Reviews For Paper

Paper ID 1382

Title Streaming Gibbs Sampling for LDA Model

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Rating	Strong accept
Reviewer confidence	Reviewer is knowledgeable
	This paper presents a streaming version of collapsed Gibbs sampling inference for LDA, which only looks at the dataset in minibatches, and thus does not have to save the "z" decisions for the entire dataset at once. Instead, it forgets old "z" assignments, and only maintains the topic-word sufficient statistics count tables; and it decays the old sufficient statistics when updating with new ones from the new minibatch. This corresponds to a nice Bayes updating viewpoint on learning from streaming data: the old data imposes a Dirichlet prior, and then you Bayes update on the new minibatch (an insight in the alternative approach they compare to, SVB by Broderick et al 2013). Interestingly, they can apply theoretical results on the consistency of this posterior estimation procedure ("conditional density filtering"). Experimental results show their approach performs favorably compared to other streaming inference approaches to LDA. They also present a "hogwild"-style distributed version (which has minibatches do inference with stale counts, then pushes updates whenever available, I think) and analyze its empirical performance. Overall this paper is great. It has nice results, and gives a concrete example of sampling methods for streaming Bayesian inference. While many recent versions of VB for topic models report better performance than CGS under limited computation resource scenarios, CGS remains very popular among practitioners because its flexibility and simplicity makes it easy to develop and implement new mixed-membership models (for just one of many examples, say the recent Paul and Dredze TACL 2015 paper, "Sprite" http://www.cs.jhu.edu/~mpaul/files/tacl2015_sprite.pdf). Therefore it's great to see the streaming approach extended to CGS. I give the caveat that I did not rigorously check the mathematical developments in this paper, and that while I am familiar with CGS for LDA, I am only passingly familiar with SVB, and have never seen CDF before (from the Guhaniyogi 2014 arxiv paper they cite). One thing that concerns me is the authors don't give CGS and CVB enough credit for having an online flavor. All the collapsed inference techniques are online in the sense that when you update the sufficient stats, you've already updated the parameters to analyze the next variable. The big computation argument for SGS is, rather, in the data access patterns. For each document, for each document, the observed data N_{dw} table, which presumably could be loaded from a slow non-RAM storage environment. CGS has to access a little more, but if you think about it, not a ton more: it also needs N_{dk} (the doc-topic counts, not too large) and also the topic-token assignments z^{(d)}. This is only a constant factor larger than the original data. You never have to load the topic-token assignments for the entire dataset at once! You

Detailed comments for the authors	could have streaming access to "z" in the same manner. The cumbersome aspect to this type of streaming implementation of CGS is, rather, the fact that **writes** are required when doing "z" updates; if you're storing them on a hard disk or other slow non-RAM storage setting, this could be a bottleneck without special care for with systems programming issues. I do think this is still a reasonable argument for the computational advantages of SGS over traditional CGS, but the authors should improve and be more careful about their arguments here. Furthermore, the authors also omit the fact that some researchers initialize CGS not with random initialization, but instead with progressive online updating, as if the data has never been seen before: start with literally 0 counts in all the suff stats tables, then build it up during the first iteration. Once you've completed the first iteration, your count tables now have their sums as the total number of tokens; all further iterations require the substraction step since they're now formally Gibbs sampling. I don't know what's a good reference for this or if anyone has given it a name, but I was under the impression it's used at least a little bit; I've done it myself and it seems to work better than random initialization. Anyways, this approach to the first iteration in CGS really is totally online just as much as SGS, since it doesn't require any access to old per-document sampling decisions. Using this type of initialization might close some of the gap between SGS and the version of CGS analyzed here. The qualitative analysis in 3.6 and 3.7 was nice to see, but not very convincing. I really love seeing qualitative analysis so I hesitate to criticize, but it's just not convincing. 3.6 isn't useful because it's only from one run of SVG and SGS each. Multiple runs of the same algorithm can give very different results, even qualitatively speaking. (e.g. see http://scholar.harvard.edu/files/dtingley/files/multimod.pdf). 3.7 isn't remarkable either. Any time you run CGS and look at a single topic over multiple iterations, it sometimes transitions between qualitatively-remarkable concepts in this way. Are we supposed to believe this conceptual transition from neural concepts to speech recognition concepts is a true semantic change in the data? It might just be a funny issue with how samplers work. The paper is well written in terms of argumentation and structure, but there are a substantial number of grammar errors. I've only pointed out a subset of them below. Finally, I'm puzzled by the leaving out of discussion about distributed Bayesian inference approaches. For example see the references here http://people.math.umass.edu/~conlon/bigdata.html including - Scott, S.L., Blocker, A.W., Bonassi, F.V. (2013) Bayes and Big Data: The consensus Monte Carlo Algorithm. Bayes 250. - Neiswanger, W., Wang, C., Xing, E. (2014) Asymptotically exact, embarrassingly parallel MCMC. arXiv:1311.4780v2. Minor notes - I like the pointing out of the fact that collapsed LDA is equivalent to using posterior mean estimates of the posterior dirichlets. - lines 187-198: authors should cite the Yao, Mimno, McCallum 2009 paper that originally introduced this breakdown. - line 236 (algo 1): make sure to be clear, in the text if not here, that these are posterior mean estimates. i've noticed confusion among EM/VB-oriented researchers about whether sampling approaches return a parameter estimate or not. the answer, of course, is they can if you want to, and these ratios are

	a type of posterior mean. - line 263: i think this is a typo, should be phi not theta ? - line 318: grammar error, sentence "Given ..." is an incomplete sentence - line 394: grammar error or several errors - line 518-522: might want to enrich notation to make the half-holdout structure within a document more clear, specifically that theta was estimated without looking at the second-half tokens. e.g. theta^{(firsthalf)} = EstimateTheta(w^{(firsthalf)}, phi) p(w^{(secondhalf)}) = phi theta^{(firsthalf)} (... i left out many terms there to give the idea) - line 672: "For visualization purpose," grammar error. delete "purpose"

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall Rating	Weak reject
Reviewer confidence	Reviewer is an expert
Detailed comments for the authors	This paper describes a streaming approach for collapsed Gibbs sampling in learning LDA. The method takes a subset of documents at each iteration, and use collapsed Gibbs sampling to infer the latent variables. The parameters are then aggregated to the global parameters with decay over time. Intuitively, it can be understood as "online" version for collapsed Gibbs sampling LDA like the work by Hoffman, et al., 2010. My concern with the paper is the lack of critical comparison and missing some technical details. 1) The paper proposes a streaming approach for Gibbs sampling, but did not show any comparison of the proposed model versus batch version of collapsed Gibbs sampling. One critically comparison of online models is against its batch counter part, both in terms of running time and convergence property, and it should not be difficult based on the paper context. 2) The baseline model for comparison used in the paper is the streaming variational Bayesian approach, which I think is less meaningful because a) the convergence of MCMC and variational method are generically different; b) the effects of model settings on these two methods are different, i.e., the settings of batch size 100 in MCMC does not necessary equivalent to batch size 100 in variational method. The paper did not include any information about the convergence of these two approaches. How many mini-batch for each model? Are these two models fully converged on training dataset? How many passes over the entire dataset? A more apple-to-apple comparison is between the best setting of both approaches using cross-validation to choose batch size and decay factor independently. Even in this case, authors did not include any time comparison between these two approaches. 3) As Canini et al. (2009) pointed out, the performance of their particle filters and sequential Monte Carlo approaches heavily relies on the initialization, which should also be the case for the proposed approach. However, the paper misses the important technical details on how to initialize it. If completely random, It is reasonable to believe the proposed model requires a large

	number of local sampling (variable N in Algorithm 2) to reduce the noise. This leads to previous questions on running time since it would definitely increase the running time. Minor comments for the reference, upper case {MCMC}, {D}irichlet, {B}ayesian, etc.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall Rating	Weak reject
Reviewer confidence	Reviewer is an expert
Detailed comments for the authors	This paper presents the streaming collapsed gibbs sampling method to learn topics in an online manner, and further develops the distributed algorithm in parallel environment. By comparing against the streaming variational Bayes inference, this paper shows the advantages of the proposed model and analyzes the extracted topics and topic evolution. In general, this paper is well organized. However, I have several concerns: - While the authors connected LDA to conditional density filtering and show CDF-LDA is guaranteed to converge, it is not the direct convergence proof for the proposed SGS. - The distributed SGS is not well introduced. It's unclear that how the batches are distributed according to time and how the different local nodes are synchronized with the global node. - Missing comparison between CGS and DCGS, while this paper declares "DSGS only suffer minor performance loss compared to the single thread version (SGS)" in the title of Figure 4(a). - Missing citation: the factorization of sampling equation in Section 2.1 has been introduced in "Limin Yao, David Mimno and Andrew McCallum. Efficient Methods for Topic Model Inference on Streaming Document Collections. KDD, 2009".