Parallel Sampling of DP Mixture Models using Sub-Cluster Splits

Jason Chang & John W. Fisher III

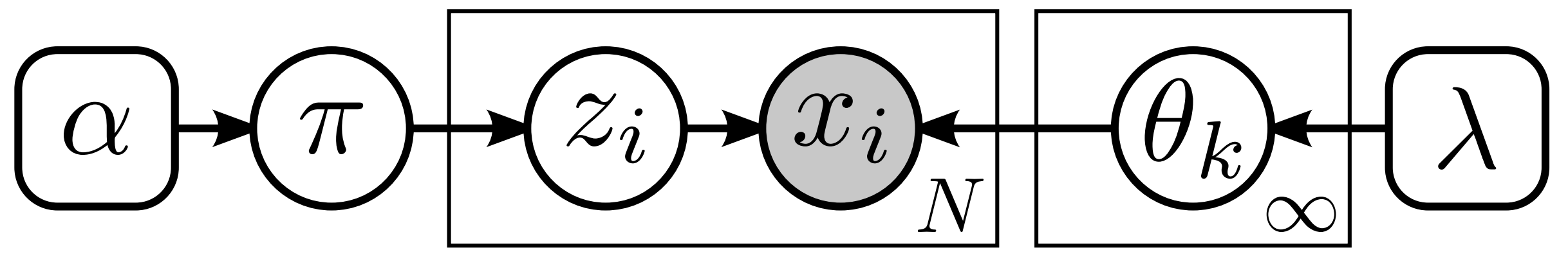
Overview

Dirichlet process mixture models (DPMMs) are widely used to model grouped data. Current Markov chain Monte Carlo (MCMC) sampling algorithms suffer from scalability, slow convergence, and/or require approximations. We present a new method that overcomes these issues.

	Exact Model	Splits & Merges	Intra-Cluster Parallel	Inter-Cluster Parallel
Collapsed Weight	☑	□	□	□
Finite Approx. [3,4]	□	□	□	☑
Slice Sampling [2,7]	☑	□	□	☑
RGSM [5] / SAMS [1]	☑	☑	□	□
Super-Clusters [6,8]	☑	□	☑	□
Sub-Cluster Method	☑	☑	☑	☑

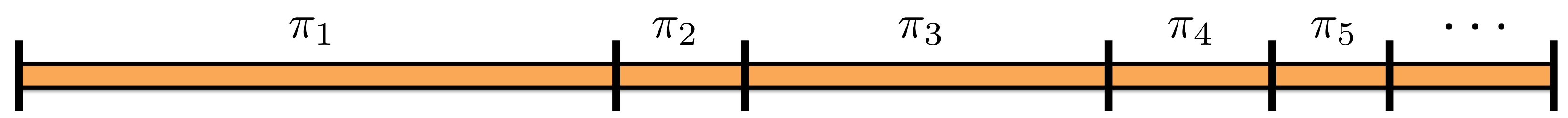
DP Mixture Models

The graphical model for a DPMM is shown below



$$\pi \sim \mathrm{GEM}(1,\alpha)$$
$$z_i \sim \mathrm{Cat}(\pi)$$
$$\theta_k \sim f_\theta(\theta;\lambda)$$
$$x_i \sim f_x(x_i;\theta_{z_i})$$

