Parallel Sampling of HDPs using Sub-Cluster Splits

Jason Chang & John W. Fisher III

Overview

A Hierarchical Dirichlet Process (HDP) models groups of data with shared cluster statistics. HDPs are used in many applications such as document analysis [10], computer vision [8], and as priors for HMMs [5].

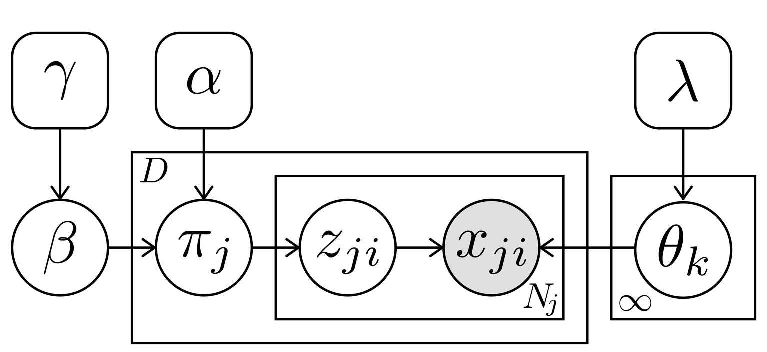
We extend the work of [3] on using Sub-Clusters in DPs to HDPs while addressing some distinct obstacles. Unlike [3], we show that the proposed ***global*** split and merge moves can drastically improve convergence.

	Infinite Model	Nonconjugate Priors	Parallel	Local Split Merge	Global Split Merge
Chinese Rest. Fran. [10]	☑	∗	☐	☐	☐
Direct Assignment [10]	☑	∗	☐	☐	☐
SAMS [4]	☑	☐	☐	☑	☐
FSD [5]	☐	☑	☑	☐	☐
Hog-Wild [7]	☑	☐	☑	☐	☐
Super-Clusters [9]	☑	∗	☑	☐	☐
Sub-Cluster Method	☑	☑	☑	☑	☑

∗ potentially possible by adapting the DP Metropolis-Hastings framework of [10]

Hierarchical Dirichlet Processes

We make use of the Direct Assignment representation of an HDP from [10]. The corresponding graphical model is below.

