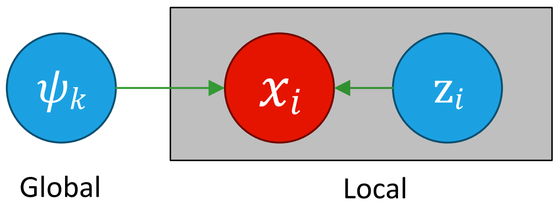
Fast Sampling Algorithms for Sparse Latent Variable Models

Manzil Zaheer, Amr Ahmed, Ha Loc Da, Jay-Yoon Lee, Sujith Ravi and Alexander J Smola

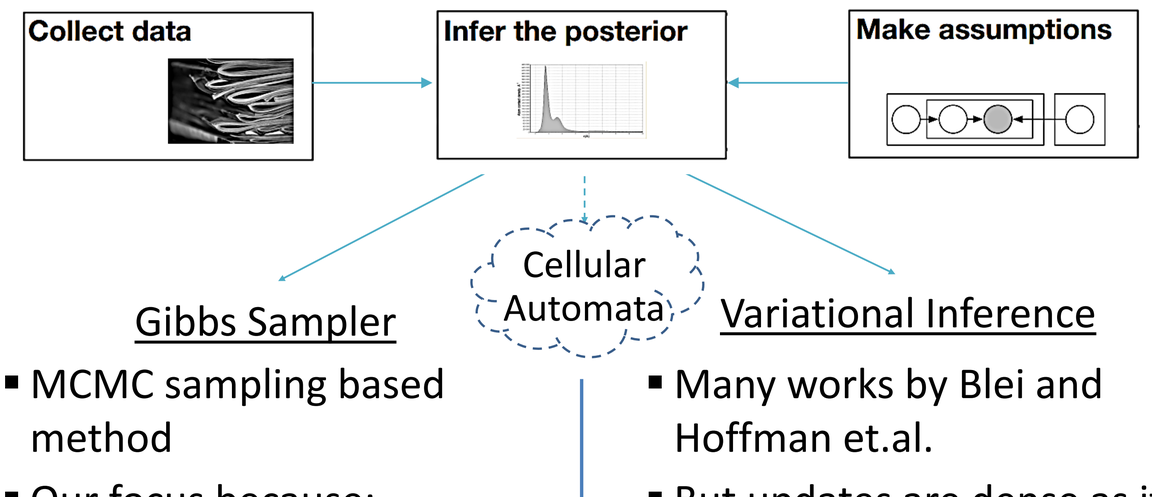
Latent Variable Models

- Latent variable models have become a staple of statistical modelling: clustering, topic models, subspace estimation
- Diverse applications range from
  - Organising text documents (e.g. news) and images, to
  - Predicting user behavior (e.g. click patterns), to
  - Targeting ads
- Versatile tools for discovering the hidden thematic structure of objects in a human-understandable format
- Share common structure
  - Global variables: themes that pervade the dataset
  - Local variables: labels for data point



Inference Strategy

- Inference is the process of estimating posterior distribution (or the most likely assignment) of all the latent variables
- Unfortunately, the task is intractable and even approximate methods can not be scaled easily due to the data dependencies introduced by global state.



Gibbs Sampler

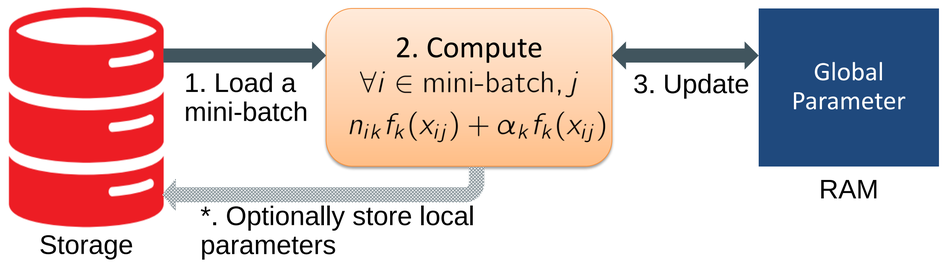
- MCMC sampling based method
- Our focus because:
- Leverage inherent sparsity present
- Use of smart data-structures for speeding up sampling

Variational Inference

- Many works by Blei and Hoffman et.al.
- But updates are dense as it is an expectation over all labels
- Lots of pressure on CPU-RAM bandwidth

