Split-Merge Augmented Gibbs Sampling for Hierarchical Dirichlet Processes

Santu Rana, Dinh Phung, and Svetha Venkatesh

Center for Pattern Recognition and Data Analytics
Deakin University
Waurn Ponds, VIC 3216

Abstract. The Hierarchical Dirichlet Process (HDP) model is an important tool for topic analysis. Inference can be performed through a Gibbs sampler using the auxiliary variable method. We propose a split-merge procedure to augment this method of inference, facilitating faster convergence. Whilst the incremental Gibbs sampler changes topic assignments of each word conditioned on the previous observations and model hyper-parameters, the split-merge sampler changes the topic assignments over a group of words in a single move. This allows efficient exploration of state space. We evaluate the proposed sampler on a synthetic test set and two benchmark document corpus and show that the proposed sampler enables the MCMC chain to converge faster to the desired stationary distribution.

1 Introduction

The hierarchical Dirichlet process (HDP) [1] is an important tool for Bayesian nonparametric topic modelling, particularly when mixed-membership exists, such as in a document collection. Each document is modelled as a group of words, generated from the underlying latent topic. It is an extension of Latent Dirichlet Allocation (LDA) [2], allowing unbounded latent dimensionality, with capacity to automatically infer the number of topics in a document set. The HDP is a hierarchial version of the Drichilet process (DP) clustering model, where a corpus of documents are assumed to be generated from a set of top-level topics with independent mixing distribution. In contrast to the DP mixture model for which the metaphor is a Chinese Resturant Process (CRP), a HDP can be expressed using a metaphor of Chinese Restaurant Franchise (CRF), where a set of dishes is shared across a collection of franchise restaurants, each having tables generated using a CRP from the customers arriving at that franchise.

As with Bayesian nonparametric models, exact posterior inference is not tractable. MCMC or variational approximation are used for approximate posterior inference. In this paper, we focus on MCMC sampling, wherein posterior is computed from the empirical distribution of samples from a Markov chain, whose stationary distribution is the posterior of interest. [1] propose two MCMC sampling schemes, one based on the CRF and the other on the auxiliary variable

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 546–557, 2013.
© Springer-Verlag Berlin Heidelberg 2013

method. In many cases the use of auxiliary variable sampling method is preferred to keep the sampling simple and easily extendable to elaborate models such as iHMM [1]. The basic MCMC sampler for the HDP is an incremental Gibbs sampler - the topic is sampled for a single observation, one at a time, conditioned on preceding observations and model hyperparameters. Since, only one state change takes place at a time, mixing may be slow, requiring many Gibbs iterations for the MCMC chain to converge to its stationary distribution. Whereas CRF based sampling is staightforward in formulation, the implementation is tedious, requiring tracking of individual table assignments for each restaurant, and then tracking the dish preference for each table. The auxiliary variable split-merge sampler is based on directly sampling the topic assignment (dish) of the words (customers) in the documents and thus straightforward in implementation.

Split-merge MCMC samplers have been proposed for Bayesian nonparametric models, such as for DPM to accelerate mixing [3]. In a split-merge setting, a group of observations are moved together in the state space based on whether splitting or merging of topics are accepted based on a Metropolis-Hastings ratio. In practice, each sampling run consists of a Gibbs sampling followed by a split-merge proposal evaluation. Since, the state change may occur for a group of points at each iteration, the MCMC chain can quickly traverse the state-space and potentially converge faster than if only the Gibbs sampler is used.

Motivated by this, we propose a split-merge procedure for the HDP to accelerate the mixing of the MCMC chain for the auxiliary variable sampling scheme called the *Split-Merge Augmented Gibbs sampler*. Assuming each word (customer) in the document corpus has been assigned to a topic(dish) at the higher level, we evaluate a split-merge proposal on the customer-dish relationship i.e. we either propose to split all the customers in all the franchise restaurants who share the same dish into two different dishes or propose merging the set of all customers sharing two different dishes. In contrast to the CRF based split-merge sampling scheme [4], we do not worry about the lower-level customer-table assignments and thus the proposed split-merge scheme is effective at both levels of the HDP.

We evaluate and analyze the proposed algorithm on synthetic data and two benchmark document corpus, - NIPS abstracts and 20 News Group data. In synthetic experiments, we generate topics with low separability and show that the incremental Gibbs sampler is unable to recover all the correct topics; however, our split-merge augmented Gibbs sampler is able to recover all topics correctly. For the document corpus, we evaluate the performance of Gibbs vs our sampler based on the perplexity of held-out data and show that our proposed method is able to produce lower perplexity in similar time.

The layout is as follows: Related background on HDP and inference techniques is described in the section 2; in the section 3, we detail the split-merge procedure after briefly reviewing the Gibbs sampling procedure based on the auxiliary variable scheme. Experimental results are discussed in section 4 and finally, section 5 concludes our discussion.

2 Related Background

Dirichlet Proess (DP) mixture model for clustering with theoretically unbounded mixture component has been first studied in [5] with [6] giving a stick-breaking construction for the DP prior. Hierarchical Dirichlet Process (HDP) extends the DP in two level where the bottom level DP uses the top-level DP as the base measure was first proposed in [1]. This is a mixed-membership model where a group is sampled from a mixture of topics and has been used extensively for document analysis [7], multi-population haplotype phasing [8], image/object retrieval [9] etc.

Split-merge sampling for DP mixture model was first proposed in [3] and splitting of a single cluster by running a Restricted Gibbs sampler on the subset of points belonging to that topic is described. Whilst a merge proposal is easy to generate, generating a split proposal takes some work as a random split will most likely to be a bad proposal and they would be rejected. Hence, the need for the Restricted Gibbs sampler. Using the same framework [10] proposed a slightly different split-merge algorithm by having a simpler split routine using a sequential allocation scheme. In contrast to running a Gibbs sampler to generate a split proposal they proposed a single run sequential allocation scheme to generate the split, thus reducing the overhead cost. Split-merge sampler for HDP based on the Chinese Restaurant Franchise sampling scheme has been proposed in [4]. This perform splitting or merging only at the top level assignments using the similar procedure for the DP with additional factors coming from the bottom level when computing the prior clustering probability.

3 Framework

3.1 Hierarchical Dirichlet Process

The hierarchical Dirichlet process is a distribution over a set of random probability measure over $(\Theta, \mathcal{B})$. It is a hierarchical version of the DP, where a set of group level random probability measures $(G_j)_{j=1}^{J}$ are defined for each group which shares a global random probability measure G_0 at the higher level. The global measure G_0 is a draw from a DP with a base measure H and a concentration parameter γ. The group specific random measures G_j are subsequently drawn from a DP with G_0 as its base measure,

$$G_0 \sim DP(\gamma, H) \tag{1}$$

$$G_j/G_0 \sim DP(\alpha_0, G_0) \tag{2}$$

with j denoting the group. Since G_j are drawn from the almost surely discrete distribution of G_0, it ensures that the top level atoms are shared across the groups. In the topic model context each document is a group of words and the

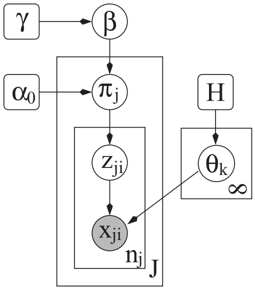


Fig. 1. The HDP model

atoms (topics) are the distribution over words. The stick-breaking representation of G_0 can be expressed as,

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{3}$$

where $\theta_k \sim H$ independently and $(\beta_k)_{k=1}^{\infty}$ admitting stick-breaking construction such that $(\beta_k)_{k=1}^{\infty} \sim Stick(\gamma)$. Since, G_0 is used as the base measure for G_j, it can be expressed as,

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \tag{4}$$

where it can be shown that [1] $\pi_j \sim DP(\alpha_0, \beta)$. The stick-breaking representation for HDP is given below,

$$\begin{aligned} \beta|\gamma &\sim Stick(\gamma) && \\ \pi_{j|\alpha_0,,\beta} &\sim DP(\alpha_0,\beta) && z_{ji}|\pi_j \sim \pi_j \\ \theta_k|H &\sim H && x_{ji}|z_{ji},(\theta_k)_{k=1}^{\infty} \sim F(\theta_{z_{ji}}) \end{aligned} \tag{5}$$

3.2 Posterior Inference with Auxiliary Variable

With the stick breaking representation of 5, the state space consists of $(\mathbf{z}, \pi, \beta, \theta)$. Since $\mathbf{z}$ and π forms a conjugate pair, π can be integrated out giving the conditional probability of $\mathbf{z}$ given β as.

$$P(\mathbf{z}|\beta) = \prod_{j=1}^{J} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_0\beta_k + n_{jk})}{\Gamma(\alpha_0\beta_k)} \tag{6}$$

From 6 the prior probability for z_{ji} given $\mathbf{z}^{-\mathbf{ji}}$ and β can be expressed as,

$$p(z_{ji} = k|\mathbf{z}^{-ji}, \beta) = \frac{(\alpha_0 \beta_k + n_{jk}^{-ji})}{\alpha_0 + n_j} \quad for \quad k = 1, ...K, u \tag{7}$$

where β =[$\beta_1 \beta_2 ... \beta_u$] such that $\beta_u = \sum_{k=K+1}^{\infty} \beta_k$. Adding the likelihood term we can have the sampling formula of z_{ji} as,

$$p(z_{ji} = k|\mathbf{z}^{-ji}, \beta) \propto (\alpha_0 \beta_k + n_{jk}^{-ji}) f(x_{ji}/\theta_k) \quad for \quad k = 1, ...K, u \tag{8}$$

where θ_u is sampled from its prior H. If a new topic $(K+1)$ is created then we set $\beta_{K+1} = b\beta_u$, where $b \sim Beta(1, \gamma)$. To sample β we use the auxiliary variable method as outlined in [Teh]. We first sample the auxiliary variable $\mathbf{m}$ from,

$$q(m_{jk} = m|\mathbf{z}, \mathbf{m}^{-jk}, \beta) \propto s(n_{jk}, m)(\alpha_0 \beta_k)^m \tag{9}$$

where $s(n_{jk}, m)$ are the unsigned Stirling numbers of the first kind. Subsequently, β is sampled from,

$$q(\beta|\mathbf{z}, \mathbf{m}) \propto \beta_u^{\gamma - 1} \prod_{k=1}^{K} \beta_k^{\sum_j m_{jk} - 1} \tag{10}$$

Eq 8910 completes the Gibbs sampling formula for HDP inference. For elaboration please refer to [1,7].

3.3 Split-Merge procedure

The split-merge proposal is a form of Metropolis-Hasting algorithm where the algorithm draws a new candidate state C^* from a distribution with density $\pi(C)$ according to a proposal density $q(C^*/C)$ and then evaluation of the proposal based on the Metropolis-Hasting ratio of

$$a(C^*, C) = min[1, \frac{q(C|C^*)\pi(C^*)}{q(C^*|C)\pi(C)}]$$

The proposal $C^*$is accepted with the probability $a(C^*, C)$. If it is accepted the state changes to C^* or it remains at C. For HDP mixture model the above formula takes the form of

$$a(C^*, C) = min[1, \frac{q(C|C^*)P(C^*)L(C^*|\mathbf{x})}{q(C^*|C)P(C)L(C|\mathbf{x})}$$

From this prior distribution of $P(\mathbf{z}|\beta)$ (Eq. 7) we can use the Polya's urn metaphor to create a sequence $[z_{ji_1} z_{ji_2} ... z_{jn_j}]$ for a particular document j given β as,

$$\begin{aligned} P(z_{ji} = k|c_1, c_2, ..., c_k; z_{j1}, z_{j2}, ..., z_{ji-1}; \beta) &= \frac{\alpha_0 \beta_k + n_{jc_k}^{<i}}{\alpha_0 + i - 1} \, for \; k <= K \\ &= \frac{\alpha_0 \beta_u}{\alpha_0 + i - 1} \, for \; k = K + 1 \end{aligned}$$

