Parallel Sampling of HDPs using Sub-Cluster Splits

Authored by:

John W. Fisher III
Jason Chang

Abstract

We develop a sampling technique for Hierarchical Dirichlet process models. The parallel algorithm builds upon [Chang & Fisher 2013] by proposing large split and merge moves based on learned sub-clusters. The additional global split and merge moves drastically improve convergence in the experimental results. Furthermore, we discover that cross-validation techniques do not adequately determine convergence, and that previous sampling methods converge slower than were previously expected.

1 Paper Body

Hierarchical Dirichlet Process (HDP) mixture models were first introduced by Teh et al. [2]. HDPs extend the Dirichlet Process (DP) to model groups of data with shared cluster statistics. Since their inception, HDPs and related models have been used in many statistical problems, including document analysis [2], object categorization [3], and as a prior for hidden Markov models [4]. The success of HDPs has garnered much interest in inference algorithms. Variational techniques [5, 6] are often used for their parallelization and speed, but lack the limiting guarantees of Markov chain Monte Carlo (MCMC) methods. Unfortunately, MCMC algorithms tend to converge slowly. In this work, we extend the recent DP Sub-Cluster algorithm [1] to HDPs to accelerate convergence by inferring ?sub-clusters? in parallel and using them to propose large split moves. Extensions to the HDP are complicated by the additional DP, which violates conjugacy assumptions used in [1]. Furthermore, split/merge moves require computing the joint model likelihood, which, prior to this work, was unknown in the common Direct Assignment HDP representation [2]. We discover that significant overlap in cluster distributions necessitates new global split/merge moves that change all clusters simultaneously. Our experiments on synthetic and real-world data validate the improved convergence of the proposed method. Additionally, our analysis of joint summary statistics suggests that other MCMC methods may converge prematurely in finite time.

2

Related Work

The seminal work of [2] introduced the Chinese Restaurant Franchise (CRF) and the Direct Assignment (DA) sampling algorithms for the HDP. Since then, many alternatives have been developed. Because HDP inference often extends methods from DPs, we briefly discuss relevant work on both models that focus on convergence and scalability. Current methods are summarized in Table 1. Simple Gibbs sampling methods, such as CRF or DA, may converge slowly in complex models. Works such as [11, 12, 13, 14] address this issue in DPs with split/merge moves. Wang and Blei [7] developed the only split/merge MCMC method for HDPs by extending the Sequentially Allocated Merge-Split (SAMS) algorithm of DPs developed in [13]. Unfortunately, reported results in [7] only show a marginal improvement over Gibbs sampling. Our experiments suggest that this is likely due to properties of the specific sampler, and that a different formulation significantly improves convergence. Additionally, SAMS cannot be parallelized, and is therefore only tested on a corpus with 263K words. By designing a parallel algorithm, we test on a corpus of 100M words. 1

Table 1: Capabilities of MCMC Sampling Algorithms for HDPs CRF [2] DA [2] SAMS [7] FSD [4] Hog-Wild [8] Super-Cluster [9] Proposed Infinite Model X X X ? X X X MCMC Guarantees X X X X ? X X Non-Conjugate Priors ? ? ? X ? ? X Parallelizable ? ? ? X X X X Local Splits/Merges ? ? X ? ? ? X Global Splits/Merges ? ? ? ? ? ? X ? potentially possible with some adapatation of the DP Metropolis-Hastings framework of [10].

There has also been work on parallel sampling algorithms for HDPs. Fox et al. [4] generalizes the work of Ishwaran and Zarepour [15] by approximating the highest-level DP with a finite symmetric Dirichlet (FSD). Iterations of this approximation can be parallelized, but fixing the model order is undesirable since it no longer grows with the data. Furthermore, our experiments suggest that this algorithm exhibits poor convergence. Newman et al. [8] present an alternative parallel approximation related to Hog-Wild Gibbs sampling [16, 17]. Each processor independently runs a Gibbs sampler on its assigned data followed by a resynchronization step across all processors. This approximation has shown to perform well on cross-validation metrics, but loses the limiting guarantees of MCMC. Additionally, we will show that cross-validation metrics are not suitable to analyze convergence. An exact parallel algorithm for DPs and HDPs was recently developed by Willamson et al. [9] by grouping clusters into independent super-clusters. Unfortunately, the parallelization does not scale well [18], and convergence is often impeded [1]. Regardless of exactness, all current parallel sampling algorithms exhibit poor convergence due to their local nature, while split/merge proposals are essentially ineffective and cannot be parallelized. 2.1

