Sparse Parallel Training of Hierarchical Dirichlet Process Topic Models

Alexander Terenin
Imperial College London (UK)
A.TERENIN17@IMPERIAL.AC.UK

Måns Magnusson
Aalto University (FI)
MANS.MAGNUSSON@AALTO.FI

Leif Jonsson
Ericsson (SE)
LEIF.JONSSON@ERICSSON.COM

Abstract

Nonparametric extensions of topic models such as Latent Dirichlet Allocation, including Hierarchical Dirichlet Process (HDP), are often studied in natural language processing. Training these models generally requires use of serial algorithms, which limits scalability to large data sets and complicates acceleration via use of parallel and distributed systems. Most current approaches to scalable training of such models either don’t converge to the correct target, or are not data-parallel. Moreover, these approaches generally do not utilize all available sources of sparsity found in natural language – an important way to make computation efficient. Based upon a representation of certain conditional distributions within an HDP, we propose a doubly sparse data-parallel sampler for the HDP topic model that addresses these issues. We benchmark our method on a well-known corpora (PubMed) with 8m documents and 768m tokens, using a single multi-core machine in under three days.

1 Introduction

Topic models are a widely-used class of models in natural language processing that allow practitioners to identify latent semantic themes occurring in large bodies of text. Hierarchical Bayesian discrete mixture models such as Latent Dirichlet Allocation (LDA) [4] and its many nonparametric extensions [5,12,20,23,24] are ubiquitous within the field. These models usually combine categorical likelihoods with conjugate Dirichlet or Dirichlet process priors, and are trained via various forms of Bayesian learning. To train these models in massively parallel settings, the training algorithm needs to satisfy two requirements.

1. Expose sufficient parallelism that the hardware can take advantage of.
2. Utilize sparsity found in natural language [31] to control memory requirements and computational complexity.

In this work, we focus on the Hierarchical Dirichlet process (HDP) topic model of Teh et al. [24], which we review in Section 2. This model is perhaps the simplest non-trivial extension of LDA to the nonparametric setting – its parallel implementation provides a blueprint for designing massively parallel training algorithms in more complicated settings such as topic models with Pitman-Yor priors [23], nonparametric dynamic topic models [1], tree-based extensions [5,12], and others.

*Website: HTTPS://AVT.IM/
Parallel approaches to training HDPs have been previously introduced by a number of authors, including Newman et al. [19], Wang et al. [28], Williamson et al. [29], and Chang and Fisher [8]. Gal and Ghahramani [10] have pointed out some of these methods can suffer from load-balancing issues which limit their available parallelism and scalability. The largest-scale benchmark of parallel HDP training performed thus far to our awareness is by Chang and Fisher [8] on the 100m-token NYTimes corpora.

Our contributions are as follows. We propose an augmented representation of the HDP under which the topic indicators can be sampled in parallel over documents. We prove that the global topic distribution $\Psi$, which is assigned a Griffiths-Engen-McCloskey (GEM) [22] prior under this representation, is conditionally conjugate given an auxiliary parameter $l$. We develop fast sampling schemes for $\Psi$ and $l$, thereby directly extending the parallel partially collapsed Gibbs sampler of Magnusson et al. [16] and Terenin et al. [25] from LDA to HDP. The proposed algorithm is doubly sparse: it has per-iteration complexity which depends on the minima of two sparsity terms, and thus takes advantage of both document-topic and topic-word sparsity simultaneously.

### 2 Partially collapsed Gibbs sampling for hierarchical Dirichlet processes

The Hierarchical Dirichlet Process topic model [24] begins with a global distribution $\Psi$ over topics. Documents are assumed exchangeable – for each document $d$, the associated topic distribution $\theta_d$ follows a Dirichlet process centered at $\Psi$. Each topic is associated with a distribution of tokens $\phi_k$. Within each document, tokens are assumed exchangeable (bag of words) and assigned to topic indicators $z_{i,d}$. For a given data set, we observe the tokens $w_{i,d}$.

Summarizing everything, we arrive at the GEM representation of a hierarchical Dirichlet process, given by equation (19) of Teh et al. [24] as

$$
\Psi \sim \text{GEM}(\gamma) \\
\theta_d \mid \Psi \sim \text{DP}(\alpha, \Psi) \\
z_{i,d} \sim \text{Discrete}(\theta_d) \\
w_{i,d} \sim \text{Discrete}(\phi_{z_{i,d}})
$$

where $\alpha, \beta, \gamma$ are prior hyperparameters.

#### 2.1 Intuition and augmented representation

At a high level, our strategy for constructing a scalable sampler is as follows. Conditional on $\Psi$, the likelihood in equation (1) is the same as that of LDA. Using this observation, the Gibbs step for $z$, which is the largest component of the model, can be handled efficiently by leveraging insights on sparse parallel sampling from the well-studied LDA literature [15][16][25][30]. For this strategy to succeed, we need to ensure that all Gibbs steps involved in the HDP under this representation are analytically tractable and can be computed efficiently. For this, the representation needs to be slightly modified.

To begin, we integrate each $\theta_d$ out of the model, which by conjugacy [3] yields a Pólya sequence for each $z_d$. By definition, given in Appendix A, this sequence is a mixture distribution with respect to a
set of Bernoulli random variables $b_d$, each representing whether $w_{i,d}$ was drawn from $\Psi$ or from a repeated draw in the Pólya urn. Thus, the HDP model admits the representation

$$\Psi \sim \text{GEM}(\gamma)$$  \hspace{1cm} (4)
$$b_{i,d} \sim \text{Ber}\left(\frac{1}{\alpha + \psi_i} \right)$$ \hspace{1cm} (5)
$$w_{i,d} \mid b_d, \Psi \sim \text{PS}(\Psi, b_d) \quad \phi_k \sim \text{Dir}(\beta)$$ \hspace{1cm} (6)

where $\text{PS}(\Psi, b_d)$ is defined in Appendix A. This representation defines a posterior distribution over $z, \Phi, \Psi, b$ for the HDP. To derive a Gibbs sampler, we proceed to calculate its full conditionals.

### 2.2 Full conditionals for $z, \Phi, \Psi$, and $b$

The full conditionals $z \mid \Phi, \Psi$ and $\Phi \mid z, \Psi$, with $b$ marginalized out, are essentially identical to those in partially collapsed LDA \[16, 25\]. They are

$$\mathbb{P}(z_{i,d} = k \mid z_{-i,d}, \Phi, \Psi) \propto \phi_{k,v(i)} \left[ \alpha \Psi_k + m_{d,k}^{-1} \right] \quad \phi_k \mid z \sim \text{Dir}(\beta + n_k)$$  \hspace{1cm} (7)

where $m_{d,k}^{-1}$ denotes the document-topic sufficient statistic with index $i$ removed. Note that the number of possible topics and full conditionals $\phi_k \mid z$ here is countably infinite. The full conditional for each $b_{i,d}$ is

$$\mathbb{P}(b_{i,d} = 1 \mid z_d, \Psi, b_{-i,d}) = \frac{\alpha \Psi_{z_{i,d}}}{\alpha \Psi_{z_{i,d}} + \sum_{i=1}^{\infty} \mathbb{I}_{z_{i,d}}(z_{i,d})}.$$  \hspace{1cm} (8)

The derivation, based on a direct application of Bayes’ Rule with respect to the probability mass function of the Pólya sequence, is given in Appendix A.

### 2.3 The full conditional for $\Psi$

To derive the full conditional for $\Psi$, we examine the prior and likelihood components of the model. It is shown in Appendix A that the likelihood term $z_d \mid b_d, \Psi$ may be written

$$p(z_d \mid b_d, \Psi) = \prod_{i=1}^{N_d} \prod_{k=1}^{D} \frac{1}{l_k} \mathbb{I}_{z_{i,d}}(z_{i,d}) \prod_{i=1}^{N_d} \prod_{k=1}^{\infty} \Psi_k^{l_k(z_{i,d})}.$$  \hspace{1cm} (9)

The first term is a multiplicative constant independent of $\Psi$ and vanishes via normalization. Thus, the full conditional $\Psi \mid z, b$ depends on $z$ and $b$ only through the sufficient statistic $l$ defined by

$$L_k = \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{I}_{z_{i,d}}(z_{i,d})$$  \hspace{1cm} (10)

and so we may suppose without loss of generality that the likelihood term is categorical. Under these conditions, we prove, via a measure-theoretic argument, that the full conditional for $\Psi$ admits a stick-breaking representation.