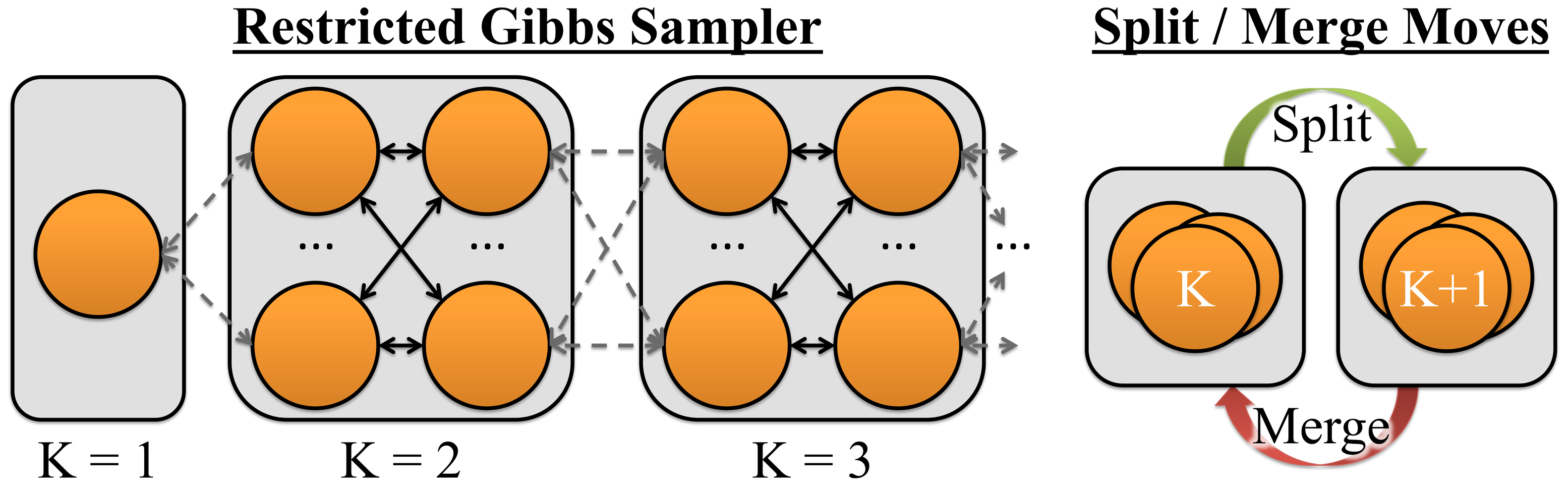


$$\beta \sim \mathrm{GEM}(1,\gamma)$$
$$\pi_j \sim \mathrm{DP}(\alpha,\beta)$$
$$z_{ji} \sim \mathrm{Cat}(\pi_j)$$
$$\theta_k \sim f_\theta(\theta;\lambda)$$
$$x_{ji} \sim f_x(x_{ji};\theta_{z_{ji}})$$

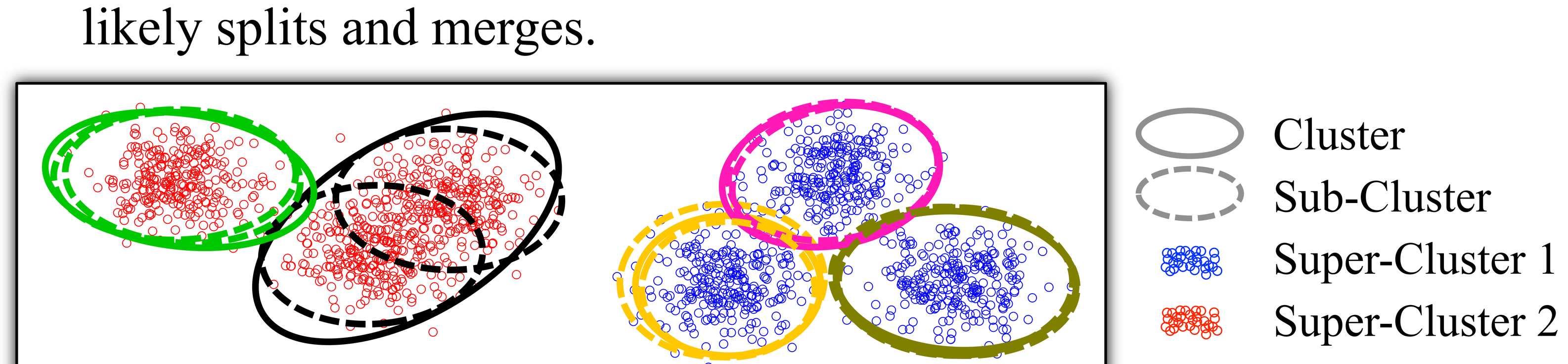
DP Sub-Clusters

The DP Sub-Clusters algorithm exploited the following two properties.

1. Combining a restricted Gibbs sampler (that does not create new clusters) with split/merge moves results in an ergodic Markov chain.