DP Sub-Clusters Algorithm

The recent DP Sub-Cluster algorithm [1] addresses these issues by combining non-ergodic Markov chains into an ergodic chain and proposing splits from learned sub-clusters. We briefly review relevant aspects of the DP Sub-Cluster algorithm here. MCMC algorithms typically satisfy two conditions: detailed balance and ergodicity. Detailed balance ensures that the target distribution is

a stationary distribution of the chain, while ergodicity guarantees uniqueness of the stationary distribution. The method of [1] combines a Gibbs sampler that is restricted to non-empty clusters with a Metropolis-Hastings (MH) algorithm that proposes splits and merges. Since any Gibbs or MH sampler satisfies detailed balance, the true posterior distribution is guaranteed to be a stationary distribution of the chain. Furthermore, the combination of the two samplers enforces ergodicity and guarantees the convergence to the stationary distribution. The DP Sub-Cluster algorithm also augments the model with auxiliary variables that learn a twocomponent mixture model for each cluster. These ?sub-clusters? are subsequently used to propose splits that are learned over time instead of built in a single iteration like previous methods. In this paper, we extend these techniques to HDPs. As we will show, considerable work is needed to address the higher-level DP and the overlapping distributions that exist in topic modeling.

3

Hierarchical Dirichlet Processes

We begin with a brief review of the equivalent CRF and DA representations of the HDP [2] depicted in Figures 1a?1b. Due to the prolific use of HDPs in topic modeling, we refer to the variables with their topic modeling names. ? is the corpus-level, global topic proportions, ?k is the parameter for topic k, and xji is the ith word in document j. Here, the CRF and DA representations depart. In the CRF, ? ?j is drawn from a stick-breaking process [19], and each ?customer? (i.e., word) is assigned to a ?table? through tji ? Categorical(? ?j). The higher-level DP then assigns ?dishes? (i.e., topics) to tables via kjt ? Categorical(?). The association of customers to dishes through the tables is equivalent to assigning a word to a topic. In the CRF, multiple tables can be assigned the same dish. The DA formulation combines these multiple instances and directly assigns a word to a topic with zji . The resulting document-specific topic proportions, ?j , aggregates multiple ? ?j values. For 2

(a) HDP CRF Model

(b) HDP DA Model

(c) HDP Augmented DA Model

Figure 1: Graphical models. (c) Hyper-parameters are omitted and auxiliary variables are dotted.

Figure 2: Visualization of augmented sample space. reasons which will be discussed, inference in the DA formulation still relies on some aspects of the CRF. We adopt the notation of [2], where the number of tables in restaurant j serving dish k is denoted mjk , and the number of customers in restaurant j at table t eating dish k is njtk . Marginal P P counts are represented with dots, e.g., nj?? , t,k njtk and mj? , k mjk represent the number of customers and dishes in restaurant j, respectively. We refer the reader to [2] for additional details.

4

Restricted Parallel Sampling

We draw on the DP Sub-Cluster algorithm to combine a restricted, parallel Gibbs sampler with split/merge moves (as described in Section 2.1). The former

is detailed here, and the latter is developed in Section 5. Because the restricted Gibbs sampler cannot create new topics, dimensions of the infinite vectors ?, ?, and ? associated with empty clusters need not be instantiated. Extending the DA sampling algorithm of [2] results in the following restricted posterior distributions: p(?—m) = Dir(m?1 , . . . , m?K , ?), p(?j —?, z) = Dir(??1 + nj?1 , . . . , ??K + nj?K , ??K+1), p(?k —x, z) ? fx (xIk ; ?k)f? (?k ; ?), PK p(zji —x, ?j , ?) ? k=1 ?jk fx (xji ; ?k)1I[zji = k], p(mjk —?, z) = fm (mjk ; ??k , nj?k) ,