where c_k is the $k'th$ topic. Given this assignement scheme, the probability of a particular configuration of word assignments $\mathbf{C} = \{n_{jc_1}, n_{jc_2}, ..., n_{jc_K}\}_{j=1}^{J}$ to the topic set $\{c_k\}_{k=1}^{K}$ can be expressed as,

$$P(\mathbf{C}|\beta) = \frac{\alpha_0^K \beta_{u_1} \beta_{u_2} ... \beta_{u_k} \prod_{j=1}^{J} \prod_{k=1}^{K} < \alpha_0 \beta_k >_{n_{jk}}}{\prod_{j=1}^{J} \prod_{i=1}^{n_j} (\alpha_0 + i - 1)} \tag{11}$$

where $\beta_{u_k} = \sum_{l=k}^{\infty} \beta_l$ and $< \alpha_0 \beta_k >_{n_{jk}} = \alpha_0 \beta_k (\alpha_0 \beta_k + 1) ... (\alpha_0 \beta_k + n_{jk} - 1)$ denotes the rising factorial and can be computed as the ratio of two gamma functions. For a split proposal a particular topic k is splitted in k_1 and k_2 and the new configuaration is denoted as $\mathbf{C}^{split}$. After we generate new latent assignements $\mathbf{z}^{split}$ corresponding to $\mathbf{C}^{split}$, we resample βusing 9 and 10 to obtain β^{split}.The configuration probability of $P(\mathbf{C}^{split}/\beta^{split})$ can now be computed as,

$$P(\mathbf{C}^{split}|\beta^{split}) = \frac{\alpha_0^{K+1} \beta_{u_1} \beta_{u_2} ... \beta_{u_{k+1}} \prod_{j=1}^{J} \prod_{k=1}^{K+1} < \alpha_0 \beta_k >_{n_{jk}}}{\prod_{j=1}^{J} \prod_{i=1}^{n_j} (\alpha_0 + i - 1)} \tag{12}$$

Now we can compute $\frac{P(\mathbf{C}^{split}|\beta^{split})}{P(\mathbf{C}|\beta)}$ as the ratio of 12 and 11. Similarly for merge proposal when topics k_1 and k_2 are merged into a single topic k and β_k^{merge} is sampled with the new $\mathbf{z}^{merge}$, then we have,

$$P(\mathbf{C}^{merge}|\beta^{merge}) = \frac{\alpha_0^{K-1} \beta_{u_1} \beta_{u_2} ... \beta_{u_{k-1}} \prod_{j=1}^{J} \prod_{k=1}^{K-1} < \alpha_0 \beta_k >_{n_{jk}}}{\prod_{j=1}^{J} \prod_{i=1}^{n_j} (\alpha_0 + i - 1)} \tag{13}$$

from which the ratio $\frac{P(\mathbf{C}^{split}|\beta^{split})}{P(\mathbf{C}|\beta)}$ can be computed from 13 and 11. In our proposed method we will use the conditional configuration probability ratio in place of $\frac{P(C^*)}{P(C)}$ as our target distribution is $\pi(\mathbf{C}|\beta)$.

The likelihood term $L(C/\mathbf{x})$ is computed over all the words of all the documents and is given as,

$$L(C|\mathbf{x}) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \int f(x_{ji}, \theta) dH_{ji,c_{ji}}(\theta)$$

where $H_{ji,c_{ji}}$is the posterior distribution of θ based on the prior G_0 and all the observations $x_{j',i'}$ such that $j' < j$ and $i' < i$. The above integral is analytically tractable if G_0 is conjugate prior. We can express the above likelihood equation as a product over topics such that,

$$L(C|\mathbf{x}) = \prod_{j=1}^{J}\prod_{c=1}^{K}\prod_{i:C_{ji}=c}\int f(x_{ji},\theta)dH_{ji,c}(\theta)$$

Expressing this way now we can compute the ratio of likelihoods between a split proposal C^{split} and the existing configuration C as,

$$\frac{L(C^{split}|\mathbf{x})}{L(C|\mathbf{x})} = \frac{\prod_{j=1}^{J}\prod_{i:C_{ji}^{split}=k1}\int f(x_{ji},\theta)dH_{ji,k1}(\theta)\prod_{j=1}^{J}\prod_{i:C_{ji}^{split}=k2}\int f(x_{ji},\theta)dH_{ji,k2}(\theta)}{\prod_{j=1}^{J}\prod_{i:C_{ji}=k}\int f(x_{ji},\theta)dH_{ji,k}(\theta)} \tag{14}$$

Similarly, for merge proposal the ratio of likelihood is,

$$\frac{L(C^{merge}|\mathbf{x})}{L(C|\mathbf{x})} = \frac{\prod_{j=1}^{J}\prod_{i:C_{ji}^{merge}=k}\int f(x_{ji},\theta)dH_{ji,k}(\theta)}{\prod_{j=1}^{J}\prod_{i:C_{ji}=k1}\int f(x_{ji},\theta)dH_{ji,k1}(\theta)\prod_{j=1}^{J}\prod_{i:C_{ji}=k2}\int f(x_{ji},\theta)dH_{ji,k2}(\theta)} \tag{15}$$

To evaluate the proposal density $q(C^*|C)$ we need to create an algorithm for creating C^* from the existing configuaration C. Here we use sequential assignment method similar to [10] for that. Let us assume that we are generating a split proposal for the topic k into two topics k_1 and k_2. We need to divide the words $S = \{n_{jc_k}\}_{j=1}^{J}$ into two sets $S_{k_1} = \{n_{j_{k_1}}\}_{j=1}^{J}$ and $S_{k_2} = \{n_{j_{k_2}}\}_{j=1}^{J}$. We start with a random word from a random document as the seed for the topic k_1 and similarly for topic k_2 i.e. $S_{k_1} = \{x_{jr_1,ir_1}\}_{(jr_1,ir_1)\in S}$ and $S_{k_1} = \{x_{jr_2,ir_2}\}_{(jr_2,ir_2)\in S}$ such that $(jr_1, ir_1) \neq (jr_2, ir_2)$. The rest of the words from the set S can be assigned by sampling from,

$$P(k_{ji} = k_1|S_{k_1}, S_{k_2}, \theta, x_{ji}) \tag{16}$$

$$= \frac{(\alpha_0\beta_{k_1}^{split}+|S_{k_1}^{j}|-1)p(x_{ji}|\theta_{S_{k_1}})}{(\alpha_0\beta_{k_1}^{split}+|S_{k}^{j}|-1)p(x_{ji}|\theta_{S_{k_1}})+(\alpha_0\beta_{k_2}^{split}+|S_{k_2}^{j}|-1)p(x_{ji}|\theta_{S_{k_2}})}$$

$$P(k_{ji} = k_2|S_{k_1}, S_{k_2}, \theta, x_{ji}) \tag{17}$$

$$= \frac{(\alpha_0\beta_{k_2}^{split}+|S_{k_2}^{j}|-1)p(x_{ji}|\theta_{S_{k_2}})}{(\alpha_0\beta_{k_1}^{split}+|S_{k}^{j}|-1)p(x_{ji}|\theta_{S_{k_1}})+(\alpha_0\beta_{k_2}^{split}+|S_{k_2}^{j}|-1)p(x_{ji}|\theta_{S_{k_2}})}$$

where, a simple allocation of $\beta_{k_1}^{split}$and $\beta_{k_2}^{split}$ can be assigned as $\beta_{k_1}^{split} = \beta_{k_2}^{split} = \beta_k/2$. The proposal probability $q(\mathbf{C}^{split}|\mathbf{C})$ is computed as a product of the above probabilities based on the actual assignment. The reverse proposal probability $q(\mathbf{C}|\mathbf{C}^{split}) = 1$ since the set of two sets of words can only be combined in a single way. We propose merge proposal as combining the two topics k_1 and $k_2$into a single topic k. In this case $q(\mathbf{C}^{merge}|\mathbf{C}) = 1$, however, to compute the reverse proposal probability $q(\mathbf{C}|\mathbf{C}^{merge})$ we need to create a dummy split proposal and compute $q(\mathbf{C}|\mathbf{C}^{merge}) = q(\mathbf{C}^{dummysplit}|\mathbf{C}^{merge})$ following the previously described split procedure.

Our split-merge procedure runs after each Gibbs iteration and at each run of aplit-merge procedure we either select to perform a split or merge. Till now we have not discussed whether a split or a merge proposal is to be evaluated. The simplest way to determine that by way of sampling two random words from the document corpus and then depending on whether they belong to the same topic or not we evaluate a split or merge proposal respectively. Whilst this scheme works fine it is understood that with the increasing number of topics we may encounter more merge proposal being evaluated than split proposals. To circumvent that we propose sampling from a binary random variable with equal probability of selecting a merge or split proposal at each run. When a split proposal has to be created we first select a topic at random and then proceed with splitting that topic, similarly when a merge proposal has to be created we select two topics at random and then proceed with the merging. From our experience this provides faster convergence than the naive method.

4 Experiments

We evaluate our proposed split-merge algorithm for HDP topic models for both synthetic and real world data. In all experiments, we run the normal conditional Gibbs sampler and the proposed split-merge augmented Gibbs sampler for the HDP model, with identical initialization of state space and variables. The normal Gibbs sampler visits each document and all words within it sequentially, assigning each to one to an existing topic or creating a new one based on the predictive likelihood of the word. The split-merge augmented Gibbs sampler runs a Gibbs iteration followed by the split-merge procedure. A split or a merge is proposed based on user-defined selection probability (a simple scheme is to have equal probability of acceptance). Depending on whether a split or merge has been selected, we pick two words randomly from a single topic or from two different topics for split and merge respectively. We then propose the split or the merge and accept them based on its acceptance probability.

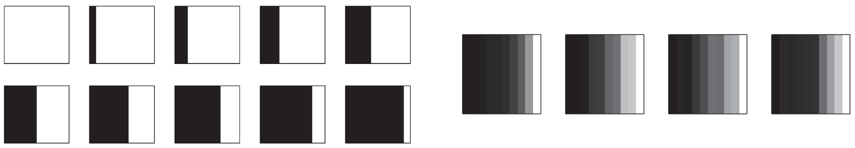
4.1 Synthetic Data

We use synthetic data to demonstrate the performance of our proposed split-merge augmented sampler in comparison to the simple conditional Gibbs sampler. We generate 10 topics from a vocabulary size of 10. The topics are created such that the first topic uses all the words with equal probability, and the rest use lesser number of words, with the last topic using only a single word, as shown in the Fig 2a. Fig 2b shows the extracted four groups. The topic mixture for each group has been generated as a random simplex.

Both the Gibbs sampler and the split-merge augmented Gibbs samplers are run for 1000 iterations and the posterior for the cluster number is shown in the Fig 3b and Fig 3a respectively. Whilst the naive conditional sampler fails to recover exact topics even after 1000 iterations, the split-merge augmented Gibbs Sampler is able to find the correct number of topics within the first 25

Algorithm 1. Split-merge augmented Gibbs sampler for HDP

```
For each iteration:
  – Perform Gibbs sampling using auxiliary variable scheme (Eq. 8,9, and10).
  – Choose a split or merge decision by sampling $t \sim Bern(0.5)$ with $t = 0$ indicating a split and $t = 1$ indicating a merge.
  – If split:
      • Randomly select a topic to split.
      • Split the chosen topic into two and generate $\mathbf{z}^{split}$ using Eq. 16 and 17.
      • Resample $\beta^{split}$ using Eq. 9 and 10.
      • Compute the proposal likelihood ratio $(\frac{P(\mathbf{C}^{split}|\beta^{split})}{P(\mathbf{C}|\beta)})$ from Eq. 12 and 11.
      • Compute likelihoods ratio $(\frac{L(C^{split}|\mathbf{x})}{L(C|\mathbf{x})})$ from Eq. 14.
      • Set $q(\mathbf{C}|\mathbf{C}^{split}) = 1$ and compute $q(\mathbf{C}^{split}|\mathbf{C})$ from Eq. 16 and 17 by multiplying the assignment probabilities.
      • Compute the Metropolis Hasting ratio
            $a(\mathbf{C}^{split}, \mathbf{C}) = min[1, \frac{q(\mathbf{C}|\mathbf{C}^{split})P(\mathbf{C}^{split}|\beta^{split})L(\mathbf{C}^{split}|\mathbf{x})}{q(\mathbf{C}^{split}|\mathbf{C})P(\mathbf{C}|\beta)L(\mathbf{C}|\mathbf{x})}$
      • Accept the split proposal with probability $a(\mathbf{C}^{split}, \mathbf{C})$.
      • Set $\mathbf{z} = \mathbf{z}^{split}$ and $\beta = \beta^{split}$.
  – if merge:
      • Randomly select two topics.
      • Merge them into two and generate $\mathbf{z}^{merge}$ and resample $\beta^{merge}$.
      • Create a dummy split following the split algorithm as outlined above to obtain
            $a(\mathbf{C}^{merge}, \mathbf{C}^{dummysplit}) =$
            $min[1, \frac{q(\mathbf{C}^{dummysplit}|\mathbf{C}^{merge})P(\mathbf{C}^{merge}|\beta^{merge})L(\mathbf{C}^{merge}|\mathbf{x})}{q(\mathbf{C}^{merge}|\mathbf{C}^{dummysplit})P(\mathbf{C}^{dummysplit}|\beta^{dummysplit})L(\mathbf{C}^{dummysplit}|\mathbf{x})}$
      • Accept the merge proposal with probability $a(\mathbf{C}^{merge}, \mathbf{C}^{dummysplit})$.
      • Set $\mathbf{z} = \mathbf{z}^{merge}$ and $\beta = \beta^{merge}$.
```

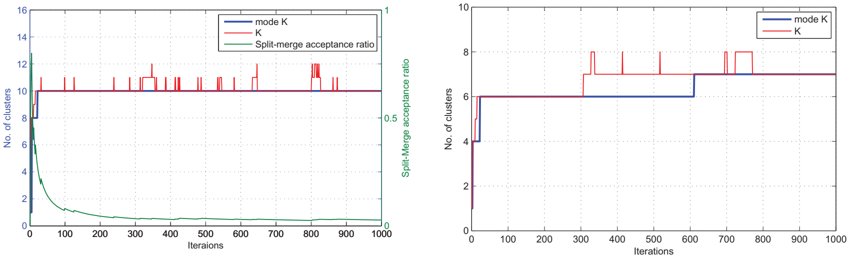


(a) The 10 topics

(b) The four groups.

Fig. 2. Synthetic experimental set up (a) the 10 topics, (b) the 4 groups represented as a bag of words

iterations. This is a significant speed up. The reason the naive Gibbs sampler fails to separate the topic is because they are not easily separable, however, our algorithm is able to split topics that are hard to separate. Fig 3a shows the split-merge acceptance ratio after each iteration. As expected the ratio falls with increasing number of samples, once all 10 topics have been recovered correctly. The confusion matrix for the topics as recovered by the two sampling algorithms is shown in Fig 4. Since the first few topics have a higher overlap, they are hard to separate.Thus it is nor surprising that the naive Gibbs sampling fails to separate them, however, our algorithm, with its capability to explore state-space in an efficient way, is able to separate the topics.



(a) Combined Gibbs and Split-Merge sampler

(b) Gibbs sampler

Fig. 3. Posterior K on synthetic data for (a) combined Gibbs and Split-Merge sampler, and (b) only the Gibbs sampler