**Proposition 1.** Without loss of generality, suppose we have

$$\Psi \sim \text{GEM}(\gamma) \quad x \mid \Psi \sim \text{Discrete}(\Psi).$$  \hspace{1cm} (11)

Then $\Psi \mid x$ is given by

$$\Psi_k = \xi_k \prod_{i=1}^{k-1} (1 - \xi_i) \quad \xi_k \sim \text{B}(a_k, b_k) \quad a_k = 1 + L_k \quad b_k = \gamma + \sum_{i=k+1}^{\infty} l_i$$  \hspace{1cm} (12)

where $l$ is the empirical distribution of $x$.

**Proof.** Appendix B.  \hspace{1cm} ■

Putting these ideas together, we define an infinite-dimensional parallel Gibbs sampler for the HDP.
Algorithm 1 (Infinite-dimensional partially collapsed Gibbs sampling for the HDP topic model).

Repeat until convergence.

- Sample $\phi_k \sim \text{Dir}(n_k + \beta)$ in parallel over topics for $k = 1, \ldots, \infty$.
- Sample $z_{i,d} \propto \phi_k(v(i)) \alpha \Psi_k + \phi_k(v(i)) m_{d,k}^{-1}$ in parallel over documents for $d = 1, \ldots, D$.
- Sample $b_{i,d}$ according to Equation (8) in parallel over documents for $d = 1, \ldots, D$.
- Sample $\Psi$ according to Proposition 1.

Algorithm 1 is completely parallelizable, but cannot be implemented as stated due to the infinite number of full conditionals for $\Phi$, as well as the infinite product used in sampling $\Psi$. We proceed to bypass these issues by introducing an approximate finite-dimensional sampling scheme.

2.4 Finite-dimensional sampling of $\Psi$ and $\Phi$

By way of assuming $\Psi \sim \text{GEM}(\gamma)$, an HDP assumes that an infinite number of topics are present in the model a priori, with the number of tokens per topic decreases rapidly with the topic’s index in a manner controlled by $\gamma$. Thus, we expect under the model that a topic with a sufficiently large index $K$ will contain no tokens with high probability.

We thus propose to approximate $\Psi$ in a finite-dimensional manner by projecting its tail onto a single flag topic $K^*$, which stands for all topics not explicitly represented as part of the computation. Following Ishwaran and James [13], this can be done by deterministically setting $\varsigma_{K^*} = 1$ in Proposition 1, which, as part of its proof, implies the approximation is convergent and well-posed.

Once this is done, $\Psi$ becomes a finite vector of length $K^*$, and only $K^*$ rows of $\Phi$ need to be explicitly instantiated as part of the computation. This allows the algorithm to be defined on a fixed finite state space, simplifying bookkeeping and implementation.

From a computational efficiency perspective, the resulting value $K^*$ takes the place of $K$ in partially collapsed LDA – however, it cannot be interpreted as the number of topics in the sense of LDA. Indeed, LDA implicitly assumes that $\Psi = U(1, \ldots, K)$ deterministically – i.e. that every topic is assumed a priori to contain the same number of tokens, whereas the HDP model assumes otherwise by letting $\Psi \sim \text{GEM}(\gamma)$.

If we allow the state space to be resized when topic $K^*$ is sampled, then following Papaspiliopoulos and Roberts [21] it is possible to develop truncation schemes which introduce no error. Since this results in more complicated bookkeeping, we instead fix $K^*$ and defer such considerations to future work. We recommend setting $K^*$ to be sufficiently large that it does not significantly affect the model’s behavior, which can be checked by tracking the number of tokens assigned to topic $K^*$.

2.5 Sparse sampling of $\Phi$ and $z$

To be efficient, a topic model needs to utilize the sparsity found in natural language as much as possible. In our case, the two main sources of sparsity are as follows.

1. **Document-topic sparsity**: most documents will only contain a handful of topics.
2. **Topic-word sparsity**: most word types will not be present in most topics.

From this, we expect the document-topic sufficient statistic $m$ and topic-word sufficient statistic $n$ to contain many zeros. We seek to use this to reduce sampling complexity. Our starting point is the Poisson Pólya Urn-based sampler of Terenin et al. [25], which presents a Gibbs sampler for LDA with computational complexity that depends on the minima of two sparsity coefficients representing document-topic and topic-word sparsity – such algorithms are termed *doubly sparse*. The key idea is to approximate the Dirichlet full conditional for $\phi_k$ with a Poisson Pólya Urn distribution defined by

$$\phi_{k,v} = \frac{\varphi_{k,v}}{\sum_{v=1}^{V} \varphi_{k,v}}$$

for $v = 1, \ldots, V$. This distribution is discrete, so $\Phi$ becomes a sparse matrix. The approximation is accurate even for small values of $n_{k,v}$, and approximation error is proven by Terenin et al. [25] to vanish for large data sets in the sense of weak convergence.
In the case where $\beta$ is uniform, we can further use sparsity to accelerate sampling $\varphi_{k,v}$. Since a sum of Poisson random variables is Poisson, we can split $\varphi_{k,v} = \varphi_{k,v}^{(j)} + \varphi_{k,v}^{(n)}$. We then sample $\varphi_{k,v}^{(j)}$ sparsely by introducing a Poisson process and sampling its points uniformly, and sample $\varphi_{k,v}^{(n)}$ sparsely by iterating over nonzero entries of $n$.

For $z$, the full conditional

$$\mathbb{P}(z_{i,d} = k \mid z_{-i,d}, \Phi, \Psi) \propto \phi_{k,v(i)} \left[ \alpha \Psi_k + m_{d,k}^{-1} \right] \sum_{j,k} \phi_{k,v(j)} \left[ \alpha \Psi_k + \phi_{k,v(j)} m_{d,k}^{-1} \right]$$

is essentially identical to the one in Pólya Urn LDA – the only difference is the presence of $\Psi_k$. As $\Psi_k$ only enters the expression through component $(a)$ and is identical for all $z_{i,d}$, it can be absorbed at each iteration directly into the Walker alias tables [15, 26] – see Magnusson et al. [16] for details. Component $(b)$ can be computed efficiently by utilizing sparsity of $\Phi$ and $\Psi$ and iterating over whichever has fewer non-zero entries – see Terenin et al. [25] for details.

### 2.6 Direct sampling of $l$

Rather than sampling $b$, whose size will grow linearly with the number of documents, we introduce a scheme for sampling the sufficient statistic $l$ directly. Observe that

$$l_k = \sum_{d=1}^{D} \sum_{i=1}^{N_d} 1_{z_{i,d} = k} = \sum_{d=1}^{D} \sum_{i=1}^{N_d} 1_{b_{i,d} = 1}.$$  

By definition of $b_{i,d}$, we have

$$\sum_{i=1}^{N_d} 1_{b_{i,d} = 1} = \sum_{j=1}^{m_{d,k}} b_{j,d,k} \sim \text{Ber} \left( \frac{\Psi_k \alpha}{\Psi_k \alpha + j - 1} \right).$$

Summing this expression over documents, we obtain

$$l_k = \sum_{j=1}^{\max_d m_{d,k}} c_{j,k} \sim \text{Bin} \left( D_{k,j}, \frac{\Psi_k \alpha}{\Psi_k \alpha + j - 1} \right)$$

where $D_{k,j}$ is the total number of documents with $m_{d,k} \geq j$. Since $m_{d,k} = 0$ for all topics $k$ without any tokens assigned, we only need to sample $I$ for topics that have tokens assigned to them. The complexity of sampling $l$ via this expression is constant with respect to the number of documents, and depends instead on the maximum number of tokens per document.

To handle the bookkeeping necessary for computing $D_{k,j}$, we introduce a sparse matrix $d$ of size $K \times \max_d N_d$ whose entries $d_{k,p}$ are the number of documents for topic $k$ that have a total of $p$ topic indicators assigned to them. We increment $d$ once $z_{d}$ has been sampled by iterating over non-zero elements in $m_{d}$. We then compute $D_{k,j}$ as the reverse cumulative sum of the rows of $d$.

Putting all of these ideas together, we obtain the following algorithm.

**Algorithm 2** (Poisson Pólya urn partially collapsed Gibbs sampling for the HDP topic model).

*Repeat until convergence.*

- **Sample** $\phi_k \sim \text{PPU}(n_k + \beta)$ in parallel over topics for $k = 1, \ldots, K^*$.
- **Sample** $z_{i,d} \propto \phi_{k,v(i)} \alpha \Psi_k + \phi_{k,v(i)} m_{d,k}^{-1}$ in parallel over documents for $d = 1, \ldots, D$.
- **Sample** $l_k$ according to Equation (18) in parallel over topics for $k = 1, \ldots, K^*$.
- **Sample** $\Psi$ according to Proposition 7 except with $\varsigma_{K^*} = 1$.

Algorithm 2 is sparse, massively parallel, and contains no infinite computations in any of its steps. The Gibbs steps for $\Phi$ converge in distribution [25] to the true Gibbs steps as $N \to \infty$, and the Gibbs step for $\Psi$ converges almost surely [13] to the true Gibbs step as $K^* \to \infty$. 

5
2.7 Computational complexity

We now examine the per-iteration computational complexity of Algorithm 2. To proceed, we fix $K^*$ and maximum document size $\max_{d} N_d$, and relate the vocabulary size $V$ with the number $N$ of total words as follows.

**Assumption (Heaps’ Law).** The number of unique words in a corpus follows Heaps’ law \[11\] $V = \xi N^\zeta$ with constants $\xi > 0$ and $\zeta < 1$.

The per-iteration complexity of Algorithm 2 is equal to the sum of the per-iteration complexity of sampling its components. The sampling complexities of $\Psi$ and $l$ are constant with respect to the number of tokens, and the sampling complexity of $\Phi$ has been shown by Magnusson et al. \[16\] to be negligible under the given assumptions. Thus, it suffices to consider $z$.

At a given iteration, let $K_{d(i)}^{(m)}$ be the number of existing topics in document $d$ associated with word token $i$, and let $K_{v(i)}^{(\Phi)}$ be the number of nonzero topics in the row of $\Phi$ corresponding to word token $i$.

It follows immediately from the argument given by Terenin et al. \[25\] that the per-iteration complexity of sampling a single topic indicator $z_i$ is \[\mathcal{O}\left[\min\{K_{d(i)}^{(m)}, K_{v(i)}^{(\Phi)}\}\right]. \tag{19}\]

Algorithm 2 is thus a doubly sparse algorithm.

3 Performance results

To study the performance of the partially collapsed sampler – Algorithm 2 – we implemented it in Java using the open-source Mallet \[17\] topic modeling framework. We ran it on the CGCBIB, NeurIPS, and PubMed corpora \[2\] which are summarized in Table 2. Prior hyperparameters were set to $\alpha = 0.1$, $\beta = 0.01$, $\gamma = 1$. We set $K^* = 1000$ and observed no tokens ever allocated to topic $K^*$.

Data was preprocessed with default Mallet stop-word removal, minimum document size of 10, and a rare word limit of 10. Following standard practice for the HDP topic model \[8, 24, 28, 29\], the algorithm was initialized with one topic. Total runtime for each experiment is given in Table 2.

To assess Algorithm 2 in a small-scale setting and compare it to the widely-studied fully collapsed direct assignment sampler of Teh et al. \[24\], which is not parallel, we ran 100 000 iterations of both methods on CGCBIB. We selected this corpora because it was among the larger corpora for which it was feasible to run our direct assignment reference implementation within one week. Computation was performed on an Apple MacBook Pro laptop, which has a 6-core i7-8850H CPU with 16GB RAM.

Trace plots for the log marginal likelihood for $z$ given $\Psi$ and number of active topics, i.e. those topics assigned at least one token, can be seen in Figure 1(a) and Figure 1(b), respectively. The direct assignment algorithm converges slower, but achieves a slightly better local optima in terms of marginal log-likelihood, compared to our method. This indicates that it may stabilize at a different local optima, and may represent a potential limitation of the partially collapsed sampler in settings where non-parallel methods are practical.

To better understand distributional differences between the algorithms, we examined the number of tokens per topic, which can be seen in Figure 1(c). The partially collapsed sampler is seen to assign

| Corpus     | $V$   | $D$   | $N$   | Iterations | Threads | Runtime  |
|------------|-------|-------|-------|------------|---------|----------|
| CGCBIB     | 6 079 | 5 940 | 570 370 | 100 000    | 12      | 2.7 hours|
| NeurIPS    | 12 419 | 1 499 | 1 894 051 | 255 500    | 8       | 24 hours |
| PubMed     | 89 987 | 8 199 999 | 768 434 972 | 20 000    | 20      | 65.2 hours|

Table 2: Corpora used in experiments, together with compute configuration.
Figure 1: Trace plots for log-likelihood, number of active topics, and additional metrics for CGCBIB, NeurIPS, and PubMed. Per-iteration scale is used for CGCBIB and PubMed, real-time scale is used for NeurIPS. Algorithms used are partially collapsed HDP for all corpora, direct assignment HDP for CGCBIB, and subcluster split-merge HDP for NeurIPS. Where multiple runs are present, individual traces are partially transparent, and their mean is opaque.

more tokens to smaller topics, indicating that it stabilizes at a local optima with slightly broader semantic themes.

To visualize the effect this has on the topics, we examined the most common words for each topic. Since the algorithms generate too many topics to make full examination practical, we instead compute a quantile summary with five topics per quantile. This is computed by ranking all topics by number of tokens, choosing the five closest topics to the 100%, 75%, 50%, 25%, 5% quantiles in the ranking, and computing their top words. This gives a representative view of the algorithm’s output for large, medium, and small topics. Results may be seen in Appendix C – we find the direct assignment and partially collapsed samplers to be largely comparable, with substantial overlap in top words for common topics.

To assess Algorithm 2 in a more demanding setting and compare against previous parallel state-of-the-art, we ran it and the parallel subcluster split-merge algorithm of Chang and Fisher [8] on the NeurIPS corpora. The subcluster split-merge algorithm is designed to converge with fewer iterations, but is more costly to run per iteration, so for both algorithms we used a fixed computational budget of 24 hours of wall-clock time. Computation was performed on a system with a 4-core i7-4790 CPU and 8GB RAM.

Results can be seen in Figure 1(d) – note that the subcluster split-merge algorithm is parametrized using sub-topic indicators and sub-topic probabilities, so its numerical log-likelihood values are not directly comparable to ours and should be interpreted purely to assess convergence. Algorithm 2 stabilizes much faster with respect to both the number of active topics in Figure 1(d), and marginal log-likelihood in Figure 1(e). The subcluster split-merge algorithm is only able to add new topics one-at-a-time, whereas our algorithm can create multiple new topics per iteration – we hypothesize this difference leads to faster convergence for Algorithm 2.

In Figure 1(f), we observe that the amount of compute time per iteration increases substantially for the subcluster split-merge method as it discovers more topics. For Algorithm 2 this stays approximately constant for its entire runtime.

To evaluate the topics produced by the algorithms, we again examined the most common words for each topic via a quantile summary, which may be seen in Appendix D. We find that the subcluster split-merge algorithm appears to generate topics with slightly more semantic overlap compared to Algorithm 2, but has otherwise produced comparable output.
Finally, to assess scalability, we ran 20,000 iterations of Algorithm 2 on PubMed, which contains 768m tokens. To our knowledge, this dataset is an order of magnitude larger than any datasets used in previous MCMC-based approaches for the HDP. Computation was performed on a standard compute node at the Triton scientific compute cluster at Aalto University, which has 2x10-core Xeon E5 2680v2 CPUs with hyper-threading disabled and 64GB of RAM. The experiment was repeated four times to assess variability. Marginal log-likelihood and the number of active topics can be seen in Figure 1(g) and Figure 1(h).

To evaluate the topics discovered by the algorithm, we examined their most common words – these may seen for every topic in Appendix E. We find by inspection that essentially all topics generated by the algorithm look reasonable and compelling. In particular, there appear to be virtually no duplicate topics, and few topics that contain multiple clearly-unrelated semantic concepts which should be split up. This suggests that the algorithm captures the underlying semantics well. We suspect the topic estimates for PubMed are particularly sharp compared to other corpora due to the large number of tokens present.

4 Discussion

In this work, we introduce the parallel partially collapsed Gibbs sampler – Algorithm 1 – for the HDP topic model, which converges to the correct target distribution. We propose a doubly sparse approximate sampler – Algorithm 2 – which allows the HDP to be implemented with per-token sampling complexity of $O\left[\min\{K^{(m)}_{\phi(i)}, K^{(\Psi)}_{\psi(i)}\}\right]$ which is the same as that of Pólya Urn LDA [25]. Compared to other approaches for the HDP, it offers the following improvements.

1. The algorithm is fully parallel in all of its steps.
2. The topic indicators $z$ utilize all available sources of sparsity to accelerate sampling.
3. All steps not involving $z$ are sampled with constant complexity with respect to data size.
4. The proposed approximate algorithm becomes exact as $N \to \infty$ and $K^* \to \infty$.

These improvements allow us to train the HDP via MCMC on larger corpora than any other published work we are aware of at the present. The data-parallel nature of our approach means that the amount of available parallelism increases with data size. This avoids the load-balancing-related scalability limitations of other methods that were recently pointed out by Gal and Ghahramani [10].