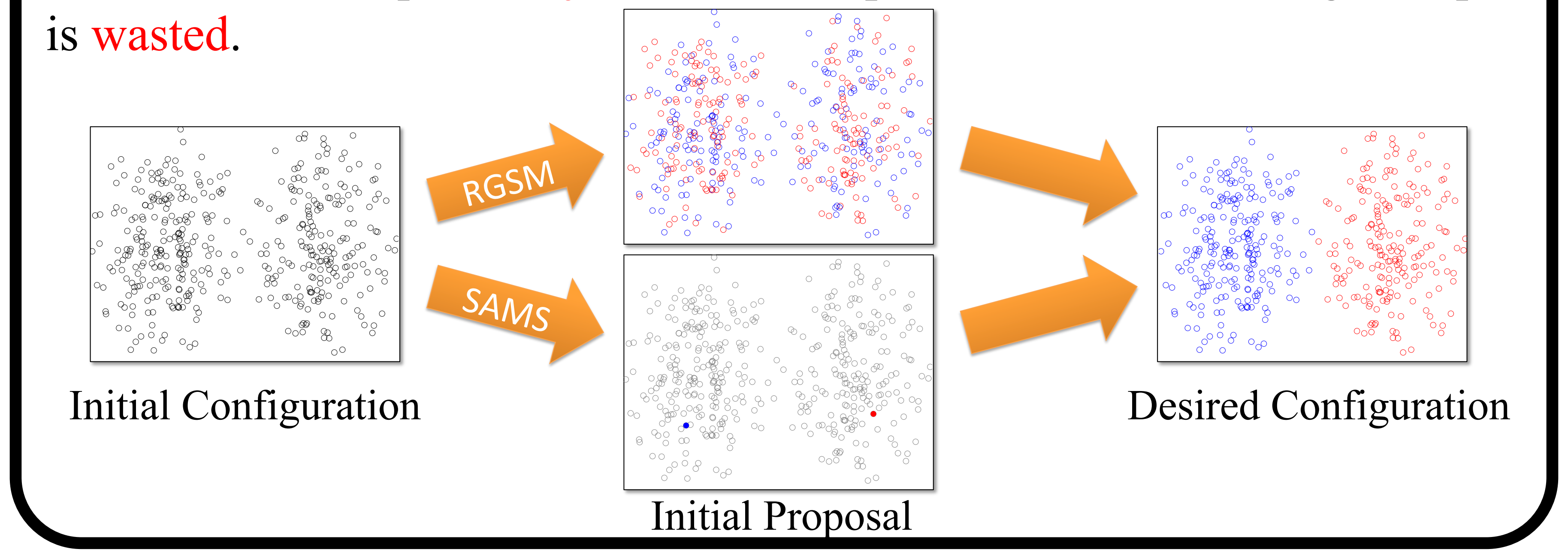
Chinese Restaurant Process based sampling methods marginalize over the infinite-length cluster weights, π. These methods cannot be parallelized.



Finite approximations ([3,4]) and slice-sampling ([2,7]) can be used to instantiate weights. Cluster assignments can then be sampled in parallel.

Split / Merge Methods

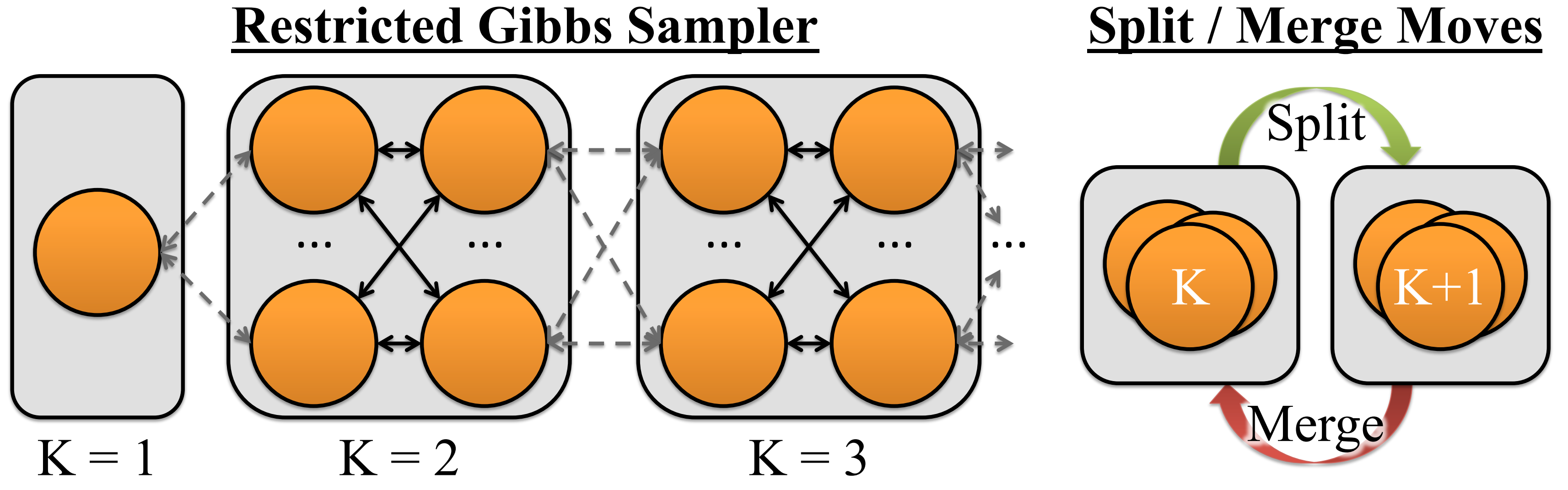
RGSM [5] construct a split by randomly partitioning a cluster and running restricted Gibbs iterations on the data points. SAMS [1] constructs a split by sequentially assigning points to one of two new clusters. Both algorithms then use a Metropolis-Hastings accept/reject framework. If a split is rejected, all computation in constructing the split is wasted.