Implementation

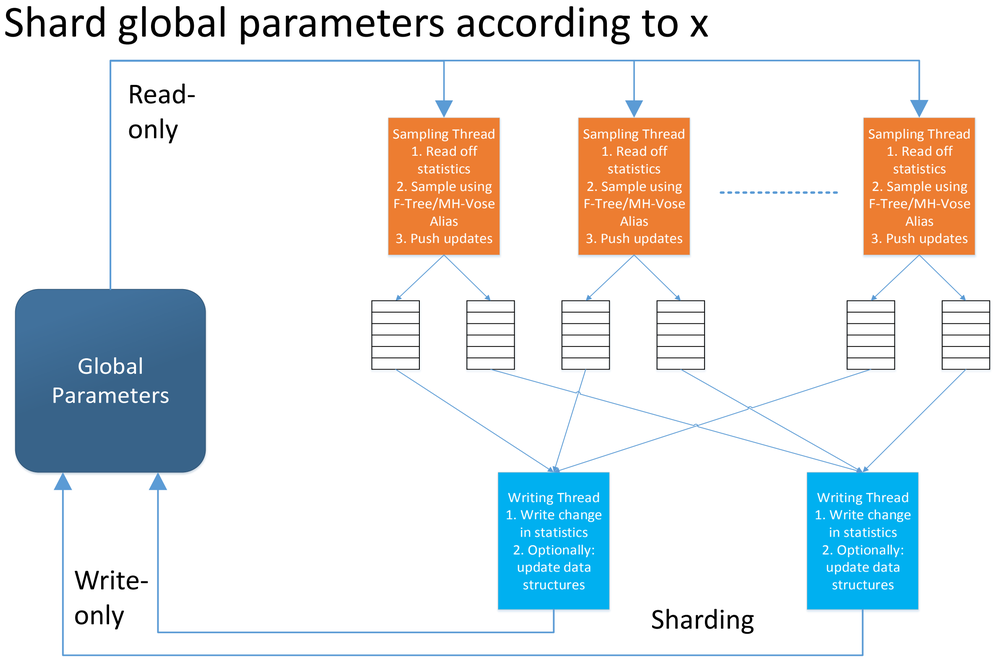
- Global state typically fits into RAM
- Managing a massive local state
  - Out of core storage
  - Effective since the typical schedule for the Gibbs sampler iterates over the variables in a fixed order



- Leveraging sparsity and better data-structures
  - After gory math, for most Latent variable models the compute step involves drawing from the conditional:

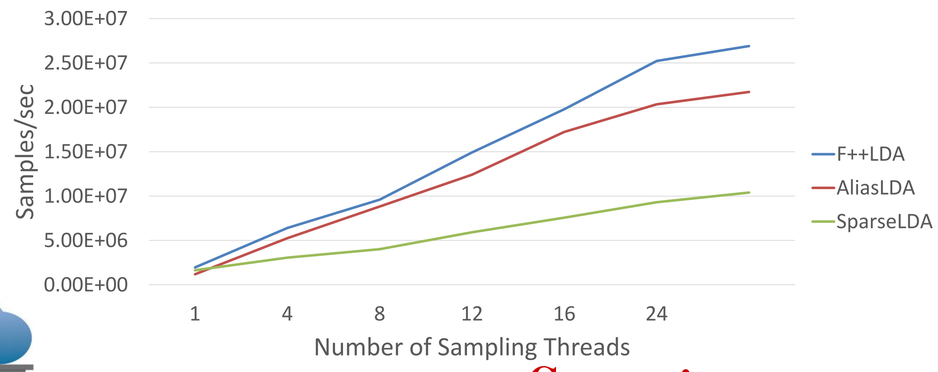
$$p(z_{ij} = k|\text{rest}) \propto \underbrace{n_{ik} f_k(x_{ij})}_{\text{sparse}} + \underbrace{\alpha_k f_k(x_{ij})}_{\text{slowly-varying}}$$

  - Divide & Conquer! Brute force sampling for sparse term
  - Smart tricks for almost constant time sampling like Alias Method or Fenwick tree for global dense slowly varying term
- Lock-free implementation
  - Sampling thread only reads from global, samples and pushes changes to circular buffers
  - Writing thread pulls from the circular buffers and modifies global
  - Shard global parameters according to x



Performance

- Throughput
  - Single instance of AWS C4.8xlarge (36 vCPI, 60 GiB RAM)



- Accuracy
  - Per-word log-likelihood after 1000 iteration of each method

Dataset	Tokens	Vanilla LDA	SparseLDA	aliasLDA	F++LDA
ACM	12M	-7.82	-7.81	-7.83	-7.82
NY Times	100M	-7.91	-7.91	-7.92	-7.91
Reuters	105M	-7.34	-7.34	-7.35	-7.35
PubMed	738M	-6.96	-6.92	-6.96	-6.95
Wikipedia	1,418M	-7.58	-7.58	-7.60	-7.59

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington