigenvectors v_i and eigenvalues λ_i of $\hat{S}_3^0$.

5: **Reconstruction**

$$\text{Result set } \hat{\Phi} = \left\{ \lambda_i \frac{C_3}{C_6} \left(W^\dagger\right)^T v_i \right\}, \tag{23}$$

where C_3 and C_6 are coefficients defined in Lemma 3. See the Appendix for a derivation of (23).

D Proof of reconstruction formula

Dimensionality reduction and whitening To reduce the computational complexity and make the statistics at the root $\hat{S}_r^0$ orthogonally decomposable, it is desirable to construct a low-dimensional

orthogonal representation of S_2^0. Specifically, we find $W \in \mathbb{R}^{d\times k}$ such

$$W^\top S_2^0 W = I_k$$

where I_k is a k-by-k identity matrix. A stable yet computationally efficient way to find W is to compute the top k eigenvectors $U \in \mathbb{R}^{d\times k}$ and eigenvalues $\Sigma \in \mathbb{R}^{k\times k}$, and to use $W = U\Sigma^{-1/2}$ as whitening matrix.

Reconstruction In this step the recovered eigenvectors are transformed back to the original space, and Φ is then recovered. Concentration of measure guarantees for the reconstructed $\hat{\Phi}$ are given in the next section. Note that due to the amount of noise in real-world data, not all columns in Φ might lie on the probability simplex. This can be addressed by projection, i.e. by removing negative entries and normalizing the sum to 1. We show in Section 5 that even with this modification, the recovered $\hat{\Phi}$ is accurate.

Define $\tilde{\Phi} := T(\sqrt{C_3}\mathrm{diag}(\sqrt{\gamma}), \Phi)$ and perform SVD on $\tilde{O}$ such that $\tilde{\Phi} = USV^T$. Then

$$S_2 = \tilde{\Phi}\tilde{\Phi}^T = USS^TU^T \tag{24}$$

$$S_3 = T\left(\frac{C_6}{C_3\sqrt{C_3}}\mathrm{diag}\left(\frac{1}{\sqrt{\gamma}}\right), \tilde{\Phi}, \tilde{\Phi}, \tilde{\Phi}\right) \tag{25}$$

Since in our algorithm $W^TS_2W = I$, $W = US^{-1}$ Thus

$$T(S_3, W, W, W) \tag{26}$$

$$= T(\frac{C_6}{C_3\sqrt{C_3}}\mathrm{diag}\left(\frac{1}{\sqrt{\gamma}}\right), W^T\tilde{\Phi}, W^T\tilde{\Phi}, W^T\tilde{\Phi})) \tag{27}$$

$$= T(\frac{C_6}{C_3\sqrt{C_3}}\mathrm{diag}\left(\frac{1}{\sqrt{\gamma}}\right), V^T, V^T, V^T) \tag{28}$$

The corresponding eigenvalues λ_i and eigenvectors v_i, with some permutation π, are

$$\lambda_i = s_i\frac{C_6}{C_3\sqrt{C_3}}\mathrm{diag}\left(\frac{1}{\sqrt{\gamma_{\pi_i}}}\right) \tag{29}$$

$$v_i = s_iV^Te_{\pi_i} \tag{30}$$

Therefore

$$W^+ = \left(W^TW\right)^{-1}W^T = SU^T \tag{31}$$

$$\left(W^+\right)^T v_i = s_i(US)V^Te_{\pi_i} = s_i\sqrt{C_3}\sqrt{\gamma_{\pi_i}}\Phi_{\pi_i} \tag{32}$$

$$\Phi_{\pi_i} = \frac{\left(W^+\right)^T v_i}{s_i\sqrt{C_3}\sqrt{\gamma_{\pi_i}}} \tag{33}$$

Rearranging Equation 29, we have

$$s_i = \frac{C_6}{C_3\sqrt{C_3}\sqrt{\gamma_{\pi_i}}\lambda_i}; \quad \Phi_{\pi_i} = \frac{\lambda_i(W^+)^Tv_i}{C_6/C_3} \tag{34}$$

E Tensor decomposition solvers

Robust tensor power method Since the estimation of S_k^0 might not be perfect, and tensors might not be orthogonally decomposable, a robust tensor power algorithm is employed here to recover the robust eigenvectors. The key step is iterate the update:

$$\theta_t \leftarrow v/\left\|v\right\| \text{ and } v \leftarrow T(\hat{S}_3^0, W, W\theta_{t-1}, W\theta_{t-1}) \tag{35}$$

until convergence to a eigenvector. In this method, s random vectors are initialized as random draws from unit sphere in $\mathbb{R}^k$, and each of them is updated t times according to (35). We then update the one with the largest eigenvalue for another t times to ensure convergence and return this one. It is straightforward to parallelize the updates of these vectors to speed up. See [11] for a detailed description and perturbation analysis of the robust tensor power method.

Alternating Least Square Another commonly used method for solving tensor decomposition is alternating least square method. The main idea is to concatenate the tensors into a matrix and then minimize the Frobenius norm of the difference:

$$\min \left\| [S_3(W,W,W)]_{(1)} - V\mathrm{diag}(\lambda)(V \odot V)^T \right\|_F \tag{36}$$

where the definition of operators used are:

$$S_{(1)} = \left[S[:,:,1] \;\; S[:,:,2] \cdots S[:,:,K] \right] \tag{37}$$

$$V \odot V = [v_1 \boxplus v_1 \;\; v_2 \boxplus v_2 \cdots v_K \boxplus v_K]. \tag{38}$$

The notation $\odot$ is the Khatri-Rao product and $\boxplus$ represents the Kronecker product. Taking the second and third V in $V\mathrm{diag}(V \odot V)^T$ as fixed, we get the closed form solution of the optimization problem as:

$$V\mathrm{diag}(\lambda) = [S_3(W,W,W)]_{(1)} (V \odot V) \left((V^T V). \wedge 2 \right)^\dagger$$

where the notation $.\wedge$ denoting point-wise power. By iteratively updating v_i until it converges, we solve the optimization problem in (36).

However, we found that the algorithm does not guarantee to converge with high dimensional data. According to experimental results, the objective function does not always decrease after each iteration. This is due to the fact that the objective function is not necessarily a convex function since there is a cubic function of the unknown parameters. To deal with the problem, following [14], alternating rank-1 update, which is an alternating version of the tensor power method, is guarantee to converge locally. Th update rule becomes

$$\tilde{v}_i = [S_3(W,W,W,)]_{(1)} (v_i \odot v_i)(v_i^T v_i)^{-2}, \tag{39}$$

$$\tilde{\lambda}_i = \|\tilde{v}_i\|_2, \quad v_i = \tilde{v}_i / \|\tilde{v}_i\|_2, \tag{40}$$

$\forall i \in [K]$. Iteratively updating the eigenvector until converges gives us the largest eigenvector with corresponding eigenvalue. Before deriving the second eigenvector, we deflate the original tensor by subtracting the mode-1 tensor constructed by the derived eigenvector and eigenvalue.

F Concentration of measure bounds

We derive theoretical guarantees for the spectral inference algorithms in an HDP. Specifically we provide guarantees for moments $M_r^{\mathbf{i}}$, tensors $S_r^{\mathbf{i}}$, and latent factors Φ. The technical challenge relative to conventional models is that the data are not drawn iid. Instead, they are drawn from a predefined hierarchy and they are only exchangeable within the hierarchy. We address this by introducing a more refined notion of effective sample size which borrows from its counterpart in particle filtering [15]. We define $n_{\mathbf{i}}$ to be the effective sample size, obtained by hierarchical averaging over the HDP tree. This yields

$$n_{\mathbf{i}} := \begin{cases} 1 & \text{for leaf nodes} \\ |c(\mathbf{i})|^2 \left[\sum_{\mathbf{j} \in c(\mathbf{i})} \frac{1}{n_{\mathbf{i}}} \right]^{-1} & \text{otherwise} \end{cases} \tag{41}$$

One may check that in the case where all leaves have an equal number of samples and where each vertex in the tree has an equal number of children $n_{\mathbf{i},1}$ is the overall sample size.

Theorem 4 *For any node* $\mathbf{i}$ *in an* L*-layer HDP with* r*-th order moment* $M_r^{\mathbf{i}}$ *and for any* $\delta \in (0,1)$ *the following bound holds for the tensorial reductions* $M_r(u) := T(M_r^{\mathbf{i}}, u, \cdots, u)$ *and its empirical estimate* $\hat{M}_r := T(\hat{M}_r^{\mathbf{i}}, u, \cdots, u)$.

$$\Pr\left\{ \sup_{u:\|u\|\le 1} \left| M_r(u) - \hat{M}_r(u) \right| \le \frac{2 + \sqrt{-\ln \delta}}{\sqrt{n_{\mathbf{i}}}} \right\} \ge 1 - \delta$$

As indicated, $n_{\mathbf{i}}$ plays the role of an effective sample size. Note that an unbalanced tree has a smaller effective sample size compared to a balanced one with same number of leaves.

Theorem 5 *Given an L-layer HDP with symmetric tensor $S_r^{\mathbf{i}}$. Assume that $\delta \in (0,1)$ and denote the tensorial reductions as before $S_r(u) := T(S_r^{\mathbf{i}}u, \cdots, u)$ and $\hat{S}_r(u) := T(\hat{S}_r^{\mathbf{i}}, u, \cdots, u)$. Then we have for $r \in \{2,3\}$*

$$\Pr\left\{\sup_{u:\|u\|\leq 1}\left|S_r - \hat{S}_r\right| \leq \epsilon_r\right\} \geq 1-\delta \tag{42}$$

where for some constant c

$$\epsilon_2 := 3n_{\mathbf{i}}^{-\frac{1}{2}}\left[2+\sqrt{\ln(3/\delta)}\right] \text{ and} \tag{43}$$

$$\epsilon_3 := cn_{\mathbf{i}}^{-\frac{1}{2}}\left[2+\sqrt{\ln(3/\delta)}\right]. \tag{44}$$

This shows that not only the moments but also the symmetric tensors directly related to the statistics Φ are directly available. The following theorem guarantees the accurate reconstruction of the latent feature factors in Φ. Again, a detailed proof is relegated to Appendix G.4.

Theorem 6 *Given an L-layer HDP with hyperparamter $\gamma_{\mathbf{i}}$ at node $\mathbf{i}$. Let $\sigma_k(\Phi)$ denote the smallest non-zero singular value of Φ, and ϕ_i denote the i-th column of Φ. For sufficiently large sample size, and for suitably chosen $\delta \in (0,1)$, i.e.*

$$3n_{\mathbf{i}}^{-\frac{1}{2}}\left[2+\sqrt{\ln(3/\delta)}\right] \leq \frac{C_3\gamma_0 \min_j \pi_{0j}\sigma_k(\Phi)^2}{6}$$

we have $\Pr\left\{\left\|\Phi_i - \hat{\Phi}_{\sigma(i)}\right\| \leq \epsilon\right\} \geq 1-\delta$ where

$$\epsilon := \frac{ck^3(\min_{\mathbf{i}}\gamma_{\mathbf{i}}+2)^2}{\delta \min_j \pi_{0j}\sigma_k(\Phi)^3} n_{\mathbf{i}}^{-\frac{1}{2}}\left[2+\sqrt{\ln(3/\delta)}\right]$$

Here $\left\{\hat{\Phi}_1, \hat{\Phi}_2, \cdots, \hat{\Phi}_k\right\}$ is the set Algorithm 1 returns, for some permutation σ of $\{1,2,\cdots,k\}$, $i \in 1,2,\cdots,k$, and some constant c.

The theorem gives the guarantees on l_2 norm accuracy for the reconstruction of latent factors. Note that all the bounds above are functions of the effective sample sizes $n_{\mathbf{i}}$. The latter are a function of both the number of data and the structure of the tree.

G Concentration of Measure

G.1 Effective sample size

In the following it will be useful to keep track of the explicit weighting inherent in the definition of the moments $M_r^{\mathbf{i}}$. In this context recall that $M_r^{\mathbf{i}} = \frac{1}{|c(\mathbf{i})|}\sum_{\mathbf{j}\in c(\mathbf{i})} M_r^{\mathbf{j}}$ and that furthermore for leaf nodes $M_r^{\mathbf{i}}$ is the weighted average over all combinations of occurring attributes.

Definition 7 (Effective sample size) *For any average $x := \sum_i \eta_i x_i$, we denote by $n_{\text{eff}} := \frac{\|\eta\|_1^2}{\|\eta\|_2^2}$ its effective sample size.*

To see that this definition is sensible, consider the case of $\eta_i = l^{-1}$ and $\eta \in \mathbb{R}^l$. In this case we obtain $n_{\text{eff}} = l$, as desired for even weighting.

Lemma 8 *Denote by $\eta_i \in \mathbb{R}^{l_i}$ normalized vectors with $\eta_{ij} \geq 0$ and $\|\eta_i\|_1 = 1$. Moreover, let $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$. Then the effective sample size of the concatenated vector $\eta := \uplus_i \lambda_i \eta_i$ satisfies*

$$\frac{1}{n_{\text{eff}}} = \sum_i \frac{\lambda_i^2}{n_{\text{eff},i}}$$

This follows by direct calculation. In particular, note that $\|\eta\|_1 = 1$. Hence $\|\eta\|_2^2 = \sum_i \lambda_i^2 \|\eta_i\|_2^2$. Taking the inverse yields the claim.

We now explicitly construct an auxiliary weighting vector $\eta^{(\mathbf{i},r)}$ of dimensionality $\rho^{(\mathbf{i},r)}$. At the leaf level we use a vector of dimensionality 1 and weights 1. As we ascend through the tree, all children are given weights $1/|c(\mathbf{i})|$ and a weighting vector $\eta^{(\mathbf{i},r)} = |c(\mathbf{i})|^{-1} \uplus_{\mathbf{j}\in c(\mathbf{i})} \eta^{(\mathbf{j},r)}$ is assembled. For convenience we will sometimes also make use of $d(\mathbf{i}, r)$, the set of all index vectors used in $\eta_{\mathbf{i}}$, which is the same as the number of documents under this node.

G.2 Proof of Theorem 4

Proof Recall that for both empirical estimate and expectation of moment at node $\mathbf{i}$, we have:

$$M_r^{\mathbf{i}} = \frac{1}{|c(\mathbf{i})|} \sum_{\mathbf{j}\in c(\mathbf{i})} M_r^{\mathbf{j}} = \sum_{\mathbf{s}\in d(\mathbf{i})} \eta_{\mathbf{s}}^{(\mathbf{i},r)} \varphi_r(x_{\mathbf{s}}) \tag{45}$$

Now define

$$\Xi[X] := \sup_{u:\|u\|\leq 1} \left| T(M_r^{\mathbf{i}}, u, \cdots, u) - T(\hat{M}_r^{\mathbf{i}}, u, \cdots, u) \right|.$$

The deviation between empirical average and expectation observed when using X. Then $\Xi[X]$ is concentrated. This follows from the inequality of [16] since for any $\mathbf{r} \in d(\mathbf{i}, k)$

$$|\Xi[X] - \Xi[(X \setminus \{x_{\mathbf{s}}\}) \cup \{x'_{\mathbf{s}}\}]| \leq \eta_{\mathbf{s}}^{(\mathbf{i},r)} \|\varphi_r(x_{\mathbf{s}}) - \varphi_r(x'_{\mathbf{s}})\| \leq \sqrt{2}\eta_{\mathbf{s}}^{(\mathbf{i},r)}. \tag{46}$$

Hence the random variable $\Xi[X]$ is concentrated in the sense that $\Pr\{\Xi[X] - \mathbf{E}_X[\Xi[X]] < \epsilon\} \geq 1 - \delta$ with $\delta = \exp\left(-\frac{\epsilon^2}{\|\eta^{(\mathbf{i},r)}\|_2^2}\right)$ or, in other words, $\epsilon = \|\eta^{(\mathbf{i},r)}\|_2 \sqrt{\ln(1/\delta)}$.

The next step is to bound the expectation of $\Xi[X]$. This is accomplished as follows:

$$\begin{aligned}
\mathbf{E}_X[\Xi[X]] \leq & \mathbf{E}_{X,X'} \left[\sup_{u:\|u\|\leq 1} \left| T(\hat{M}_r^{\mathbf{i}}, u, \cdots, u) - T(\tilde{M}_r^{\mathbf{i}}, u, \cdots, u) \right| \right] \\
= & \mathbf{E}_\sigma \mathbf{E}_{X,X'} \left[\sup_{u:\|u\|\leq 1} \left| \sum_{\mathbf{s}\in d(\mathbf{i})} \sigma_{\mathbf{s}} \eta_{\mathbf{s}}^{(\mathbf{i},r)} \left(T(\varphi_r(x_{\mathbf{s}}), u, \cdots, u) - T(\varphi_r(x'_{\mathbf{s}}), u, \cdots, u) \right) \right| \right] \\
\leq & 2\mathbf{E}_\sigma \mathbf{E}_X \left[\sup_{u:\|u\|\leq 1} \left| \sum_{\mathbf{s}\in d(\mathbf{i})} \sigma_{\mathbf{s}} \eta_{\mathbf{s}}^{(\mathbf{i},r)} T(\varphi_r(x_{\mathbf{s}}), u, \cdots, u) \right| \right] \\
\leq & 2\mathbf{E}_\sigma \mathbf{E}_X \left[\left\| \sum_{\mathbf{s}\in d(\mathbf{i})} \sigma_{\mathbf{s}} \eta_{\mathbf{s}}^{(\mathbf{i},r)} \varphi_r(x_{\mathbf{s}}) \right\| \right] \\
\leq & 2\mathbf{E}_X \left[\mathbf{E}_\sigma \left[\left\| \sum_{\mathbf{s}\in d(\mathbf{i})} \sigma_{\mathbf{s}} \eta_{\mathbf{s}}^{(\mathbf{i},r)} \varphi_r(x_{\mathbf{s}}) \right\|^2 \right] \right]^{\frac{1}{2}} \leq 2\|\eta^{(\mathbf{i},r)}\|_2,
\end{aligned}$$

Here the first inequality follows from convexity of the argument. The subsequent equality is a consequence of the fact that X and X' are drawn from the same distribution, hence a swapping permutation with the ghost-sample leaves terms unchanged. The following inequality is an application of the triangle inequality. Next we use the Cauchy-Schwartz inequality, convexity and last the fact that $\|\varphi_r(x)\| \leq 1$. Combining both bounds yields $\epsilon_{\mathbf{i}} \geq \|\eta^{(\mathbf{i},r)}\|_2 \left(2 + \sqrt{\ln(1/\delta)}\right)$. For the definition of efficient number, since $\|\eta\|_1 = 1$, we have $n_{\mathbf{i},r} = 1/\|\eta^{(\mathbf{i},r)}\|_2^2$. Thus we obtained the theorem. ■

G.3 Proof of Theorem 5

Proof By theorem 4 and the definition of $S_2^{\mathbf{i}}$ and $S_3^{\mathbf{i}}$, the bounds for tensors can be easily obtained. For $S_2^{\mathbf{i}}$, we have:

$$\begin{aligned}\left\|\hat{S}_2^{\mathbf{i}} - S_2^{\mathbf{i}}\right\| &= \left\|\hat{M}_2^{\mathbf{i}} - C_2 \cdot \hat{S}_1^{\mathbf{i}} \otimes \hat{S}_1^{\mathbf{i}} - M_2^{\mathbf{i}} + C_2 \cdot S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}}\right\| \\ &\leq \left\|\hat{M}_2^{\mathbf{i}} - M_2^{\mathbf{i}}\right\| + C_2 \left\|\hat{S}_1^{\mathbf{i}} \otimes \hat{S}_1^{\mathbf{i}} - S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}}\right\| \\ &\leq \left\|\hat{M}_2^{\mathbf{i}} - M_2^{\mathbf{i}}\right\| + \left\|\hat{S}_1^{\mathbf{i}} - S_1^{\mathbf{i}}\right\| \left\|\hat{S}_1^{\mathbf{i}} - S_1^{\mathbf{i}}\right\| + 2\|S_1\| \left\|\hat{S}_1 - S_1\right\| \\ &\leq \left[\|\eta^{(\mathbf{i},2)}\|_2 + 2\|\eta^{(\mathbf{i},1)}\|_2\right] \left(2 + \sqrt{\ln(3/\delta)}\right) + \left[\|\eta^{(\mathbf{i},1)}\|_2 \left(2 + \sqrt{\ln(3/\delta)}\right)\right]^2\end{aligned} \tag{47}$$

For S_3, expanding the

$$\begin{aligned}\left\|\hat{S}_3^{\mathbf{i}} - S_3^{\mathbf{i}}\right\| &\leq \|M_3^{\mathbf{i}} - C_4 \cdot S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}} \otimes S_1^{\mathbf{i}} - C_5 \cdot \mathfrak{S}_3\left[S_2^{\mathbf{i}} \otimes S_1^{\mathbf{i}}\right] - \hat{M}_3^{\mathbf{i}} + C_4 \cdot \hat{S}_1 \otimes \hat{S}_1 \otimes \hat{S}_1 \\ &\quad + C_5 \cdot \mathfrak{S}_3\left[\hat{S}_2 \otimes \hat{S}_1\right]\| \\ &\leq \left\|M_3^{\mathbf{i}} - \hat{M}_3^{\mathbf{i}}\right\| + C_4\left(\left\|S_1^{\mathbf{i}} - \hat{S}_1^{\mathbf{i}}\right\|^3 + 3S_1^{\mathbf{i}}\left\|S_1^{\mathbf{i}} - \hat{S}_1^{\mathbf{i}}\right\|^2\right) + 3C_4 {S_1^{\mathbf{i}}}^2 \left\|S_1^{\mathbf{i}} - \hat{S}_1^{\mathbf{i}}\right\| \\ &\quad + 3C_5\left\|S_2^{\mathbf{i}} - \hat{S}_2^{\mathbf{i}}\right\| \left\|S_1^{\mathbf{i}} - \hat{S}_1^{\mathbf{i}}\right\| + 3C_5\left(S_1^{\mathbf{i}}\left\|S_2^{\mathbf{i}} - \hat{S}_2^{\mathbf{i}}\right\| + \cdot S_2^{\mathbf{i}}\left\|S_1 - \hat{S}_1\right\|\right) \\ &\leq \left[\|\eta^{(\mathbf{i},3)}\|_2 + 3C_5\|\eta^{(\mathbf{i},2)}\|_2\right]\left(2 + \sqrt{\ln(3/\delta)}\right) + (C_4 + 3C_5)\|\eta^{(\mathbf{i},1)}\|_2\left(2 + \sqrt{\ln(3/\delta)}\right) \\ &\quad + 3(C_4 + C_5)\|\eta^{(\mathbf{i},1)}\|_2^2\left(2 + \sqrt{\ln(3/\delta)}\right)^2 + C_4\|\eta^{(\mathbf{i},1)}\|_2^3\left(2 + \sqrt{\ln(3/\delta)}\right)^3\end{aligned} \tag{48}$$

■

G.4 Proof of Theorem 6

Proof We follow the similar steps for complexity analysis in [4]. Using the definition of tensor structure we stated in Lemma 2, we define:

$$\tilde{\Phi} := T(\sqrt{C_3^0} \cdot \sqrt{\pi_0}, \Phi), \tag{49}$$

where $H = [H_1, H_2, \cdots, H_k]$ is a normalized vector. So we have:

$$\begin{aligned}\sqrt{C_3^0 \min_j \pi_{0j}}\sigma_k(\Phi) \leq \sigma_k\left(\tilde{\Phi}\right) \leq 1, \\ \sigma_1\left(\tilde{\Phi}\right) \leq \sigma_1(\Phi)\sqrt{C_3^0\gamma_0}.\end{aligned} \tag{50}$$

Thus, S_2^0 and S_3^0 can be transformed to:

$$\begin{aligned}S_2^0 &= T(C_3^0 \cdot \mathrm{diag}(\pi_0), \Phi, \Phi) = \tilde{\Phi}\tilde{\Phi}^T \\ S_3^0 &= T(\frac{C_6^0}{C_3^0\sqrt{C_3^0}}\mathrm{diag}\left(\frac{1}{\sqrt{\pi_0}}\right), \tilde{\Phi}, \tilde{\Phi}, \tilde{\Phi})\end{aligned} \tag{51}$$

Let λ_i be the singular values of S_3, we have:

$$\lambda_i = \frac{C_6^0}{C_3^0\sqrt{C_3^0}}\frac{1}{\sqrt{\pi_{0i}}} \tag{52}$$

such that, for $i \in [K]$,

$$\frac{C_6^0}{C_3^0\sqrt{C_3^0}}\frac{1}{\sqrt{\gamma_0}} \leq \lambda_i \leq \frac{C_6^0}{C_3^0\sqrt{C_3^0}}\frac{1}{\sqrt{\min_j \pi_{0j}}}. \tag{53}$$

Next, as in Algorithm 1, $\hat{W}$ whitens a rank k approximation to S_2^0, U. Here we define $\hat{S}_{2,k}^0$ to be the best rank k approximation of $\hat{S}_2^0$. Besides, define:

$$M := W^T\tilde{\Phi}, \quad \hat{M} = \hat{W}^T\tilde{\Phi}. \tag{54}$$

Lemma 9 *(Lemma C.1 in [4]) Let Π_W be the orthogonal projection onto the range of W and Π be the orthogonal projection onto the range of 0. Suppose $\left\|\hat{S}_2^0 - S_2^0\right\| \leq \sigma_k\left(S_2^0\right)/2$,*

$$\begin{aligned}
&\left\|M^0\right\| = 1, \quad \left\|\hat{M}^0\right\| \leq 2, \quad \left\|\hat{W}\right\| \leq \frac{2}{\sigma_k\left(\tilde{\Phi}\right)}, \\
&\left\|\hat{W}^+\right\| \leq 2\sigma_1\left(\tilde{\Phi}\right), \quad \left\|W^+\right\| \leq 3\sigma_1\left(\tilde{\Phi}\right) \\
&\left\|M^0 - \hat{M}^0\right\| \leq \frac{4}{\sigma_k\left(\tilde{\Phi}\right)^2}\left\|\hat{S}_2^0 - S_2^0\right\|, \\
&\left\|\hat{W}^+ - W^+\right\| \leq \frac{6\sigma_1\left(\tilde{\Phi}\right)}{\sigma_k\left(\tilde{\Phi}\right)^2}\left\|\hat{S}_2^0 - S_2^0\right\|, \\
&\left\|\Pi - \Pi_W\right\| \leq \frac{4}{\sigma_k\left(\tilde{\Phi}\right)}\left\|\hat{S}_2^0 - S_2^0\right\|.
\end{aligned} \tag{55}$$

By using the upper bound of λ_i and **Lemma C.1** and **Lemma C.2** in [4], we get:

Lemma 10 *Suppose $E_{S_2^0} \leq \sigma_k\left(S_2^0\right)/2$. For $\|\theta\| = 1$, we have:*

$$\begin{aligned}
&\left\|T\left(S_3^0, W, W, W\theta\right) - T\left(\hat{S}_3^0, \hat{W}, \hat{W}, \hat{W}\theta\right)\right\| \\
&\leq c\left(\frac{C_6\left\|S_2^0 - \hat{S}_2^0\right\|}{C_3^0\sqrt{C_3^0}\sqrt{\min_j \pi_{0j}}\sigma_k\left(\tilde{\Phi}\right)^2} + \frac{\left\|S_3^0 - \hat{S}_3^0\right\|}{\sigma_k\left(\tilde{\Phi}\right)^3}\right)
\end{aligned} \tag{56}$$

Following the similar steps in **Lemma C.3** in [4], we have

Lemma 11 *(SVD Accuracy) Suppose $\left\|S_2^0 - \hat{S}_2^0\right\| \leq \sigma_k\left(S_2^0\right)/2$, with probability greater than $1-\delta'$*

$$\|v_i - \hat{v}_i\| \leq c\frac{k^3 C_3^0\sqrt{C_3^0\gamma_0}}{\delta' C_6^0}c_1 \tag{57}$$

where

$$c_1 = \frac{C_6^0\left\|S_2^0 - \hat{S}_2^0\right\|}{C_3^0\sqrt{C_3^0}\sqrt{\min_j \pi_{0j}}\sigma_k\left(\tilde{\Phi}\right)^2} + \frac{\left\|S_3^0 - \hat{S}_3^0\right\|}{\sigma_k\left(\tilde{\Phi}\right)^3} \tag{58}$$

Combine everything together, we have:

Lemma 12 *(Lemma C.6 in [4]) Suppose $\left\|S_2^0 - \hat{S}_2^0\right\| \leq \sigma_k\left(S_2^0\right)/2$, with probability greater than $1-\delta'$, we have*

$$\left\|\Phi_i - \frac{1}{\hat{Z}_i}\left(\hat{W}^+\right)^T\hat{v}_i\right\| \leq c\frac{k^3\gamma_0}{\delta'\min_j \pi_{0j}^2 C_3}\epsilon \tag{59}$$

where

$$\epsilon = \frac{\left\|S_2^0 - \hat{S}_2^0\right\|}{\sigma_k\left(\Phi\right)^2} + \frac{C_3^0\left\|S_3^0 - \hat{S}_3^0\right\|}{C_6^0\sigma_k\left(\Phi\right)^3} \tag{60}$$

Using the bounds for tensor in Theorem 5, we finish the proof. ■