Speeding up Gibbs sampling by variable grouping

Evgeniy Bart
Palo Alto Research Center
Palo Alto, CA 94304
bart@parc.com

Abstract

Gibbs sampling is a widely applicable inference technique that can in principle deal with complex multimodal distributions. Unfortunately, it fails in many practical applications due to slow convergence and abundance of local minima. In this paper, we propose a general method of speeding up Gibbs sampling in probabilistic models. The method works by introducing auxiliary variables which represent assignments of the original model variables to groups. Our experiments indicate that the groups converge early in the sampling. After they have converged, the original variables no longer need to be sampled, and it becomes possible to re-sample an entire group at a time, greatly speeding up the sampler. The proposed ideas are illustrated on LDA and are applicable to many other topic models.

1 Introduction

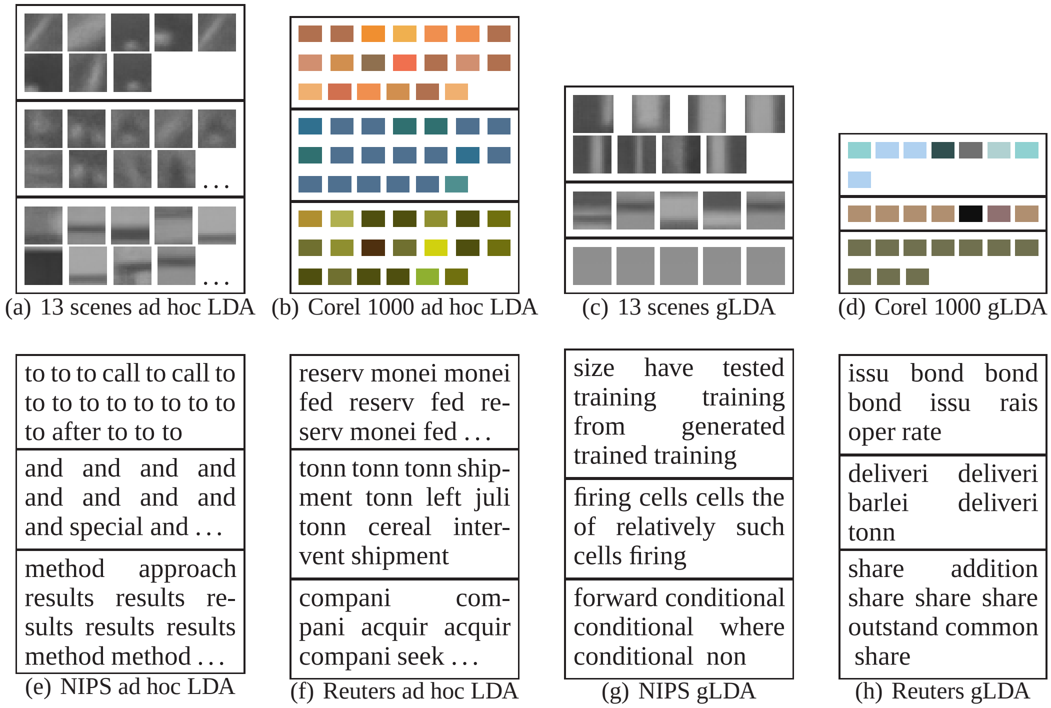
A fascinating aspect of human problem solving is how humans are capable of discovering the correct primitives at a level of abstraction suitable for the given problem. In contrast, most machine learning algorithms only work with user-supplied primitives. For example, when Gibbs sampling is used for inference in a probabilistic model, these primitives are the variables of the model. Typically, they represent the most basic building blocks of the problem, such as pixels in an image or words in a text document. As the problem size increases, more variables in the model become necessary. Gibbs sampling often fails under these circumstances due to slow convergence and abundance of local minima. The reason is that standard primitives become too fine-grained for large problems.

In this paper, we propose to combine the original model variables into groups (as in Figure 1), such that all variables within a group are likely to have the same assigned value. We show that establishing such groups is feasible very early in the sampling. After the groups are known, it becomes possible to re-sample an entire group at a time, thus speeding up sampling. Due to space constraints, only a brief description of variable grouping is given here; more details can be found in [2].

1.1 Brief survey of previous work

Variational inference is a popular method generally considered to be faster than sampling. However, a significant drawback of variational inference is that it finds only a single solution. In many cases, drawing multiple samples from the posterior distribution is desirable, for example, to find multiple modes or to deal with local maxima in the posterior.

In sampling, one class of relevant approaches includes augmentation samplers and similar methods [12, 8, 1]. Typically, several variables are sampled as a group to improve convergence. A disadvantage is that finding a good augmentation for a new problem may be difficult. The approach proposed here is more general and is readily applicable to a broader class of models. In addition, traditional augmentation samplers generally do not find persistent groups; instead, the groups change at every iteration. In our approach, the groups are persistent across multiple iterations and therefore are valuable by themselves. Similar comments pertain to split-merge methods (e. g. [7]).