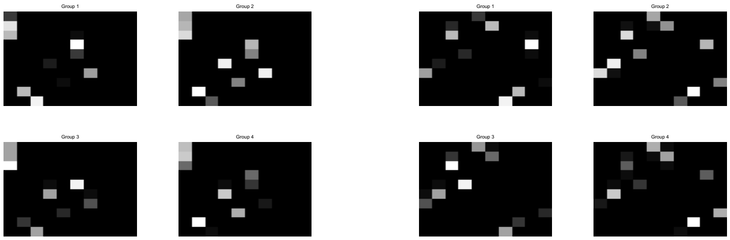
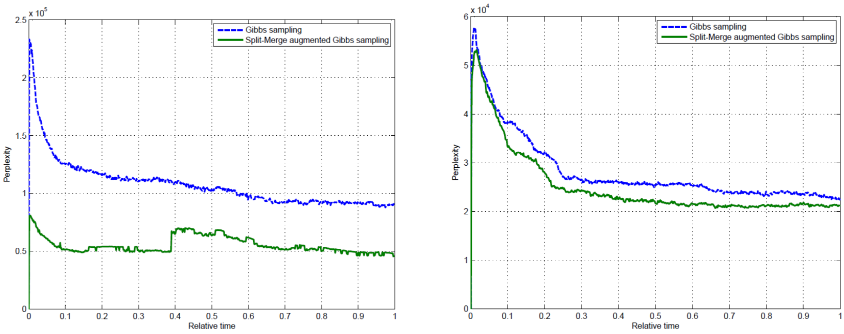


Fig. 4. Confusion matrix for topic mixtures for the four synthetic groups. Naive Gibbs sampler is in left and the split-merge augmented sampler is in right

4.2 Document Corpus

We used NIPS abstract data and 20 News Group data to study the convergence of our proposed method. NIPS0-12 data is a collection of abstracts published in the NIPS conference from the year 1988-1999. We select 1392 abstracts consisting of 263K words. The Dirichlet prior is set at $Dir(0.5)$. Both the Gibbs sampler and our sampler was initialized with the same initial topic distribution. We used random 80% of the data for topic modelling and the rest 20% data for perplexity computation. We run them for the same time and plot the the perplexity at each iteration in Fig 5a.



(a) NIPS corpus

(b) 20 News Group corpus

Fig. 5. Perplexity on the held-out data between the Split-Merge augmented Gibbs sampler and the Gibbs sampler on (a) NIPS corpus and (b) on 20 News Group corpus

The 20 News Group data contains 16242 documents with vocabulary size of 100. The Dirichlet parameter is set at 0.05. Similar to above setting, we learn our model with a random set of 80% of documents and the remaining 20% are used for perplexity computation. Both the Gibbs and our algorithm are run with the same initialization. Perplexity at each iteration is reported in Fig 5b. Superior perplexity is observed, although the algorithms ran for the same time.

5 Conclusion

In this paper we proposed a novel split-merge algorithm for HDP based on the direct conditional assignement of words-to-topics. The incremental Gibbs sampler can often be slow to mix and may often fail to provide a good posterior estimate in a limited time. The split-merge sampler with its ability to make a bigger move across the state-space mixes faster and often lead to very good posterior estimates. We experimented on both synthetic and real world data and demonstrate the convergence speedup of the proposed combined Gibbs and split-merge sampler over the plain Gibbs sampling method.

References

1. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
2. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of MachineResearch 3, 993–1022 (2003)
3. Jain, S., Neal, R.: 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)
4. Wang, C., Blei, D.: A split-merge mcmc algorithm for the hierarchical dirichlet process. Arxiv preprint arXiv:1201.1657 (2012)
5. Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2(6), 1152–1174 (1974)
6. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica 4(2), 639–650 (1994)
7. Teh, Y., Jordan, M.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N., Holmes, C., Müller, P., Walker, S. (eds.) Bayesian Nonparametrics: Principles and Practice, p. 158. Cambridge University Press (2009)
8. Xing, E., Sohn, K., Jordan, M., Teh, Y.: Bayesian multi-population haplotype inference via a hierarchical dirichlet process mixture. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 1049–1056. ACM (2006)
9. Li, L., Fei-Fei, L.: Optimol: automatic online picture collection via incremental model learning. International Journal of Computer Vision 88(2), 147–168 (2010)
10. Dahl, D.: Sequentially-allocated merge-split sampler for conjugate and nonconjugate dirichlet process mixture models. Journal of Computational and GraphicalStatistics (2005)