Mixing Non-Ergodic Chains

MCMC sampling algorithms typically consider the case where the induced Markov chain is ergodic and where the transition distribution satisifed detailed balance. One key observation of this work is that mixing multiple non-ergodic chains can result in a ergodic chain.

For mixture models, a restricted Gibbs sampler that does not allow the introduction of new clusters can be mixed with any split/merge sampler.



Restricted Gibbs-Sampler

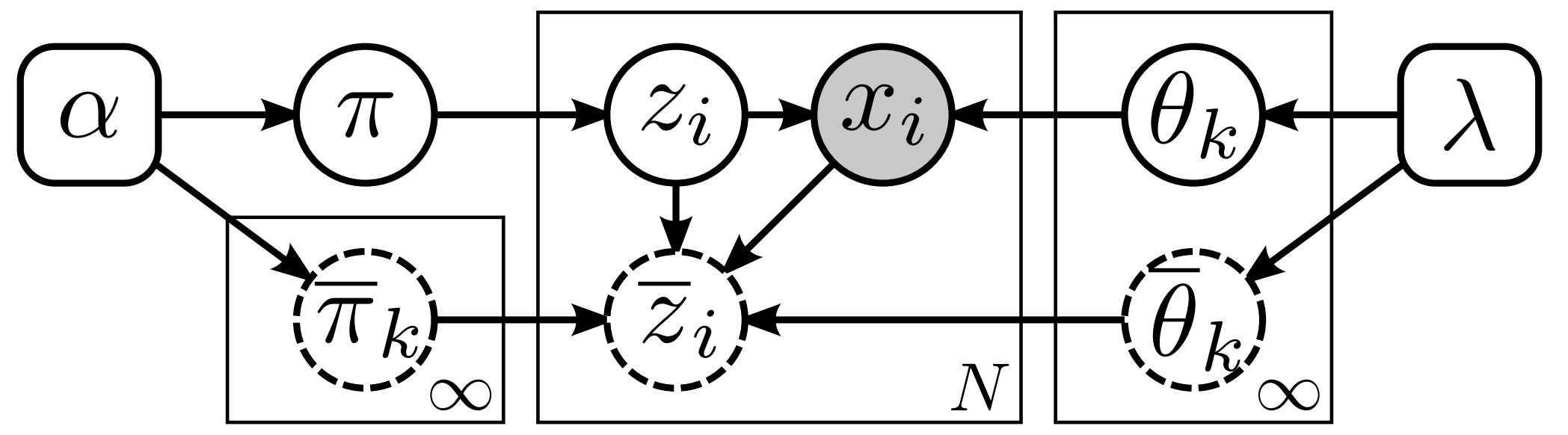
Because the restricted Gibbs sampler does not need to consider creating new clusters, it can be efficiently and exactly parallelized. The posterior cluster weights are then distributed according to

$$(\pi_1,\ldots,\pi_K,\tilde{\pi}_{K+1}) \sim \mathrm{Dir}(N_1,\ldots,N_K,\alpha)$$

Sampling cluster assignments can be parallelized without approximations since the K cluster weights and parameters are instantiated.

Fitting Sub-Clusters

We augment the space with auxiliary variables representing sub-clusters.



