Towards a Highly Efficient Online Inference Algorithm for Latent Dirichlet Allocation

Rashidul Islam
IS Department, UMBC
islam.rashidul@umbc.edu

James R. Foulds
IS Department, UMBC
jfoulds@umbc.edu

Topic models such as latent Dirichlet allocation (LDA) are powerful tools for analyzing today's massive, constantly expanding digital text information by representing high dimensional data in a low dimensional subspace. We can uncover the main themes of a corpus by using LDA to help organize, understand, search, and explore the documents. The central computational problem of LDA is to infer the latent variables, distributions over vocabularies for topics, and distributions over topics for documents, given an observed set of documents. In this work, we propose a highly efficient and scalable algorithm for LDA inference.

Traditional inference techniques such as variational Bayes and collapsed Gibbs sampling do not readily scale to corpora containing millions of documents. To scale up inference, the main approaches are distributed algorithms (Newman et al., 2008) and stochastic algorithms (Hoffman et al., 2010). Stochastic algorithms, such as stochastic variational inference (SVI), operate in an online fashion, and hence do not need to see all of the documents before updating the topics, so they can be applied to corpora of any size, without expensive distributed hardware (Hoffman et al., 2010). The "collapsed" representation of LDA is also frequently important, as it leads to faster convergence, efficient updates, and lower variance in estimation. The stochastic collapsed variational Bayesian inference (SCVB0) algorithm, proposed by Foulds et al. (2013), combines the benefits of stochastic and collapsed inference.

Larger corpora typically support more topics, which brings the additional efficiency challenges of training larger models. This challenge has been addressed by exploiting sparsity to perform updates in time sublinear in the number of topics. The Metropolis Hastings Walker (MHW) method, developed by Li et al. (2014), scales well in the number of topics, and uses a collapsed inference algorithm, but operates in batch setting, which is not scalable to large corpora. The MHW algorithm uses an efficient data structure, known as an alias table, for sampling from the collapsed conditional distributions in amortized constant time, combined with a Metropolis-Hastings correction for the use of cached, slightly stale samples. A sparse variant of the SVI algorithm, SSVI, was proposed by Mimno et al. (2012), which is scalable to large numbers of topics, but does not fully exploit the collapsed representation of LDA.

In this work, we aim to develop an online inference algorithm for LDA which leverages stochasticity to scale well in the number of documents, sparsity to scale well in the number of topics, and which operates in the collapsed representation of LDA. We thereby combine the individual benefits of SVI, SSVI, SCVB0, and MWH into a single algorithm. Our approach is to develop a sparse version of SCVB0. Inspired by SSVI, we use a Monte Carlo inner loop to approximate the SCVB0 variational distribution updates in a sparse and efficient way, which we accomplish via the MHW method. To get full benefit from the sparsity assumption we apply a sparsification heuristic after processing a mini-batch of documents.

Surprisingly, preliminary results indicate that our method converges to a better solution compared to the traditional SCVB0 approach on the NIPS corpus, as shown in Figure 1 in the Appendix. We believe that the sparsification caused by our approach may lead the inference to more quickly reach certainty in the variational distributions over latent variables, which is the regime in which convergence is rapid. We are currently working on clever updates of model parameters so that we can achieve our targeted speed for scalable online inference, to efficiently process today's constantly expanding massive digital text data.

References

James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. 2013. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 446–454. ACM.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. 2014. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 891–900. ACM.

David Mimno, Matt Hoffman, and David Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. *Proceedings of the 29th International Conference on Machine Learning*.

David Newman, Padhraic Smyth, Max Welling, and Arthur U Asuncion. 2008. Distributed inference for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1081–1088.

A Preliminary Results

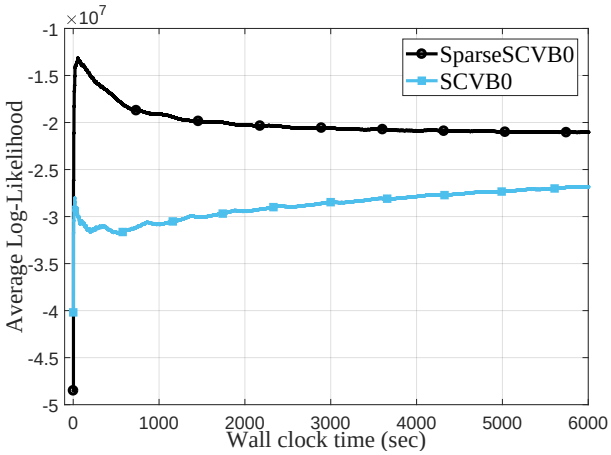

Figure 1: Comparison of our proposed MHW-based sparse SCVB0 and traditional SCVB0 indicates that our proposed sparse SCVB0 method converges faster than traditional SCVB0, and to a better solution.