Fast parallel MCMC for large scale topic models

Måns Magnusson, Leif Jonsson, Mattias Villani and David Broman
Linköping University, Ericsson AB, KTH Royal Institute of Technology, UC Berkley
mans.magnusson@liu.se

LINKÖPING UNIVERSITY

Abstract

Latent Dirichlet allocation (LDA) is a model widely used for unsupervised probabilistic modeling of text and images. MCMC sampling from the posterior distribution is typically performed using a collapsed Gibbs sampler or fast Metropolis-Hastings samplers. We propose a sparse parallel partially collapsed Gibbs and Metropolis-Hastings samplers. We show on well-known corpora that the expected increase in statistical inefficiency from only partial collapsing is smaller than commonly assumed, and can be more than compensated by the speed-up from parallelization for larger corpora. We also prove that the partially collapsed sampler scale well with the size of the corpus. The algorithm is fast, efficient, and exact.

Introduction

Topic modeling or Latent Dirichlet allocation (LDA) is an immensely popular way to model text probabilistically using the following generative model:

1. For each topic k in K:
   (a) $\phi_k \sim \text{Dir}_V(\beta)$
2. For each document d in D:
   (a) $\theta \sim \text{Dir}_K(\alpha)$
   (b) For each word token i in document d:
      i. $z \sim \text{Multinomial}_K(\theta)$
      ii. $w \sim \text{Multinomial}_V(\phi_z)$

The standard sampling scheme for the topic indicators is the collapsed Gibbs sampler of [1]:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \underbrace{\frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + V\beta}}_{topic-word} \underbrace{\left(n^{(d_i)}_{-i,j} + \alpha\right)}_{document-topic},$$

where $n^{(w)}$ is the counts of $\mathbf{z}$ by topic and word type and $n^{(d)}$ is the counts of $\mathbf{z}$ by document and topic.

This sampler is sequential in nature due to conditioning on *all other* topic indicators in the whole corpus.

References

[1] Griffiths, T. L., Steyvers, M., 2004. *Finding scientific topics.*, Proceedings of the National Academy of Sciences 101 (suppl 1), 5228-5235.

[2] Li, A. Q., Ahmed, A., Ravi, S., Smola, A. J., 2014. *Reducing the sampling complexity of topic models.* In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 891-900.

[3] Newman, D., Asuncion, A., Smyth, P., Welling, M., 2009. *Distributed algorithms for topic models.* The Journal of Machine Learning Research 10, 1801-1828.

[4] Yao, L., Mimno, D., McCallum, A., 2009. *Efficient methods for topic model inference on streaming document collections.* Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 09, 937.

[5] Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E. P., Liu, T.-Y., Ma, W.-Y., Dec. 2014. *LightLDA: Big Topic Models on Modest Compute Clusters.* ArXiv e-prints.

Partially collapsed samplers

We propose partially collapsed approach where only Θ is integrated out. This gives the following basic sampler where we first sample the topic indicators for each document in parallel as

$$p(z_i = j|\mathbf{z}^{(d)}_{-i}, \Phi, w_i) \propto \phi_{j,w} \cdot \left(n^{(d_i)}_{-i,j} + \alpha\right)$$

and then sample the rows of Φ in parallel as

$$\phi_k \sim \text{Dir}(n^{(\mathbf{w})}_k + \beta).$$

Reducing sampling complexity
We construct Gibbs-PC-LDA by using Walker-Alias multinomial sampling from [2] and exploit that document local sparsity can reduce the complexity of the sampler by decomposing

$$p(z_i = j|\mathbf{z}^{(d)}_{-i}, \Phi, w_i) \propto \underbrace{\phi_{j,w} \cdot \alpha}_{a} + \underbrace{\phi_{j,w} \cdot n^{(d_i)}_{-i,j}}_{b}$$

where a is sampled using Walker-Alias tables and b by iterating over existing topics ($n^{(d_i)}_{-i,j} > 0$).

A light partially collapsed sampler
In a way similar to light-LDA [5] we can create a partially collapsed sampler (Light-PC-LDA) with using a cyclical Metropolis-Hastings algorithm that is altering between the word proposal

$$p_w(z^*) \propto \phi_{z^*,v}$$

and the document proposal

$$p_d(z^*) \propto \alpha + n^{(d_i)}_{-i,j}$$

Related Work

Parallelism
The Approximate Distributed LDA (AD-LDA) [3] is currently the most common way to parallelize topic models. The idea is that each processor works in parallel with a given set of topic counts in the word-topic count matrix $n^{(w)}$ and the different processors are synced after each iteration. The resulting algorithm is not guaranteed to converge to the target posterior, and will in general not do so.