?(??k) mjk . ?(??k +nj?k) s(nj?k , mjk)(??k)

(1) (2) (3) (4) (5)

Since p(?—?) is not known analytically, we use the auxiliary variable, mjk , as derived by [2, 20]. Here, s(n, m) denotes unsigned Stirling numbers of the first kind. We note that ? and ? are now (K + 1)?length vectors partitioning the space, where the last components, ?K+1 and ?j(K+1) , aggregate the weight of all empty topics. Additionally, Ik , {j, i; zji = k} denotes the set of indices in topic k, and fx and f? denote the observation and prior distributions. We note that if f? is conjugate to fx , Equation (3) stays in the same family of parametric distributions as f? (?; ?). Equations (1?5), each of which can be sampled in parallel, fully specify the restricted Gibbs sampler. The astute reader may notice similarities with the FSD approximation used in [4]. The main differences are that the ? distribution in Equation (1) is exact, and that sampling z in Equation (4) is explicitly restricted to non-empty clusters. Unlike [4], however, this sampler is guaranteed to converge to the true HDP model when combined with any split move (cf. Section 2.1).

5

Augmented Sub-Cluster Space for Splits and Merges

In this section we develop the augmented, sub-cluster model, which is aimed at finding a twocomponent mixture model containing a likely split of the data. As demonstrated in [1], these splits perform well in DPs because they improve at every iteration of the algorithm. Unfortunately, because these splits perform poorly in HDPs, we modify the formulation to propose more flexible moves. For each topic, k, we fit two sub-topics, k‘ and kr, referred to as the ?left? and ?right? sub-topics. Each topic is augmented with auxiliary global sub-topic proportions, ? k = {? k‘ , ? kr }, document3

level sub-topic proportions, ? jk = {? jk‘ , ? jkr }, and sub-topic parameters, ?k = {?k‘ , ?kr }. Furthermore, a sub-topic assignment, z ji ? {‘, r} is associated with each word, xji . The augmented space is summarized in Figure 1c and visualized in Figure 2. These auxiliary variables are denoted with the same symbol as their ?regular-topic? counterparts to allude to their similarities. Extending the work of [1], we adopt the following auxiliary generative and marginal posterior distributions: Generative Distributions

Marginal Posterior Distributions p(? k —?) = Dir(? + m?k‘ , ? + m?kr),

p(? k) = Dir(?, ?), p(? jk —? k) = Dir(?? k‘ , ?? kr), Y Y p(?k —?, z, x) = f? (?kh ; ?) Zji (?, ?, z, x),

Zji (?, ?, z, x) ,

YK Y k=1

p(? jk —?) = Dir(?? k‘ +nj?k‘ ,?? kr +nj?kr), (7) p(?kh —?) ? fx (xIkh ; ?kh)f? (?kh ; ?),

(8)

p(z ji —?) ? ? jzji zji fx (xji ; ?zji zji)

(9)

p(mjkh —?) = fm (mjkh ; ?? kh , nj?kh),

(10)

j,i?Ik

h?{‘,r}

p(z—?, ?, z, x) =

(6)

? jkzji fx (xji ;? kzji)

j,i?Ik Zji (?,?,z,x)

X h?{‘,r}

,

? jzji h fx (xji ; ?zji h),

where ? denotes all other variables. Full derivations are given in the supplement. Notice the similarity between these posterior distributions and Equations (1?5). Inference is performed by interleaving the sampling of Equations (1?5) with Equations (6?10). Furthermore, each step can be parallelized. 5.1