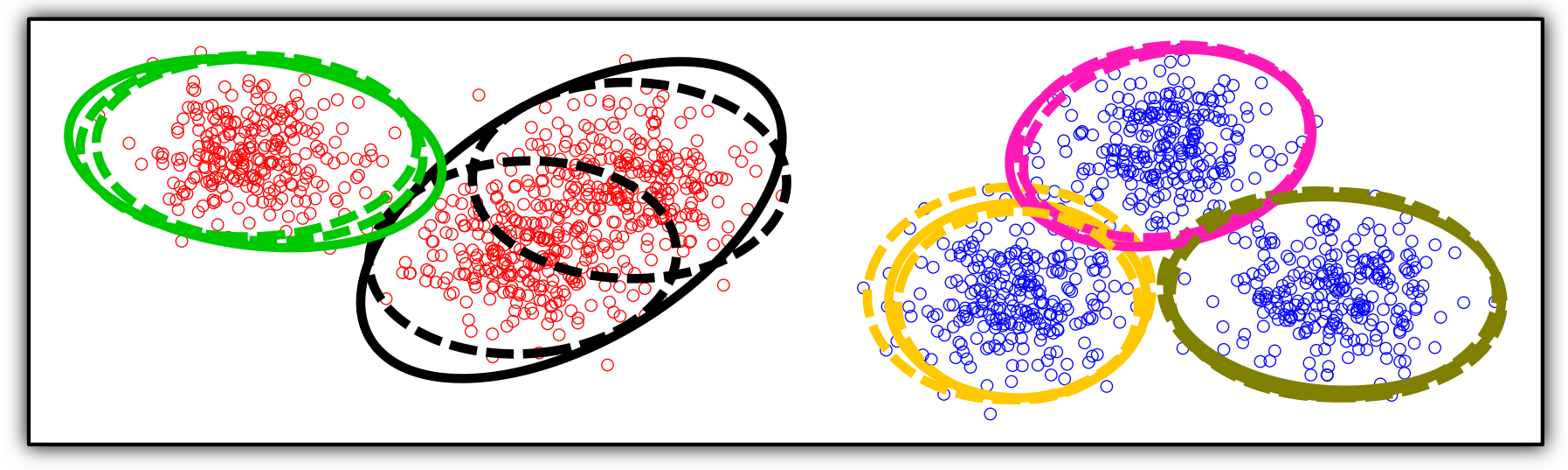
Sub-cluster variables are denoted with a bar and have similar meaning to regular-clusters.

Each regular-cluster is augmented with two sub-clusters, denoted the "left" and "right" sub-clusters. Suitable priors are chosen for the sub-cluster parameters such that their posterior distributions are as follows:

$$p(\overline{\pi}_{k\ell},\overline{\pi}_{kr}|\bullet) = \mathrm{Dir}(\overline{\pi}_{k\ell},\overline{\pi}_{kr};\tfrac{\alpha}{2}+N_{k\ell},\tfrac{\alpha}{2}+N_{kr})$$

$$p(\overline{\theta}_{ks}|\bullet) \propto f_\theta(\overline{\theta}_{ks};\lambda) \prod_{\{i;z_i=k,\overline{z}_i=s\}} f_x(x_i;\overline{\theta}_{ks}), \quad s\in\{\ell,r\}$$

$$p(\overline{z}_i|\bullet) \propto \overline{\pi}_{z_i\overline{z}_i} f_x(x_i;\overline{\theta}_{z_i\overline{z}_i})$$



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

Sub-Cluster Splits

Since sub-clusters contain likely splits, we propose to split a cluster into its two sub-clusters. Conditioned on the new cluster assignments, all other regular-cluster parameters are proposed from their posterior distributions. New auxiliary variables are deferred to the restricted Gibbs sampler.

Random Splits / Merges

Because a sub-cluster split is constructed deterministically from its sub-clusters, the reverse merge move is rejected with very high probability. We therefore include a set of "random" split/merge moves. A "random" split is sampled independent of data from a 2-component Dir-Mult.

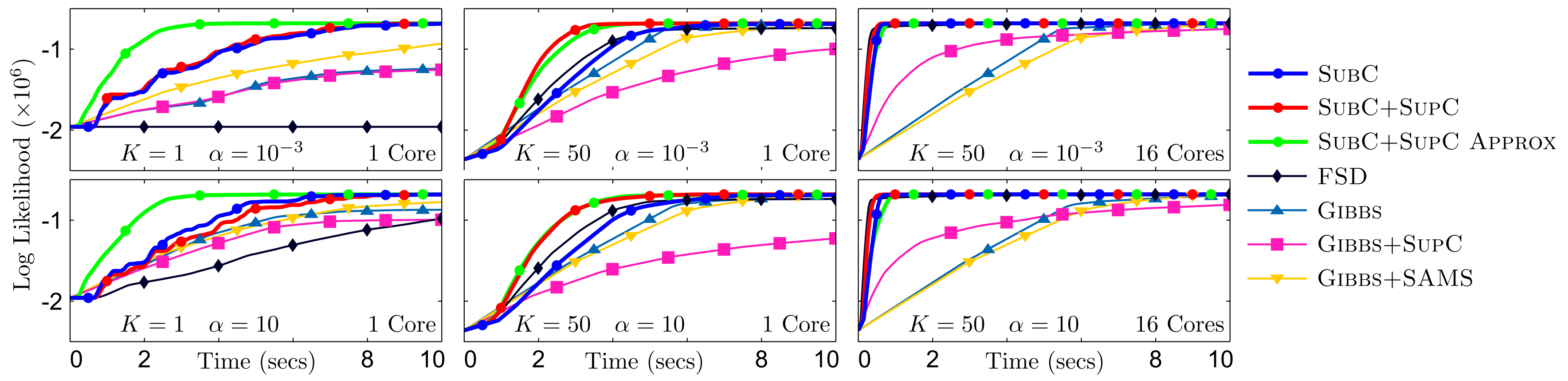
	Good Splits	Good Merges
Sub-Cluster Split/Merge Moves	☑	□
Random Split/Merge Moves	□	☑