Reducing sampling complexity
There are three main approaches to speed up sampling in topic models. All methods are based on the collapsed approach where both Θ and Φ is integrated out. Sparse-LDA is a Gibbs sampler while Alias- and Light-LDA are based on Metropolis-Hastings sampling.

Sampler	Complexity
Sparse-LDA [4]	$O\left(\sum_i^N \max(K_{d(i)}, K_{w(i)})\right)$
Alias-LDA [2]	$O\left(\sum_i^N K_{d(i)}\right)$
Light-LDA [5]	$O(N)$

Our samplers	Complexity
Gibbs-PC-LDA	$O\left(\sum_i^N K_{d(i)}\right)$
Light-PC-LDA	$O(N)$

Data

We have used two corpora in the experiments.

Corpus	N	D	V
NIPS	1.9m	1499	12375
PUBMED 10%	78.5m	820000	50000

Experiments

The experiments were conducted using 2 socket 4-core Intel Xeon E5520 processors at 2.2GHz and 2 socket 8-core Intel Xeon E5-2660 "Sandy Bridge" processors at 2.2GHz. The code can be found at: **https://github.com/lejon/PartiallyCollapsedLDA**

Figure 1. Effect of parallelism using AD-LDA and PC-LDA sampling (NIPS 100 topics)

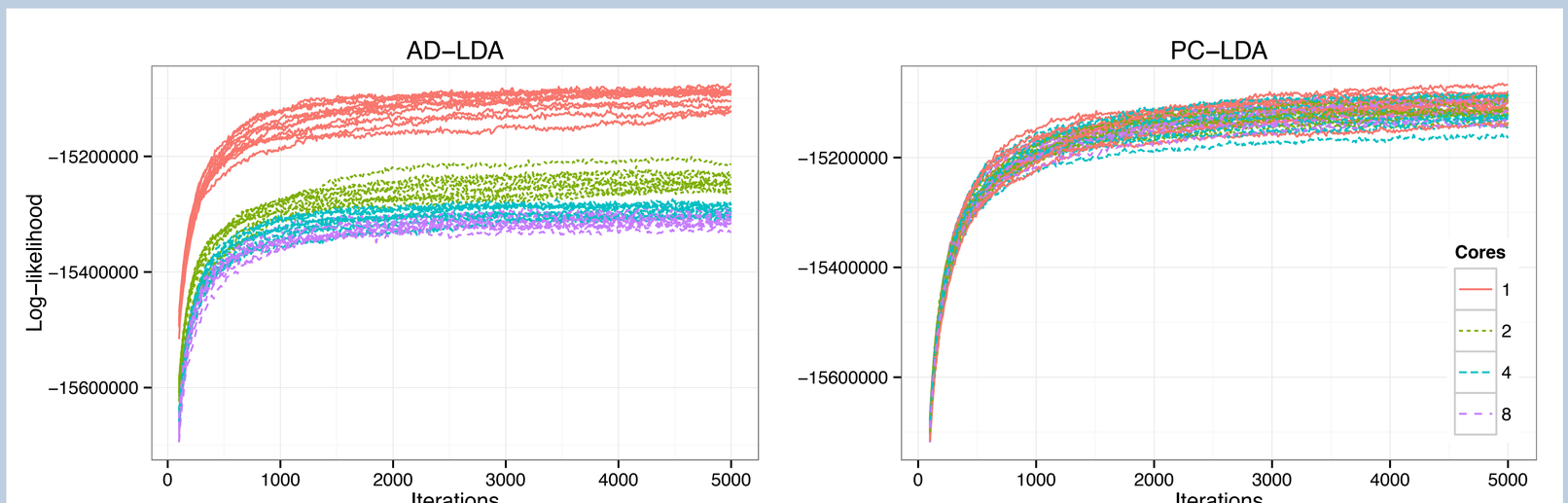

Figure 2. Time to convergence PUBMED 10% 10 topics (left) and 1000 topics (right)