Sub-Topic Split/Merge Proposals

We adopt a Metropolis-Hastings (MH) [21] framework that proposes a split/merge from the subtopics and either accepts or rejects it. Denoting v , {?, ?, z, ?} and v , {?, ?, z, ?} as the set of regular and auxiliary variables, a sampled proposal, {? v , v?} ? q(? v , v?—v) is accepted with probability h i ? v) q(v—x,? ? v)q(v—x,v,v) ? = min 1, p(x,?v)p(v—x,? ? Pr[{v, v} = {? v , v}] = min [1, H] . (11) ? p(x,v)p(v—x,v) q(? v —x,v)q(v—x,v,? v) H, is known as the Hastings ratio. Algorithm 1 outlines a general split/merge MH framework, where steps 1?2 propose a sample from q(? v —x, v)q(v?—x, v, v, v?). Sampling the variables other than z? is detailed here, after which we discuss three versions of Algorithm 1 with variants on sampling z?. Algorithm 1 Split-Merge Framework ? document proportions, ? ? 1. Propose assignments, z?, global proportions, ?, ? , and parameters, ?. 2. Defer the proposal of auxiliary variables to the restricted sampling of Equations (1?10). 3. Accept/reject the proposal with the Hastings ratio. ? In Metropolis-Hastings, convergence typically improves as the proposal distribution is (Step 1: ?): closer to the target distribution. Thus, it would be ideal to propose ?? from p(?—? z). Unfortunately, p(?—z) cannot be expressed analytically without conditioning on the dish counts, m?k , as in Equation (1). Since the distribution of dish counts depends on ? itself, we approximate its value with m ? jk (z) , arg maxm p(m—? = 1/K , z) = arg maxm

?(1/K) 1 m ?(1/K +nj?k) s(nj?k , m)(K) ,

(12)

where the global topic proportions have essentially been substituted with 1/K . We note that the dependence on z is implied through the counts, n. We

then propose global topics proportions from ? z) = p(?— ? m(? ?? ? q(?—? ? z)) = Dir (m ? ?1 (? z), ? ? ? , m ? ?K (? z), ?) .

(13)

We will denote m ? jk , m ? jk (z) and m ?? jk , m ? jk (? z). We emphasize that the approximate m ?? jk is only used for a proposal distribution, and the resulting chain will still satisfy detailed balance. (Step 1: ? ?): Conditioned on ? and z, the distribution of ? is known to be Dirichlet. Thus, we ? z?) by sampling directly from the true posterior distribution of Equation (2). propose ? ? ? p(? ? —?, ? If f? is conjugate to fx , we sample ?? directly from the posterior of Equation (3). If (Step 1: ?): non-conjugate models, any proposal can be used while adjusting for it in the Hastings ratio. 4

(Step 2): We use the Deferred MH sampler developed in [1], which sets q(v?—x, v?) = p(v?—x, v?) by deferring the sampling of auxiliary variables to the restricted sampler of Section 5. Splits and merges are then only proposed for topics where auxiliary variables have already burned-in. In practice burnin is quite fast, and is determined by monitoring the sub-topic data likelihoods. (Step 3): Finally, the above proposals results in the following the Hastings ratio: H=

? z)p(x—? p(?,? z) p(?,z)p(x—z)

?

? q(z—? v ,v)q(?—z) ? z) . q(? z —v,v)q(?—?

(14)

The data likelihood, p(x—z) is known analytically, and q(?—z) can be calculated according to Equation 13. The prior distribution, p(?, z), is expressed in the following proposition: Proposition 5.1. Let z be a set of topic assignments with integer values in {1, . . . , K}. Let ? be a (K +1)?length vector representing global topic weights, and ?K+1 be the sum of weights associated with empty topics. The prior distribution, p(?, z), marginalizing over ?, can be expressed as h i h YD YK YK ?(?? +n) i ?(?) ??1 k j?k p(?, z) = ??K+1 ?k?1 ? . (15) ?(?+nj??) ?(??k) k=1

k=1

j=1

Proof. See supplemental material. The remaining term in Equation (14), q(? z —v, v), is the probability of proposing a particular split. In the following sections, we describe three possible split constructions using the sub-clusters. Since ? the other steps remain the same, we only discuss the proposal distributions for z? and ?. 5.1.1

Deterministic Split/Merge Proposals

The method of [1] constructs a split deterministically by copying the subcluster labels for a single cluster. We refer to this proposal as a local split, which only changes assignments within one topic, as opposed to a global split (discussed shortly), which changes all topic assignments. A local deterministic split will essentially be accepted if the joint likelihood increases. Unfortunately, as we show in the supplement, samples from the typical set of an HDP do not have high likelihood. Deterministic split and merge proposals are, consequently,

very rarely accepted. We now suggest two alternative pairs of split and merge proposals, each with their own benefits and drawbacks. 5.1.2

Local Split/Merge Proposals

Here, we depart from the approach of [1] by sampling a local split of topic a into topics b and c. Temporary parameters, {? ?b , ? ?c , ??b , ??c }, and topic assignments, z?, are sampled according to) Y X (? ?b , ? ?c) = ?a ? (? a‘ , ? ar), ? ?k fx (xji ; ??k)1I[? zji = k]. (16) =? q(? z —v, v) ? (??b , ??c) = (?a‘ , ?ar), j,i?Ia k?{b,c} We note that a sample from q(? z —v, v) is already drawn from the restricted Gibbs sampler described in Equation (9). Therefore, no additional computation is needed to propose the split. If the split is rejected, the z? is simply used as the next sample of the auxiliary z for cluster a. A ?? is then drawn by splitting ?a into ??b and ??c according to a local version of Equation (13): ?? ?b , m ?? ?c). q(??b , ??c —? z , ?a) = Dir(??b /?a , ??c /?a ; m

(17)

The corresponding merge move combines topics b and c into topic a by deterministically performing q(? zji —v) = 1I[? zji = a],

q(??a —v) = ?(??a ? (?b + ?c)).

?j, i ? Ib ? Ic ,

This results in the following Hastings ratio for a local split (derivation in supplement): ? ? Y Y m ? +m ? ?c ? ? QM ?(???k +? nj?k) p(x—? z) m ? ?b)?(m ? ?c) ?a ?b ?(??a) 1 K+1 , H = ??(? ? m ? ?b m ? ? z —v,v) QS ?(??a +nj?a) ? ?c p(x—z) q(? ?(m ? +m ?) ?(???) ?b

?c

??b

??c

K

j

k?{b,c}

(18)

(19)

k

where QSK and QM K are the probabilities of selecting a specific split or merge with K topics. We record q(? z —v, v) when sampling from Equation (9), and all other terms are computed via sufficient statistics. We set QSK = 1 by proposing all splits at each iteration. QM K will be discussed shortly. 5

The Hastings ratio for a merge is essentially the reciprocal of Equation (19). However, the reverse ? and ??, which are not readily split move, q(z—? v , v?), relies on the inferred sub-topic parameters, ? available due to the Deferred MH algorithm. Instead, we approximate the Hastings ratio by substituting the two original topic parameters, ?b and ?c , for the proposed sub-topics. The quality of this approximation rests on the similarity between the regular-topics and the sub-topics. Generating the reverse move that splits topic a into b and c can then be approximated as Y ?zji fx (xji ;?zji) bb Lcc q(z—? v , v?) ? =L (20) Lbc Lcb , j,i?Ib ?Ic ?b fx (xi ;?b)+?c fx (xi ;?c) Y Y Lkk , ?k fx (xji ; ?k), Lkl , [?k fx (xji ; ?k) + ?l fx (xji ; ?l)] . (21) j,i?Ik

j,i?Ik

All of the terms in Equation (20) are already calculated in the restricted Gibbs steps. When aggregated correctly in the K ? K matrix, L, the Hastings ratio for any proposed merge is evaluated in constant time. However, if topics b and c are merged into a, further merging a with another cluster cannot be efficiently computed without looping through the data. We therefore only propose bK/2c merges by generating a random permutation of the integers [1, K], and proposing to merge disjoint neighbors. For example, if the random permutation for K = 7 is { 3 1 7 4 2 6 5}, we propose to 2bK/2c merge topics 3 and 1, topics 7 and 4, and topics 2 and 6. This results in QM K = K(K?1) . 5.1.3