While a random split is typically meaningless, the random merge is often accepted. Thus, the random split/merge moves complement the sub-cluster split/merge moves.

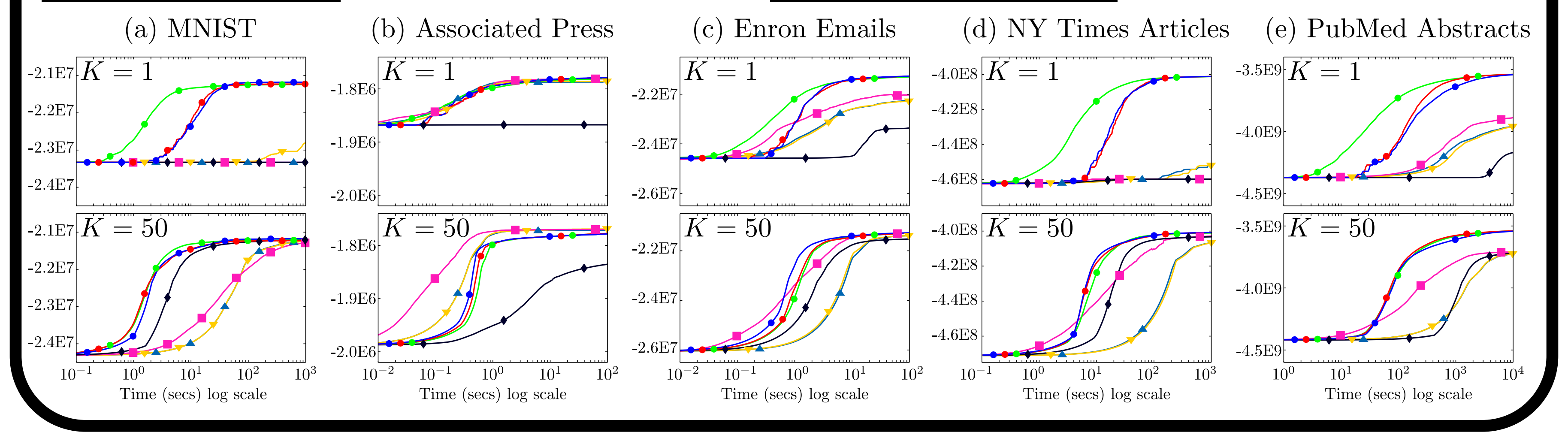
Results

We consider three new samplers: the basic method (SubC), with data-dependent super-cluster (SubC+SupC), and an approximate method (SubC+SupC Approx). We compare with the Finite Symmetric Dirichlet (FSD) of [4], a collapsed Gibbs sampler (Gibbs), the super-clusters (Gibbs+SupC) of [6], and the split/merge (Gibbs+SAMS) work of [1].

Synthetic Gaussian Data with and without Parallelization



Real Gaussian Data **Real Multinomial Data**


(a) MNIST (b) Associated Press (c) Enron Emails (d) NY Times Articles (e) PubMed Abstracts

[1] D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. Technical report, University of Wisconsin - Madison Dept. of Statistics, 2003.
[2] S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. Statistical Science, 2013.
[3] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association,2001.
[4] H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. Canadian Journal of Statistics, 2002.
[5] 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, 2000.
[6] D. Lovell, R. P. Adams, and V. K. Mansingka. Parallel Markov chain Monte Carlo for Dirichlet process mixtures. In Workshop on Big Learning, NIPS, 2012.
[7] O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika, 2008.
[8] S. A. Williamson, A. Dubey, and E. P. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In ICML, 2013.