2. Augmenting the sample space with sub-clusters helps to propose likely splits and merges.

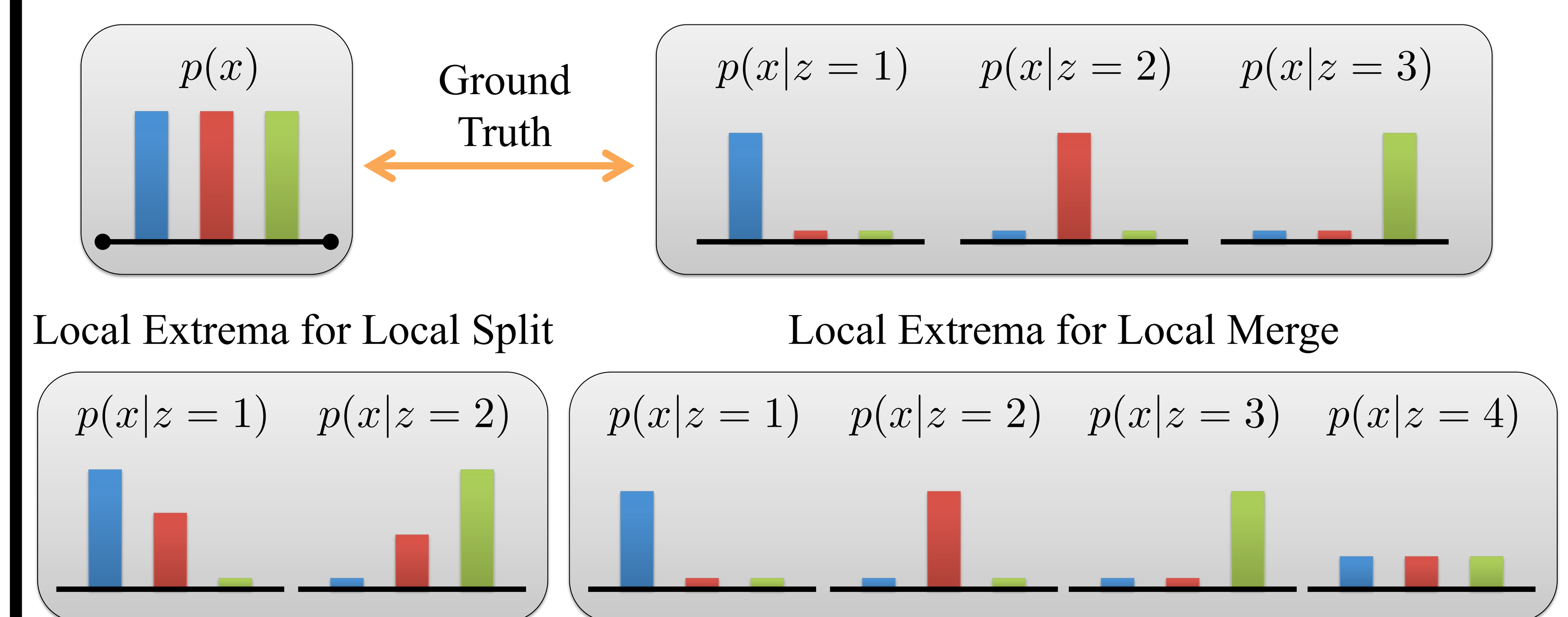


Overlapping Distributions

Overlapping distributions often result in rejecting "local" splits and merges because they are too restrictive.

Local splits and merges only act on a subset of the data.

- A local split only reassigns data belonging to the cluster being split.
- A local merge simply merges two clusters into one.



Solution: Propose global splits and merges which act on all the data.

- A global split splits one parameter into two new parameters and reassigns all data.
- A global merge combines two parameters into one new parameter and reassigns all data.

Calculating Joint Probabilities

Splits and merges require calculating the full, joint probability for Metropolis-Hastings, which is unknown.

$$p(\beta)\left[\prod_k p(\theta_k)\right]\left[\prod_j p(\pi_j|\beta)\prod_i p(z_{ji}|\pi_j)p(x_{ji}|z_{ji},\theta)\right]$$

$$\stackrel{?}{=} p(\beta,z)\left[\prod_k p(\theta_k)\right]\left[\prod_j p(\pi_j|\beta,z)\prod_i p(x_{ji}|z_{ji},\theta)\right]$$

Assuming that K non-empty clusters exist in z, we show that

$$p(\beta,z) = \gamma\beta_{K+1}^{\gamma-1}\prod_{k=1}^{K}\beta_k^{-1}\left[\prod_j \frac{\Gamma(\alpha)}{\Gamma(\alpha+n_{j\cdot\cdot})}\prod_{k=1}^{K}\frac{\Gamma(\alpha\beta_k+n_{j\cdot k})}{\Gamma(\alpha\beta_k)}\right]$$

HDP Sub-Clusters

The remaining formulation follows closely with the DP Sub-Clusters algorithm [3].

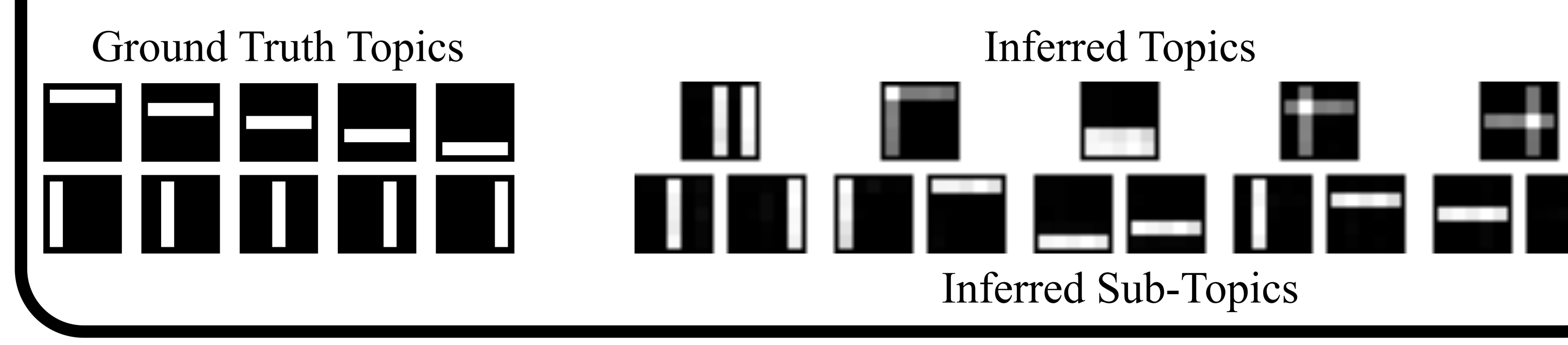
Restricted Gibbs Sampling – Instantiate and sample all variables in the HDP, but do not allow the creation of new clusters. Infer 2 sub-clusters for each regular-cluster within the restricted Gibbs sampling.

Propose Splits and Merges – Use the inferred sub-clusters to propose local and global splits and merges.

Code available at **http://people.csail.mit.edu/jchang7/**

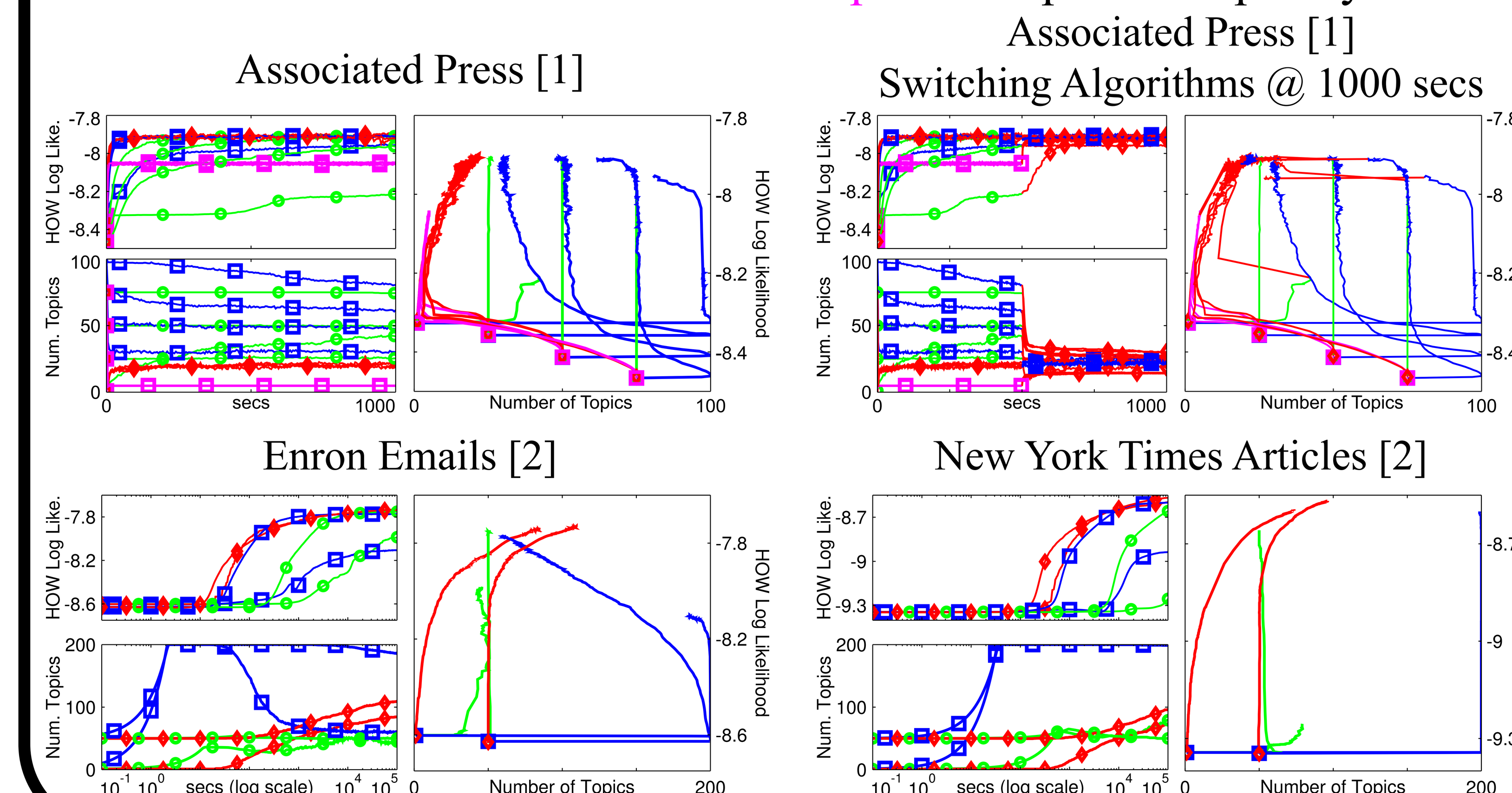
Visualizing Sub-Topics

We generate 10 ground truth topics from the "bars" dataset [6]. Inferred results when initializing to 5 topics and now proposing splits are below.



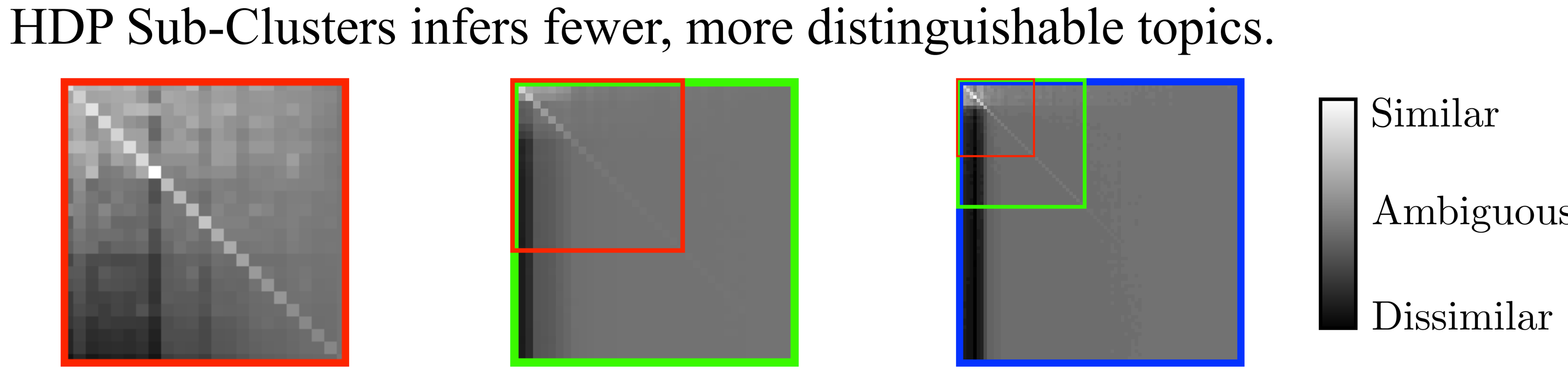
Topic Modeling Results

We compare the HDP Sub-Clusters algorithm with the Finite Symmetric Dirichlet approximation of [5] and the Direct Assignment sampler of [10] on three datasets. For each algorithm, we plot the held-out word (HOW) log likelihood and the number of topics vs. time. Notice that HOW log likelihood does not adequately determine convergence. Notice also that the HDP Sub-Clusters without Global Splits also performs poorly.



Topic Modeling Visualizations

Confusion matrices on the Associated Press dataset are shown below. HDP Sub-Clusters infers fewer, more distinguishable topics.



HDP Sub-Clusters   Direct Assignment   Finite Symmetric Dirichlet

Word clouds inferred from the New York Times dataset:

"Basketball"   "Wine"   "War on Terror"   "Cloning"

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. JMLR, 2003.
[2] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.
[3] J. Chang and J. W. Fisher, III. Parallel sampling of DP mixture models using sub-clusters splits. NIPS 2013.
[4] D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. Technical report, University of Wisconsin – Madison, 2003.
[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An HDP-HMM for systems with state persistence. ICML, 2008.
[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. National Academy of Sciences, 2004.
[7] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. JMLR, 2009.
[8] E. B. Sudderth. Graphical Models for Visual Object Recognition and Tracking. PhD thesis, MIT, 2006.
[9] S. Williamson, A. Dubey, and E. P. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. ICML, 2013.
[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. JASA, 2006.