Nonparametric topic models are less straightforward to evaluate empirically than ordinary topic models. In particular, we found topic coherence scores [18] to be strongly affected by the number of active topics $K$, which causes preference for models with fewer topics and more semantic overlap per topic. We view development of summary statistics which are $K$-agnostic, as well as those measuring other aspects of topic quality such as overlap, to be a promising direction for future work. We are particularly interested in techniques that can be used to compare algorithms for sampling from the same model which are defined over fully disjoint state spaces, such as Algorithm 2 and the subcluster split-merge algorithm in Section 3.

Partially collapsed HDP can stabilize at a different local mode than fully collapsed HDP as proposed by Teh et al. [24]. There have been many attempts to improve mixing in that sampler [8] [13] [27], including the use of Metropolis-Hastings steps for jumping between modes [14] and related ideas. These techniques are largely complementary to our own, and can be explored in combination with the ideas presented here. HDP is a heavily multimodal target for which full posterior exploration is known to be difficult [8] [10], and sampling schemes are generally used more in the spirit of optimization than of traditional MCMC. We view MAP estimation-based analogues of ideas presented here as highly promising direction, since these may allow additional theoretical flexibility that may enable faster training.

Many of the ideas in this work are generic, and applicable to any topic model which is structurally similar to HDP’s GEM representation [24] given in Section 2. For example, one could consider using an informative prior for $\Psi$ in lieu of $GEM(\gamma)$, potentially improving convergence and topic quality, or developing parallel schemes for other nonparametric topic models such as Pitman-Yor models [23] or tree-based models [5] [12] [20]. We leave such considerations to future work.
By leveraging parallelism and sparsity, our proposed techniques allow researchers to scale non-parametric topic models to larger datasets than previously considered feasible for MCMC or other methods possessing similar convergence guarantees. Our algorithm for the HDP is truly parallel and scales to a 768m-token corpus (PubMed) on a single multicore machine in under three days. We hope these contributions enable wider use of nonparametric Bayesian methods for large collections of text.

Acknowledgments

The research was funded by the Academy of Finland (grants 298742, 313122). The calculations presented above were performed using compute resources within the Aalto University School of Science and Department of Computing at Imperial College London. We also acknowledge the support of Ericsson AB.

References

[1] A. Ahmed and E. P. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, 2010. Cited on page 1.
[2] L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows: in metric spaces and in the space of probability measures. Springer, 2008. Cited on page 12.
[3] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. The Annals of Statistics:353–355, 1973. Cited on page 2.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(1):993–1022, 2003. Cited on page 1.
[5] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The Nested Chinese Restaurant Process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57(2):7, 2010. Cited on pages 1, 8.
[6] V. I. Bogachev. Measure Theory: Volume I. Springer, 2007. Cited on page 12.
[7] V. I. Bogachev. Measure Theory: Volume II. Springer, 2007. Cited on page 12.
[8] J. Chang and J. W. Fisher III. Parallel sampling of HDPs using sub-cluster splits. In Advances in Neural Information Processing Systems, pages 235–243, 2014. Cited on pages 2, 6, 8.
[9] R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association, 64(325):194–206, 1969. Cited on page 13.
[10] Y. Gal and Z. Ghahramani. Pitfalls in the use of parallel inference for the Dirichlet process. In International Conference on Machine Learning, pages 208–216, 2014. Cited on pages 2, 8.
[11] H. S. Heaps. Information Retrieval: Computational and Theoretical Aspects. Academic Press, 1978. Cited on page 6.
[12] Y. Hu and J. Boyd-Graber. Efficient tree-based topic modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 275–279, 2012. Cited on pages 1, 8.
[13] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association, 96(455):161–173, 2001. Cited on pages 9, 14, 15.
[14] S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics, 13(1):158–182, 2004. Cited on page 8.
[15] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola. Reducing the sampling complexity of topic models. In Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining, pages 891–900, 2014. Cited on pages 2, 5.
[16] M. Magnusson, L. Jonsson, M. Villani, and D. Broman. Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models. Journal of Computational and Graphical Statistics, 27(2):449–463, 2018. Cited on pages 2, 3, 5, 9.
[17] A. K. McCallum. MABEL: A Machine Learning for Language Toolkit, 2002. URL: HTTP://MALLET.CS.UMASS.EDU. Cited on page 6.
[18] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272, 2011. Cited on page 8.
Appendix A: sufficiency of \( l \) and full conditional for \( b \)

Recall that the one-step-ahead conditional probability mass function in a Pólya sequence taking values in \( \mathbb{N} \) with concentration parameter \( \alpha \) and base probability mass function \( \Psi \) is

\[
p(z_i \mid z_{i-1}, \ldots, z_1, \Psi) = \sum_{j=1}^{i-1} \frac{1}{i-1+\alpha} \mathbb{1}_{z_j}(z_i) + \frac{\alpha}{i-1+\alpha} \Psi_{z_i}.
\]  

(20)

Introducing the random variable

\[
b_i \sim \text{Ber} \left( \frac{\alpha}{i-1+\alpha} \right)
\]

we can express the one-step-ahead conditional distribution as

\[
p(z_i \mid z_{i-1}, \ldots, z_1, b_i, \Psi) = \mathbb{1}_{b_i=0} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) + \mathbb{1}_{b_i=1} \Psi_{z_i}.
\]  

(22)

The joint probability mass function for \( z \mid b, \Psi \) is then

\[
p(z \mid b, \Psi) = \prod_{i=1}^{N} p(z_i \mid z_{i-1}, \ldots, z_1, b_i, \Psi) = \prod_{i=1}^{N} \left[ \sum_{j=1}^{i-1} \mathbb{1}_{b_i=0} \mathbb{1}_{z_j}(z_i) + \mathbb{1}_{b_i=1} \Psi_{z_i} \right].
\]  

(23)
Note that $\mathbb{1}_{b_i=0} = 1 \iff \mathbb{1}_{b_i=1} = 0$ and vice versa. Thus each term in the product for $z \mid b, \Psi$ only has one component, and we may express $z \mid b, \Psi$ as

$$p(z \mid b, \Psi) = \prod_{i=1}^{N} \prod_{b_i \neq 1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) \prod_{b_i = 1}^{\infty} \Psi_k(z_i)$$  \hfill (24)

where we have re-expressed the probability mass function of $\Psi$ in a form that emphasizes conjugacy. Thus for any prior, the posterior will only depend on the likelihood of the values of $z_i$ for which $b_i = 1$. The sufficient statistic is

$$l_k = \sum_{b_i = 1}^{N} \mathbb{1}_{z_i=k}.$$  \hfill (25)

Next, for a given $i' \in \{1, \ldots, N\}$, we can calculate the posterior of a component $b_{i'}$ as

$$P(b_{i'} = 1 \mid z, \Psi, b_{\neq i'}) \propto \left( \frac{\alpha}{i'-1 + \alpha} \right)^N \prod_{j=1}^{N} \frac{1}{j-1} \mathbb{1}_{z_j}(z_i) \prod_{b_i = 1}^{\infty} \Psi_i$$

$$\propto \alpha \Psi_i$$

$$P(b_{i'} = 0 \mid z, \Psi, b_{\neq i'}) \propto \left( \frac{i'-1}{i'-1 + \alpha} \right)^N \prod_{j=1}^{N} \frac{1}{j-1} \mathbb{1}_{z_j}(z_i) \prod_{b_i = 1}^{\infty} \Psi_i$$

$$\propto \sum_{i=1}^{i'-1} \mathbb{1}_{z_i}(z_i')$$  \hfill (29)

where we have divided both expressions by

$$\frac{1}{i'-1 + \alpha} \prod_{j=1}^{N} \frac{1}{j-1} \mathbb{1}_{z_j}(z_i) \prod_{b_i = 1}^{\infty} \Psi_i$$  \hfill (30)

which is constant with respect to $b_{i'}$. Note that full conditionally, we have $b_i \perp b_{i'}$ for $i \neq i'$. This gives the desired expressions and concludes the derivation.

**Appendix B: proof of Proposition**

Here we calculate the posterior distribution under a GEM prior and discrete likelihood. Since we are working in a nonparametric setting, we begin by making precise what we actually mean when we speak of conditional probability.

**Definition 2** (Regular conditional probability measure). Let $(\Theta, \mathcal{G})$ and $(Y, \mathcal{Y})$ be measurable spaces. A function $\pi_{\theta \mid y} : \Theta \times Y \rightarrow [0, 1]$, written $\pi_{\theta \mid y}(\cdot \mid y)$, is called a regular conditional probability measure iff the following hold.

1. For all $A_\theta \in \mathcal{G}$, the map $\pi_{\theta \mid y}(A_\theta \mid y)$ is $\mathcal{Y}$-measurable.

2. For all $y \in Y$, the map $\pi_{\theta \mid y}(\cdot \mid y)$ is a probability measure.

**Result 3.** Let $(\Theta, \mathcal{G})$ and $(Y, \mathcal{Y})$ be measurable spaces. Let $\pi_\theta : \Theta \rightarrow [0, 1]$ be a probability measure and $\pi_{y \mid \theta} : \mathcal{Y} \times \Theta \rightarrow [0, 1]$ be a regular conditional probability measure. Then there exists a unique probability measure $\pi_{y, \theta} : \Theta \times \mathcal{Y} \rightarrow [0, 1]$ such that

$$\pi_{y, \theta}(A_\theta \times A_y) = \int_{A_\theta} \pi_{y \mid \theta}(A_y \mid \theta) \, d\pi_\theta(\theta).$$  \hfill (31)

Moreover, for all measurable $f$ we have that

$$\int_{\Theta \times \mathcal{Y}} f(\theta, y) \, d\pi_{y, \theta}(\theta, y) = \int_{\Theta} \int_{\mathcal{Y}} f(\theta, y) \, d\pi_{y \mid \theta}(y \mid \theta) \, d\pi_\theta(\theta).$$  \hfill (32)
Moreover, for all measurable $f$ we have that
\[ \int_{\Theta \times Y} f(\theta, y) \, d\pi_{\theta,y}(\theta, y) = \int_{\Theta} \int_{Y} f(\theta, y) \, d\pi_{\theta | y}(\theta | y) \, d\pi_y(y). \]  

**Proof.** Ambrosio et al. [2], Theorem 5.3.1. See also Bogachev [6], Theorem 3.3.1.

**Result 4 (Disintegration Theorem).** Let $(\Theta, \mathcal{E})$ and $(Y, \mathcal{Y})$ be complete separable metric spaces. Let $\pi_{\theta,y} : \Theta \otimes \mathcal{Y} \to [0, 1]$ be a probability measure, and let $\pi_{\theta} = \pi_{\theta,y}(\cdot \times Y)$. Then there exists a $\pi_{\theta,y}$-a.e. unique regular conditional probability measure $\pi_{\theta | y} : \Theta \times Y \to [0, 1]$ such that
\[ \pi_{\theta,y}(A_{\theta} \times A_y) = \int_{A_{\theta}} \pi_{\theta | y}(A_y | y) \, d\pi_y(y). \]  

Moreover, for all measurable $f$ we have that
\[ \int_{\Theta \times Y} f(\theta, y) \, d\pi_{\theta,y}(\theta, y) = \int_{Y} \int_{\Theta} f(\theta, y) \, d\pi_{\theta | y}(\theta | y) \, d\pi_y(y). \]  

**Proof.** Ambrosio et al. [2], Theorem 5.3.1. See also Bogachev [6], Corollary 10.4.15.

Together, these results say that given a prior and a likelihood, the joint distribution is uniquely determined. If our state space is sufficiently regular, this in turn gives existence of the posterior distribution, and almost everywhere uniqueness with respect to the marginal measure. For conjugate models, we may calculate the posterior via finite-dimensional distributions, because the conditionals of the marginals are equal to the marginals of the conditionals. Our posterior’s finite-dimensional distributions are not analytic, so we instead work with an approximating sequence. Specifically, we show that if we introduce a sequence of approximating priors, and consider the induced sequence of approximating posteriors, then with sufficient continuity this sequence converges to the true posterior.

**Proposition 5.** Let $\Theta = \times_{k=1}^\infty \Theta_k$ be a complete separable metric space, let $\Theta$ be its cylindrical $\sigma$-algebra, and let $(Y, \mathcal{Y})$ be a countable discrete space. Let $\pi_\theta$ be a probability measure, and let $\pi_{\theta | y}$ be a regular conditional probability measure. Let $\pi_{\theta | y}$ be their associated regular conditional probability measure given by disintegration. Suppose we have a sequence of measures $(\pi^{(K)}_\theta)_{K \in \mathbb{N}}$ such that $\pi^{(K)}_\theta \to \pi_\theta$ weakly. For each $\pi^{(K)}_\theta$, let $\pi^{(K)}_{\theta | y}$ be the associated regular conditional probability measures given by disintegration. Assume for any bounded function $f(y)$ that the bounded function $\theta \mapsto \int_Y f(y) \, d\pi_{\theta | y}(y | \theta)$ is continuous. Then we have $\pi^{(K)}_{\theta | y} \to \pi_{\theta | y}$ weakly for all $y$.

**Proof.** By definition of weak convergence, we have
\[ \int_\Theta g(\theta) \, d\pi^{(K)}_\theta(\theta) \to \int_\Theta g(\theta) \, d\pi_\theta(\theta) \]  

for all bounded continuous $g$. Let $f(y)$ be a bounded function, and let $f(\theta)$ be a bounded continuous function. Taking $g(\theta) = f(\theta) \int_Y f(y) \, d\pi_{\theta | y}(y | \theta)$, which is bounded continuous by assumption, and applying Result 3 yields
\[ \int_{\Theta \times Y} f(\theta) f(y) \, d\pi^{(K)}_{\theta,y}(\theta, y) \to \int_{\Theta \times Y} f(\theta) f(y) \, d\pi_{\theta | y}(\theta, y) \]  

and in particular, by letting $f(\theta) f(y) = f(y)$ and integrating out $\theta$, we have $\pi^{(K)}_{\theta,y} \to \pi_{\theta,y}$ weakly. Every discrete metric space is complete, and since $Y$ is countable it is separable. Applying the Disintegration Theorem – Result 4 – yields
\[ \int_Y \int_{\Theta} f(\theta) f(y) \, d\pi^{(K)}_{\theta,y}(\theta | y) \, d\pi^{(K)}_y(y) \to \int_Y \int_{\Theta} f(\theta) f(y) \, d\pi_{\theta | y}(\theta | y) \, d\pi_y(y) \]  

for any bounded function $f$. Since $Y$ is a discrete space, we may rewrite the above integrals as summations with respect to the probability mass functions $p^{(K)}_y$ and $p_y$, which gives
\[ \sum_Y f(y) p^{(K)}_y(y) \int_{\Theta} f(\theta) \, d\pi^{(K)}_\theta(\theta | y) \to \sum_Y f(y) p_y(y) \int_{\Theta} f(\theta) \, d\pi_{\theta | y}(\theta | y). \]  

Taking $f(y)$ to be an indicator, and noting $\pi^{(K)}_y \to \pi_y$ implies $p^{(K)}_y \to p_y$ pointwise, we conclude
\[ \int_{\Theta} f(\theta) \, d\pi^{(K)}_{\theta | y}(\theta | y) \to \int_{\Theta} f(\theta) \, d\pi_{\theta | y}(\theta | y). \]  

Thus $\pi^{(K)}_{\theta | y} \to \pi_{\theta | y}$ weakly for all $y$ and the claim follows.
Our strategy now is to find a sequence of finite-dimensional priors which converge to our infinite-dimensional prior, use them to calculate a sequence of finite-dimensional posterior distributions, and evaluate the limit of that sequence. We proceed to define such a sequence, beginning with some preliminary definitions for the given problem.

**Definition 6** (Preliminaries). Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. Let \(S^{\infty}(\mathbb{N}) = \{x \in [0, 1]^\mathbb{N} : \sum_{i=1}^{\infty} x_i = 1\} \subset \ell^1\) be the probability simplex over \(\mathbb{N}\). Let \(N \in \mathbb{N}\). Let \(x \in \mathbb{N}^N\) be a vector. Let \(l \in N^\infty\) be its empirical distribution, i.e. \(l = \sum_{i=1}^{N} 1_{x_i}\), where \(1_{x_i}\) is equal to 1 for coordinate \(x_i\) and 0 for all other coordinate. Let \(\gamma > 0\). Recall that \(N^N\) and \(S^{\infty}(\mathbb{N})\) are complete separable metric spaces when endowed with the discrete and \(\ell^1\) metrics, respectively. We associate each random variable \(y : \Omega \rightarrow Y\) with its image probability measure \(\pi_y(A_y) = \mathbb{P}[y^{-1}(A_y)]\), and each conditional random variable \(\theta \mid y : \Omega \times Y \rightarrow \Theta\) with its image regular conditional probability measure \(\pi_y\mid\theta(A_y \mid \theta) = \mathbb{P}[(y \mid \theta) \in A_y] = \mathbb{P}[(y \mid \theta)^{-1}(A_y)]\), where the preimage is taken with respect to \(y\).

**Definition 7** (Discrete likelihood). For all \(\Psi \in S^{\infty}(\mathbb{N})\), define the conditional random variable \(x \mid \Psi : \Omega \times S^{\infty}(\mathbb{N}) \rightarrow \mathbb{N}^N\) by its probability mass function
\[
p(x \mid \Psi) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} \Psi_k(\mathbb{1}_{x_i}^{k}).
\]
We say \(x \mid \Psi \sim \text{Discrete}(\Psi)\).

**Definition 8** (GEM). Let \(\Psi : \Omega \rightarrow S^{\infty}(\mathbb{N})\) be a random variable defined by
\[
\Psi_k = \zeta_k \prod_{i=1}^{k-1} (1 - \zeta_i) \quad \zeta_k \sim \text{B}(1, \gamma).
\]
We say \(\Psi \sim \text{GEM}(\gamma)\).

We now introduce our finite-dimensional approximating prior and compute the posterior under it.

**Definition 9** (Finite GEM). Let \(\Psi : \Omega \rightarrow S^{\infty}(\mathbb{N})\) be a random variable defined by
\[
\Psi_k = \zeta_k \prod_{i=1}^{k-1} (1 - \zeta_i) \quad \zeta_k \sim \text{B}(1, \gamma) \quad \zeta_K = 1.
\]
We say \(\Psi \sim \text{FGEM}(\gamma, K)\).

We now introduce our finite-dimensional approximating prior and compute the posterior under it.

**Result 10.** Let \(x \mid \Psi \sim \text{Discrete}(\Psi)\) and \(\Psi \sim \text{FGEM}(\gamma, K)\). Then for any \(x\) with associated counts \(l\), we have that \(\Psi \mid x : \Omega \times \mathbb{N}^N \rightarrow S^{\infty}(\mathbb{N})\) is a conditional random variable defined by
\[
\Psi_k = \zeta_k \prod_{i=1}^{K-1} (1 - \zeta_i) \quad \zeta_k \sim \text{B}(\alpha_k(\Psi), \beta_k(\Psi)) \quad \zeta_K = 1
\]
where
\[
\alpha_k(\Psi) = 1 + l_k \quad \beta_k(\Psi) = \gamma + \sum_{i=k+1}^{K} l_i
\]
and \(l\) is the empirical distribution of \(x\). We say \(\Psi \mid x \sim \text{PGEM}(\gamma, l, K)\).

**Proof.** It is shown by Connor and Mosimann [9] that \(\Psi \sim \text{FGEM}(\gamma, K)\) is in fact a special case of the generalized Dirichlet distribution, which admits a general stick-breaking representation. Thus, its probability density function is
\[
f(\Psi) \propto \Psi^{-1}_{\nu_\infty} \prod_{k=1}^{K-1} \sum_{k'=1}^{K} \Psi_{k'}^{-1}
\]

13
which we have expressed in a simplified form. By conjugacy, for a given \( x \) and associated \( l \) the posterior probability density is

\[
f(\Psi \mid l) \propto \Psi^{(\gamma + l_K) - 1} \prod_{k=1}^{K-1} \left[ \sum_{k'=k}^{K} \Psi_{k'}^{(1 + l_{k'}) - 1} \right]^{\gamma + \sum_{i=k}^{K} l_i - [(1 + l_k) + \gamma + \sum_{i=k+1}^{K} l_i]} \tag{46}
\]

which is again a generalized Dirichlet admitting the necessary stick-breaking representation, which we have expressed in a form that emphasizes its posterior hyperparameters.

**Remark 11.** It is now clear that the assumption \( x \mid \Psi \sim \text{Discrete}(\Psi) \) is indeed taken without loss of generality, because if we instead took \( x \mid \Psi \) to be given by a Pólya sequence, then by sufficiency the prior-to-posterior map induced by disintegration would be identical.

**Lemma 12.** Let \( \Psi^{(K)} \sim \text{FGEM}(\gamma, K) \), let \( \Psi \sim \text{GEM}(\gamma) \), and let \( \pi^{(K)}_{\Psi} \) and \( \pi_{\Psi} \) be their respective image measures. Then we have \( \pi^{(K)}_{\Psi} \to \pi_{\Psi} \) weakly.

**Proof.** It is shown by Ishwaran and James [13] that \( \Psi^{(K)} \to \Psi \) almost surely, which implies weak convergence of their image measures by the Portmanteau Theorem.

**Lemma 13.** For any bounded continuous function \( f : \mathbb{N}^N \to \mathbb{R} \), the bounded real-valued function \( \Psi \to \int_{\mathbb{N}^N} f(x) \, d\pi_{x \mid \Psi}(x \mid \Psi) \) is continuous with respect to the \( \ell^1 \) topology.

**Proof.** We prove it is Lipschitz. Since \( x \mid \Psi \) is discrete, it admits a probability mass function, and we may write

\[
\int_{\mathbb{N}^N} f(x) \, d\pi_{x \mid \Psi}(x \mid \Psi) = \sum_{x \in \mathbb{N}^N} f(x) \prod_{i=1}^{N} \prod_{k=1}^{\infty} \Psi_k^{l_k(x_i)} = \sum_{x_N = 1}^{\infty} \cdots \sum_{x_1 = 1}^{\infty} f(x) \prod_{i=1}^{N} \Psi_{x_i}. \tag{47}
\]

Note first that, by the triangle inequality,

\[
\left| \prod_{i=1}^{N} \Psi_{x_i} - \prod_{i=1}^{N} \Psi'_{x_i} \right| = \left| \prod_{i=2}^{N} \Psi_{x_i} - \prod_{i=2}^{N} \Psi'_{x_i} \right| \tag{48}
\]

\[
= \left| \Psi_{x_1} \prod_{i=2}^{N} \Psi_{x_i} - \Psi'_{x_1} \prod_{i=2}^{N} \Psi_{x_i} \right| + \left| \Psi_{x_1} \prod_{i=2}^{N} \Psi_{x_i} - \Psi_{x_1} \prod_{i=2}^{N} \Psi'_{x_i} \right| \tag{49}
\]

\[
= \left| \Psi_{x_1} - \Psi'_{x_1} \right| \prod_{i=2}^{N} \Psi_{x_i} + \left| \Psi_{x_1} - \Psi'_{x_1} \right| \prod_{i=2}^{N} \Psi_{x_i} \tag{50}
\]

\[
\leq \left| \Psi_{x_1} - \Psi'_{x_1} \right| + \prod_{i=2}^{N} \Psi_{x_i} - \prod_{i=2}^{N} \Psi'_{x_i} \tag{51}
\]

\[
\leq \left| \Psi_{x_1} - \Psi'_{x_1} \right| + \prod_{i=2}^{N} \Psi_{x_i} - \prod_{i=2}^{N} \Psi'_{x_i} \tag{52}
\]

\[
\leq \sum_{i=1}^{N} |\Psi_{x_i} - \Psi'_{x_i}| \tag{53}
\]
where the second-to-last line follows because \(0 \leq \Psi_i \leq 1\) for all \(i\), and the last line follows by repeating the calculation inductively. Using this, we may write

\[
\left| \int_{[N]} f(x) \, d\pi_{x|\Psi}(x) - \int_{[N]} f(x) \, d\pi_{x|\Psi'}(x) \right| = \left| \sum_{x_N=1}^{\infty} \cdots \sum_{x_i=1}^{\infty} f(x) \prod_{i=1}^{N} \Psi_{x_i} - \sum_{x_N=1}^{\infty} \cdots \sum_{x_i=1}^{\infty} f(x) \prod_{i=1}^{N} \Psi'_{x_i} \right| = (54)
\]

\[
\leq \|f\|_{\infty} \sum_{x_N=1}^{\infty} \cdots \sum_{x_i=1}^{\infty} |N \prod_{i=1}^{N} \Psi_{x_i} - N \prod_{i=1}^{N} \Psi'_{x_i}| = (55)
\]

\[
\leq \|f\|_{\infty} \sum_{i=1}^{\infty} \sum_{x_i=1}^{\infty} \left| \Psi_{x_i} - \Psi'_{x_i} \right| = (56)
\]

\[
\leq \|f\|_{\infty} N \left\| \Psi - \Psi' \right\|_{\ell_1} = (57)
\]

which establishes the claim. ■

The sequence of random variables \((FGEM(\gamma, K))_{K \in \mathbb{N}}\) defined above gives the sequence of probability measures needed to apply Proposition 5 and we are now ready to prove the main result.

**Proposition 1.** Without loss of generality, suppose we have

\[
\Psi \sim GEM(\gamma) \quad x \mid \Psi \sim \text{Discrete}(\Psi).
\]

Then \(\Psi \mid x\) is given by

\[
\Psi_k = \varsigma_k \prod_{i=1}^{k-1} (1 - \varsigma_i) \quad \varsigma_k \sim \text{B}(a_k(\Psi), b_k(\Psi)) \quad a_k(\Psi) = 1 + l_k \quad b_k(\Psi) = \gamma + \sum_{i=k+1}^{\infty} l_i = (60)
\]

where \(l\) is the empirical distribution of \(x\).