Global Split/Merge Proposals

In many applications where clusters have significant overlap (e.g., topic modeling), local splits may be too constrained since only points within a single topic change. We now develop a global split and merge move, which reassign the data in all topics. A global split first constructs temporary topic ? followed by proposing topic assignments for all words with: proportions, ? ? , and parameters, ?,) Y ? (? ?b , ? ?c) = ?a ? (? a‘ , ? ar), ? ?k = ?k , ?k 6= a, ?z?ji fx (xji ; ??z?ji) =? q(? z —v, v) = . (22) P (??b , ??c) = (?a‘ , ?ar), ??k = ?k , ?k 6= a, ? ?k fx (xji ; ??k) j,i

k

Similarly, the corresponding merge move is constructed according to) Y ? ? ? a = ?b + ?c , ? ?k = ?k , ?k 6= b, c, ?z?ji fx (xji ; ??z?ji) . (23) =? q(? z —v, v) = P ??a ? q(??a —z, x), ??k = ?k , ?k 6= b, c, ? ?k fx (xji ; ??k) j,i

k

The proposal for ??a is written in a general form; if priors are conjugate, one should propose directly from the posterior. After Equations (22)?(23), ?? is sampled via Equation (13). All remaining steps follow Algorithm 1. The resulting Hastings ratio for a global split (see supplement) is expressed as K D K+1 D m ? Y Y Y ?(m Y ? ??a —z) QM ? ?k ?k ?(???k +? nj?k) z) q(z—? ?(??k) ? ?k) m ? ??) p(x—? v ,v)q(K+1 . (24) H = ??(?+ ? m ? ? q(? z —v,v) ?(m ? ?k) ?(??k +nj?k) QS ?(???k) ?(?+m ? ??) p(x—z) K ??k ?k j=1

k=1

k=1

j=1

Similar to local merges, the Hastings ratio for a global merge depends on the proposed sub-topics parameters. We approximate these with the main-topic parameters prior to the merge. Unlike the local split/merge proposals, proposing z? requires significant computation by looping through all data points. As such, we only propose a single global split and merge each iteration. Thus, QSK = 1/K and QM K = 2/(K(K ? 1)). We emphasize that the developed global moves are very different from previous local split/merge moves in DPs and HDPs (e.g., [1, 7, 11, 13, 14]). We conjecture that this is the reason the split/merge moves in [7] only made negligible improvement.

6

Experiments

We now test the proposed HDP Sub-Clusters method on topic modeling. The algorithm is summarized in the following steps: (1) initialize ? and z randomly; (2) sample ?, ?, ?, and ? via Equations (2, 3, 7, 8); (3) sample z and z via Equations (4, 9); (4) propose b K 2 c local merges followed by K local splits; (5) propose a global merge followed by a global split; (6) sample m and m via Equations (5, 10); (7) sample ? and ? via Equations (1, 6); (8) repeat from Step 2 until convergence. We fix the hyper-parameters, but resampling techniques [2] can easily be incorporated. All results are averaged over 10 sample paths. Source code can be downloaded from http://people.csail.mit.edu/jchang7. 6

(a) Visualizing Topics
20 10 0 -2 10
secs (log scale)
10
1
20
Num. Topics
1 Proc. 2 Procs.
Global Combined
4 Procs. 8 Procs.
10
(b) Split/Merge Moves
0 -2 10
10
secs (log scale)
(c) Parallelization
10 0 -2
Combined
HOW Log Like.
Det. Local
Num. Topics
Num. Topics
20
1
-2.5 -3 -3.5 -4 -3 10 10
secs (log scale)
10
2
10
3
(d) Algorithm Comparison
-8 -8
-8.4
Num. Topics
100
-8.2
50 -8.4 0

0
secs
1000 0
Number of Topics
-8 -8.2
-8
-8.4 100
Num. Topics
-8.2
-7.8
-7.8
50 -8.4 0
100
(a) AP Results with Different Initializations
-8.2
HOW Log Likelihood
HOW Log Like.
-7.8
-7.8
HOW Log Likelihood
HOW Log Like.

Figure 3: Synthetic ?bars? example. (a) Visualizing topic word distributions without splits/merges for K = 5. (b)?(c) Number of inferred topics for different split/merge proposals and parallelizations. (d) Comparing sampling algorithms with a single processor and initialized to a single topic.

0
secs
2000 0
Number of Topics
100
(b) AP Results with Switching Algorithms

Figure 4: Results on AP. (a) 1, 25, 50, and 75 initial topics. (b) Switching algorithms at 1000 secs.

6.1

Synthetic Bars Dataset

We synthesized 200 documents from the ?bars? example of [22] with a dictionary of 25 words that can be arranged in a 5x5 grid. Each of the 10 true topics forms a horizontal or vertical bar. To visualize the sub-topics, we initialize to 5 topics and do not propose splits or merges. The resulting regular- and sub-topics are shown in Figure 3a. Notice how the sub-topics capture likely splits. Next, we consider different split/merge proposals in Figure 3b. The ?Combined? algorithm uses local and global moves. The deterministic moves are often rejected resulting in slow convergence. While global moves are not needed in such a well-separated dataset, we have observed that the make a significant impact in real-world datasets. Furthermore, since every step of the sampling algorithm can be parallelized, we achieve a linear speedup in the

number of processors, as shown in Figure 3c. Figure 3d compares convergence without parallelization to the Direct Assignment (DA) sampler and the Finite Symmetric Dirichlet (FSD) of order 20. Since all algorithms should sample from the same model, the goal here is to analyze convergence speed. We plot two summary statistics: the likelihood of a single held-out word (HOW) from each document, and the number of inferred topics. While the HOW likelihood for FSD converges at 1 second, the number of topics converges at 100 seconds. This suggests that cross-validation techniques, which evaluate model fit, cannot solely determine MCMC convergence. We note that FSD tends to first create all L topics and slowly remove them. 6.2

Real-World Corpora Datasets

Next, we consider the Associated Press (AP) dataset [23] with 436K words in 2K documents. We manually set the FSD order to 100. Results using 16 cores (except DA, which cannot be parallelized) with 1, 25, 50, and 75 initial topics are shown in Figure 4a. All samplers should converge to the same statistics regardless of the initialization. While HOW likelihood converges for 3/4 FSD initializations, the number of topics indicates that no DA or FSD sample paths have converged. Unlike the well-separated, synthetic dataset, the Sub-Clusters method that only uses local splits and merges does not converge to a good solution here. In contrast, all initializations of the Sub-Clusters method have converged to a high HOW likelihood with only approximately 20 topics. The path taken by each sampler in the joint HOW likelihood / number of topics space is shown in the right panel of Figure 4a. This visualization helps to illustrate the different approaches taken by each algorithm. Figure 5aPshows confusion matrices, C, of the inferred topics. Each element of C is defined as: Cr,c = x fx (x; ?r) log fx (x; ?c), and captures the likelihood of a random word from topic r 7

(a) Confusion Matrices for AP

(b) Four Topics from NYTimes

-8.6 -8.2

Num. Topics

200 100

-8.6 0

-1

10 10

0

secs (log scale)

4

5

10 10 0

Number of Topics

-8.7 -8.7

-9 -9.3

-9

200 Num. Topics

-7.8

-8.2
100 -9.3 0
200
(a) Enron Results
HOW Log Likelihood
HOW Log Like.
-7.8 HOW Log Likelihood
HOW Log Like.

Figure 5: (a) Confusion matrices on AP for S UB -C LUSTERS, DA, and FSD (left to right). Outlines are overlaid to compare size. (b) Four inferred topics from the NYTimes articles.

-1
10 10
0
secs (log scale)
4
5
10 10 0
Number of Topics
200
(b) NYTimes Results

Figure 6: Results on (a) Enron emails and (b) NYTimes articles for 1 and 50 initial topics.

evaluated under topic c. DA and FSD both converge to many topics that are easily confused, whereas the Sub-Clusters method converges to a smaller set of more distinguishable topics. Rigorous proofs about convergence are quite difficult. Furthermore, even though the approximations made in calculating the Hastings ratios for local and global splits (e.g., Equation (20)) are backed by intuition, they complicate the analysis. Instead, we run each sample path for 2,000 seconds. After 1,000 seconds, we switch the Sub-Clusters sample paths to FSD and all other sample paths to SubClusters. Markov chains that have converged should not change when switching the sampler. Figure 4b shows that switching from DA, FSD, or the local version of Sub-Clusters immediately changes the number of topics, but switching Sub-Clusters to FSD has no effect. We believe that the number of topics is slightly higher in the former because the Sub-Cluster method struggles to create small topics. By construction, the splits make large moves, in contrast to DA and FSD, which often create single word topics. This suggests that alternating between FSD and Sub-Clusters may work well. Finally, we consider two large datasets from [24]: Enron Emails with 6M words in 40K documents and NYTimes Articles with 100M words in 300K documents. We note that the NYTimes dataset is 3 orders of magnitude larger than those considered in the HDP split/merge work of [7]. Again, we manually set the FSD order to 200. Results are shown in Figure 6 initialized to 1 and 50 topics. In such large datasets, it is difficult to predict convergence times; after 28 hours, it seems as though no algorithms have converged. However, the Sub-Clusters method seems to be approaching a solution, whereas FSD has yet

to prune topics and DA has yet to to achieve a good cross-validation score. Four inferred topics using the Sub-Clusters method on the NYTimes dataset are visualized in Figure 5b. These words seem to describe plausible topics (e.g., music, terrorism, basketball, and wine).

7

Conclusion

We have developed a new parallel sampling algorithm for the HDP that proposes split and merge moves. Unlike previous attempts, the proposed global splits and merges exhibit significantly improved convergence in a variety of datasets. We have also shown that cross-validation metrics in isolation can lead to the erroneous conclusion that an MCMC sampling algorithm has converged. By considering the number of topics and held-out likelihood jointly, we show that previous sampling algorithms converge very slowly. Acknowledgments This research was partially supported by the Office of Naval Research Multidisciplinary Research Initiative program, award N000141110688 and by VITALITE, which receives support from Army Research Office Multidisciplinary Research Initiative program, award W911NF-11-1-0391. 8

2 References

[1] J. Chang and J. W. Fisher, III. Parallel sampling of DP mixture models using sub-clusters splits. In Advances in Neural Information and Processing Systems, Dec 2013. [2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566?1581, 2006. [3] E. B. Sudderth. Graphical Models for Visual Object Recognition and Tracking. PhD thesis, Massachusetts Institute of Technology, 2006. [4] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An HDP-HMM for systems with state persistence. In International Conference on Machine Learning, July 2008. [5] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In Advances in Neural Information Processing Systems, volume 20, 2008. [6] M. Bryant and E. Sudderth. Truly nonparametric online variational inference for Hierarchical Dirichlet processes. In Advances in Neural Information Processing Systems, 2012. [7] C. Wang and D Blei. A split-merge MCMC algorithm for the Hierarchical Dirichlet process. arXiv:1207.1657 [stat.ML], 2012. [8] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. Journal of Machine Learning Research, 10:1801?1828, December 2009. [9] S. Williamson, A. Dubey, and E. P. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In International Conference on Machine Learning, 2013. [10] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249?265, June 2000. [11] S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics, 13:158?182, 2000. [12] P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. Scandinavian

Journal of Statistics, pages 355?375, 2001. [13] D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. Technical report, University of Wisconsin - Madison Dept. of Statistics, 2003. [14] S. Jain and R. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. Bayesian Analysis, 2(3):445?472, 2007. [15] H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. Canadian Journal of Statistics, 30:269?283, 2002. [16] F. Niu, B. Recht, C. R?e, and S. J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In Advances in Neural Information Processing Systems, 2011. [17] M. J. Johnson, J. Saunderson, and A. S. Willsky. Analyzing hogwild parallel gaussian gibbs sampling. In Advances in Neural Information Processing Systems, 2013. [18] Y. Gal and Z. Ghahramani. Pitfalls in the use of parallel inference for the Dirichlet process. In Workshop on Big Learning, NIPS, 2013. [19] J. Sethuraman. A constructive definition of Dirichlet priors. Statstica Sinica, pages 639?650, 1994. [20] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics, 2(6):1152?1174, 1974. [21] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97?109, 1970. [22] T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101:5228?5235, April 2004. [23] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993?1022, March 2003. [24] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.

9