**Proof.** In Lemma 12, we have given the convergent sequence \(\Psi^{(K)} \sim FGEM(\gamma, K)\) needed to apply Proposition 5 and have verified the continuity condition in Lemma 13. From the proposition, conclude that the sequence of regular conditional probability measures associated with the conditional random variables \(\Psi^{(K)} \mid x \sim PGEM(\gamma, l, K)\) given in Result 10 for any \(x\) converges weakly \(\pi_x\)-a.e. to the desired posterior distribution \(\Psi \mid x\). The limit of this sequence is known: it is shown by Ishwaran and James [13] to converge almost surely to the conditional random variable given in equation (60). Since almost sure convergence of random variables implies weak convergence of their image measures, the claim follows. ■
Appendix C: quantile summary of topics for *CGCBIB*

Here we display a multi-quantile summary for CGCBIB. This is computed by ranking all topics with at least 100 tokens by their total number of tokens, computing the $\varpi = 100\%, 75\%, 50\%, 25\%,$ and $5\%$ quantiles. We then compute the five topics closest to each quantile by number of tokens, and display their top-eight words.

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| elegans | elegans | elegans | gene | gene | gene |
| caenorhabditis | caenorhabditis | protein | genetic | sequence | mutants |
| nematode | caenorhabditis | development | protein | genes | mutator |
| results | gene | caenorhabditis | amino | caenorhabditis | elegans |
| found | function | nematode | cDNA | acid | alleles |
| show | proteins | studies | cdna | | |
| observed | required | model | | | |
| specific | show | | | | |

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| germ | ngl | emb | spe | wnt |
| germline | egg | temperature | sperm | mom |
| granules | laying | mutants | spermatozoa | signaling |
| cells | serotonin | sensitive | membrane | bar |
| embryos | cat | maternal | spermatids | pathway |
| somatic | dopamine | expression | spermatogenes | lin |
| line | mutants | embryonic | pseudopod | wrm |

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| vit | binding | kinesin | growth | eat |
| yolk | affinity | klp | survival | pharyngeal |
| vitellogenin | site | transport | mortality | pharynx |
| genes | activity | motor | population | pumping |
| yp | sites | ift | rate | inx |
| proteins | avermectin | cilia | populations | gap |
| vpe | elegans | dynein | parameter | feeding |
| lrp | membrane | movement | size | junctions |

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| mic | dom | innate | vha | ife |
| mel | effects | immune | atpase | cap |
| myosin | humic | immunity | subunit | eife |
| nmy | pyrene | abf | genes | capping |
| chain | effect | lys | vacuolar | cel |
| elongation | bioconcentrat | toll | subunits | gtp |
| rho | dissolved | antimicrobial | atpases | isoforms |
| phosphatase | substances | pathway | type | rma |

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| ubq | asp | da | ion | hcf |
| ge | salmonella | cl | diet | cehcf |
| tcp | poona | fli | relative | vp |
| footprints | enterica | gs | xpa | ldb |
| oscillin | clp | db | groups | cell |
| tlf | seroype | glu | carbon | mammalian |
| ubiquitin | necrotic | phospholipid | characteristi | phosphorylati |
| tata | mug | tg | atoms | neural |

| CGCBIB | $\varpi = 100\%$ | $\varpi = 75\%$ | $\varpi = 50\%$ | $\varpi = 25\%$ | $\varpi = 5\%$ |
|--------|------------------|------------------|------------------|------------------|------------------|
| 19 832 | 2 040 | 2 045 | 2 025 | 141 | 136 |
| ahh  | bhh  | chh  | dhh  | ehh  |
|------|------|------|------|------|
|     |      |      |      |      |
|      |      |      |      |      |

| elegans | elegans | elegans | mutations | elegans | gene |
| caenorhabditis | genetic | caenorhabditis | elegans | gene | sequence |
| protein | nematode | results | mutants | caenorhabditis | protein |
| gene | function | observed | genes | protein |
| proteins | development | high | caenorhabditis | amino |
| required | studies | type | mutant | cdna |
| show | model | effect | function | acid |

| 2 442 | 2 409 | 2 394 | 2 332 | 2 284 |
|-------|-------|-------|-------|-------|
| oxygen | germ | genes | elt | males |
| oxidative | cells | gene | ges | male |
| mev | line | expression | expression | sperm |
| sod | germline | elegans | gut | hermaphrodite |
| stress | glp | rai | gene | hermaphrodite |
| superoxide | gld | genome | gata | species |
| elegans | proliferation | analysis | specific | sex |
| mitochondrial | cell | rna | esterase | mating |

| 1 213 | 1 211 | 1 207 | 1 203 | 1 143 |
|-------|-------|-------|-------|-------|
| klp | ceh | proteins | smg | centrosome |
| kinesin | pha | protein | mrna | zyg |
| transport | pharyngeal | mass | merns | microtubule |
| motor | gene | dimensional | splicing | centrosomes |
| dynein | expression | ms | genes | tubulin |
| cilia | pharynx | gel | sr | embryos |
| unc | organ | fkh | electrophores | nonsense |
| | | | | tac |
| | | | | | |
| 664 | 658 | 642 | 638 | 630 |
| pkc | chemotaxis | mel | nsy | vha |
| tpa | elegans | mle | gcy | atpase |
| kinase | nacl | myosin | asymmetry | jnk |
| phorbol | defective | nmy | asel | isoforms |
| protein | nematode | elongation | left | subunit |
| apkc | chemotactic | rho | asel | ftu |
| pkcb | response | maternal | str | vacuolar |
| ester | stimuli | phosphatase | asymmetric | genes |

| 400 | 390 | 390 | 389 | 389 |
|------|------|------|------|------|
| irp | elements | bli | duplication | nex |
| iron | ltr | ia | sas | hcf |
| mammalian | retrotranspos | pc | gene | annexin |
| dehydrogenase | nase | kex | locus | cehf |
| zn | open | kpc | centrosome | mammalian |
| dehydrogenase | reading | proprotein | duplications | ahr |
| adh | frame | ida | zyg | mitochondrial |
| cytosolic | element | egl | cis | mitop |
Appendix D: quantile summary of topics for NeurIPS

Here we display a multi-quantile summary for NeurIPS. This is computed by ranking all topics with at least 100 tokens by their total number of tokens, computing the $\omega = 100\%, 75\%, 50\%, 25\%,$ and $5\%$ quantiles. We then compute the five topics closest to each quantile by number of tokens, and display their top-eight words.

| NeurIPS partially collapsed | $\omega = 100\%$ | $\omega = 75\%$ | $\omega = 50\%$ | $\omega = 25\%$ | $\omega = 5\%$ |
|----------------------------|------------------|------------------|------------------|------------------|------------------|
| system function number model training | 182 743 | 162 355 | 129 745 | 52 356 | 44 155 |
| information case result neural set | | | | | |
| approach result small result data | | | | | |
| set term values system test | | | | | |
| problem parameter order activity performance | | | | | |
| research neural large input number | | | | | |
| computer form effect pattern result | | | | | |
| single defined high function error | | | | | |

| NeurIPS partially collapsed | $\omega = 75\%$ | $\omega = 50\%$ | $\omega = 25\%$ | $\omega = 5\%$ |
|----------------------------|------------------|------------------|------------------|------------------|
| genetic delay bengio fig matching | 2 585 | 2 585 | 2 574 | 2 559 |
| algorithm bifurcation output properties model | | | | |
| population oscillation dependencies proc point | | | | |
| fitness point input step correspondenc | | | | |
| string stability experiment range match | | | | |
| generation fixed frasconi structure problem | | | | |
| bit limit term calculation set | | | | |
| function hopf information illinois set | | | | |

| NeurIPS partially collapsed | $\omega = 50\%$ | $\omega = 25\%$ | $\omega = 5\%$ |
|----------------------------|------------------|------------------|------------------|
| responses path male window agent | 1 310 | 1 309 | 1 309 |
| anastasio network normalization transition states | | | |
| pan packet feature connection loop | | | |
| rotation shortest mntn information history | | | |
| vestibular policy ntn temporal mdp | | | |

| NeurIPS partially collapsed | $\omega = 25\%$ | $\omega = 5\%$ |
|----------------------------|------------------|------------------|
| composite limited hypothesis tau cmm | 748 | 748 |
| mdp robot interconnect hypothesis speed | | |
| action camera fan mansour particle | | | |
| elemental set shunting growth particles | | | |
| payoff coordinates collectivity function pattern | | | |
| solution basis linear stem method | | | |
| mdt ritter unit large card | | | |

| NeurIPS partially collapsed | $\omega = 5\%$ |
|----------------------------|------------------|
| morph minimal visualization periodic machine | 396 |
| kernel root high period capacity | | |
| parent biases low coefficient path | | |
| human attribute diagram primitive trouble | | |
| busey remove visualizing homogeneous high | | |
| similar remelhart graphic tst task | | |
| exemplar row fund mhaskar increasing measures | | |
| Network | Model | Data | Function | Bound | Training | Input | Parameter | Network | Dimension |
|---------|-------|------|----------|-------|----------|-------|-----------|---------|-----------|
| Learning | Input | Parameter | Algebra | Learning | Function | Neural | Unit | Data | Operation |
| Neural | Weight | System | Mixture | Result | Function | Number | |
| Algorithm | Output | Visual | Gaussian | Set | |

| Learning | Movement | Motion | Learning | Cell |
|----------|----------|--------|----------|------|
| Critic   | Visual   | Unit   | Algorithm | Correlation |
| Function | Vector   | Direction | Action | Neuron |
| Actor    | Image    | Model | Advantage | Model |
| Algorithm | Model | Stage | System | Unit |
| System | Location | Input | Function | Interaction |
| Control | Eye | Network | Policy | Firing |
| Model | Map | Cell | Control | Set |

| Cell | Model | Form | Component |
|------|-------|------|-----------|
| Spike | Response | Word | Algorithm |
| Neural | Unit | Phone | Sources |
| Function | Escape | Input | Analysis |
| Response | Firing | Interneuron | Network | Data |
| Model | Result | Cockroach | System | Noise |
| Point | Transfer | Leg | Training | Orientation |
| Fixed | Sorting | Input | Meaning | Spatial |

| Aspect | Element | Network | Input | Traffic |
|--------|---------|---------|-------|---------|
| Object | Pairing | Neural | Unit | Waiting |
| View   | Grouping | Constraint | Spike | Elevator |
| Node   | Group   | Match   | Layer | Appeared |
| Learning | Saliency | Learn | Learning | Application |
| Network | Contour | Problem | Model | Compared |
| Weight | Computation | Initial | Predict | Department |
| Equation | Optimal | Row | Prediction | Found |

| Input | Network | Network | Network | Network |
|-------|---------|---------|---------|---------|
| Output | Neural | Symbol | Equation | Function |
| Activation | Task | VTP | Neuron | Adaptation |
| Data   | Link   | Learning | Moment | Algorithm |
| Encoded | Food   | Phrases | Neural | Prediction |
| Function | Nodes | Sentences | Approximation | Projection |
| Hidden | Output | VPP | Ohira | Neural |
| Model | Recurrent | Classification | Stochastic | Training |
Appendix E: topics produced by Algorithm 2 on PubMed

Here we show top eight words for each topic together with total number of tokens assigned, which is shown at the top of each table. We display all topics containing at least eight unique word tokens.

| Topic | Tokens |
|-------|--------|
| care  | 47322709 |
| age   | 40229486 |
| model | 34685122 |
| cell  | 30795166 |
| gene  | 30707144 |
| health |        |
| risk  |        |
| data  |        |
| expression |    |
| protein |       |
| patient |       |
| children |     |
| system |        |
| growth |        |
| dna    |        |
| medical |       |
| year  |        |
| time  |        |
| protein |       |
| expression |    |
| research |       |
| women |        |
| analysis |      |
| factor |        |
| sequence |      |
| clinical |       |
| patient |       |
| effect |        |
| receptor |      |
| genes  |        |
| system |        |
| factor |        |
| test  |        |
| kinase |        |
| rna    |        |
| system |        |
| factor |        |
| test  |        |
| kinase |        |
| rna    |        |
| cost   |        |
| population |    |
| field  |        |
| beta  |        |
| region |        |
| PubMed |        |
| cell  | 28510997 |
| cancer | 27277306 |
| patient | 26709116 |
| rat   | 26408263 |
| cell  | 25200662 |
| il    |        |
| tumor |        |
| patient |      |
| treatment |    |
| receptor |      |
| electron |      |
| cd    |        |
| patient |       |
| mg    |        |
| effect |        |
| muscle |        |
| mice  |        |
| carcinoma |     |
| drug  |        |
| neuron |        |
| tissue |        |
| antigen |       |
| cell  |        |
| effect |        |
| brain  |        |
| fiber  |        |
| human |        |
| breast |        |
| therapy |       |
| activity |      |
| rat    |        |
| lymphocytes |    |
| survival |      |
| dose  |        |
| stimulation |    |
| development |    |
| immune |        |
| tumour |        |
| day   |        |
| induced |       |
| microscopy |     |
| PubMed |        |
| patient | 24856624 |
| patient | 24750437 |
| blood | 24607618 |
| patient | 24482090 |
| infection |      |
| surgery |        |
| artery |        |
| pressure |      |
| disease |        |
| virus  |        |
| complication |     |
| heart  |        |
| flow  |        |
| clinical |      |
| hiv   |        |
| surgical |       |
| coronary |      |
| min   |        |
| diagnosis |     |
| strain |        |
| treatment |       |
| ventricular |     |
| effect |        |
| lesion |        |
| infected |      |
| year   |        |
| myocardial |     |
| exercise |      |
| brain  |        |
| patient |        |
| postoperative |     |
| cardiac |        |
| arterial |      |
| syndrome |      |
| positive |       |
| operation |       |
| left  |        |
| heart  |        |
| imaging |        |
| viral  |        |
| ca    | 22095623 |
| effect | 21838239 |
| receptor | 21363408 |
| channel | 20887061 |
| cell  | 20828980 |
| acid  |        |
| interaction |    |
| assay  |        |
| fetal  |        |
| alpha  |        |
| concentration |     |
| compound |      |
| detection |      |
| infant |        |
| antibody |       |
| na    |        |
| site  |        |
| system |        |
| concentration |    |
| gel    |        |
| PubMed |        |
| rat   | 20106260 |
| cell  | 19788488 |
| effect | 18675096 |
| liver  | 17163327 |
| mice  | 16440018 |
| dose  |        |
| hip   |        |
| disease |       |
| genes  |        |
| rat    |        |
| drug  |        |
| year  |        |
| acute  |        |
| dna    |        |
| enzymes |       |
| mg    |        |
| implant |       |
| chronic |       |
| polymorphism |     |
| activity |       |
| acid  |        |
| enzyme |        |
| liver  |        |
| analysis |      |
| concentration |    |
| rat    |        |
| enzymes |       |
| synthesis |      |
| PubMed | 16 136 164 | 14 201 063 | 13 706 016 | 13 191 158 | 13 105 245 |
|--------|------------|------------|------------|------------|------------|
|        | effect     | diet       | patient    | strain     | protein    |
|        | platelet   | weight     | disease    | plant      | membrane   |
|        | induced    | intake     | gastric    | growth     | cell       |
|        | oxide      | food       | asthma     | acid       | domain     |
|        | rat        | body       | test       | bacteria    | binding    |
|        | cell       | effect     | pylori     | activity    | receptor   |
|        | endothelial| acid       | arthritis  | cell       | lipid      |
|        | activity   | vitamin    | chronic    | species     | membranes  |

| PubMed | 12 705 261 | 12 624 252 | 10 422 885 | 9 850 167 | 7 027 660 |
|--------|------------|------------|------------|------------|------------|
|        | insulin    | species    | exposure   | skin       | level      |
|        | glucose    | population | concentration | patient | patient       |
|        | diabetes   | infection  | iron       | eyes       | ml         |
|        | cholesterol| animal     | level      | eye        | control    |
|        | level      | egg        | water      | retinal    | serum      |
|        | diabetic   | host       | effect     | laser      | plasma     |
|        | plasma     | parasite   | exposed    | visual     | factor     |
|        | lipoprotein | malaria    | lead       | corneal    | concentration |

| PubMed | 6 130 945 | 644 182 | 2 264 | 1 325 | 104 |
|--------|------------|--------|------|------|-----|
|        | dental     | sleep  | ppr  | pac  | feather |
|        | oral       | caffeine | csc  | foal | tieg |
|        | teeth      | tea    | stretch | cpr  | sorghum |
|        | tooth      | effect | pthrp | pacap | coi |
|        | periodontal| theophylline | response | edm  | phycocyanin |
|        | treatment  | night  | br    | speck | vanx |
|        | salivary   | coffee | gei   | branchial | midrib |
|        | gland      | green  | pth   | lth   | ifi |

| PubMed | 104 |
|--------|-----|
|        | steer |
|        | mca |
|        | persistency |
|        | buckwheat |
|        | dnak |
|        | eset |
|        | branding |
|        | akr |