(a) 13 scenes ad hoc LDA (b) Corel 1000 ad hoc LDA (c) 13 scenes gLDA (d) Corel 1000 gLDA

to to to call to call to to to to to to to to to to after to to to

and and and and and and and and and special and . . .

method approach results results results results results method method . . .

(e) NIPS ad hoc LDA

reserv monei monei fed reserv fed reserv monei fed . . .

tonn tonn tonn shipment tonn left juli tonn cereal intervent shipment

compani compani acquir acquir compani seek . . .

(f) Reuters ad hoc LDA

size have tested training training from generated trained training

firing cells cells the of relatively such cells firing

forward conditional conditional where conditional non

(g) NIPS gLDA

issu bond bond bond issu rais oper rate

deliveri deliveri barlei deliveri tonn

share addition share share share outstand common share

(h) Reuters gLDA

Figure 1: Example groups learned by ad hoc LDA and gLDA on image and text data. Three groups per experiment are shown. For each group, the tokens it contains are displayed within a frame. To save space, only a randomly selected subset of tokens is shown for some groups (marked by '. . . '). For the 13 scenes dataset, the average image patch is shown for each vector-quantized SIFT descriptor. Note that no word similarity measures were used; the models learned to assign similar words to the same group based solely on sampler behavior.

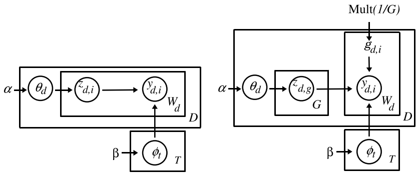
In [10], individual image pixels are combined into groups, called 'superpixels', and subsequent processing is done in terms of these superpixels. Superpixels are defined by color similarity and proximity. This and similar approaches that use persistent groups [6] generally define the groups using properties of the objects the original variables represent (for example, color similarity of image pixels). Creating such groups could therefore be difficult if the objects do not have an obvious similarity measure (e. g., in a recommendation system). In addition, grouping by similarity may not allow the method to deal with polysemy. The method proposed here is more general, as it defines similarity directly by the behavior of the original model. It can also deal with polysemy successfully.

2 Variable grouping

Latent Dirichlet Allocation (LDA) is a popular model for text and image data [3, 5] (Figure 2(a)). LDA represents documents as bags of words. The basic observed units in this representation are called *tokens*. Each token is an instance of some word in the vocabulary. Distinctive patterns of word co-occurence are represented by 'topics', which are multinomial distributions over the vocabulary.

In LDA, there is one variable for every token in every document. For large datasets, the total number of variables may reach millions. The computational cost of Gibbs sampling thus may become prohibitive. If several tokens could be identified as belonging to the same topic, they could be resampled as a group, saving computation time.

We have developed an ad hoc implementation (called ad hoc LDA) which is very general and doesn't require modifying the model. Due to space constraints, only the results are shown in section 3. A more principled approach to variable grouping is described below. The model is called gLDA, for 'group LDA' (Figure 2(b)). In gLDA, a set of G token groups for every document is introduced. Each token in a document is assigned to a group, and each group is assigned to a topic.



(a) LDA (b) gLDA

Figure 2: (a): the LDA model [3]. (b): the gLDA model. The main difference from LDA is that the topic for each token is determined not directly, but rather via a group variable $g_{d,i}$. The conditional distributions are: $\theta_d \sim \mathrm{Dir}^T[\alpha]$, $\phi_t \sim \mathrm{Dir}^Y[\beta]$, $g_{d,i} \sim \mathrm{Mult}(1/G)$, $z_{d,g} \sim \mathrm{Mult}(\theta_d)$, $y_{d,i} \sim \mathrm{Mult}(\phi_{z_{d,g_{d,i}}})$.

Dataset	Method	Initial perplexity	Final perplexity	Computational cost
Synthetic	regular LDA	4.26	4.00	100%
	Ad hoc LDA	4.26	4.03	24%
	gLDA	4.21	4.01	12%
NIPS papers	regular LDA	5.52	5.20	100%
	Ad hoc LDA	5.52	5.26	19%
	gLDA	5.47	5.23	19%
Reuters	regular LDA	6.24	5.17	100%
	Ad hoc LDA	6.24	5.24	16%
	gLDA	5.86	5.29	35%
Corel 1000	regular LDA	4.17	2.97	100%
	Ad hoc LDA	4.17	3.00	10%
	gLDA	3.47	2.97	23%
13 scenes	Gibbs sampling	6.56	5.99	100%
	Ad hoc LDA	6.56	6.09	29%
	gLDA	6.40	6.03	23%
102 flowers	regular LDA	4.99	3.37	100%
	Ad hoc LDA	4.97	3.36	18%
	gLDA	4.99	3.32	10%

Table 1: Comparison of perplexity and computational cost. Lower perplexity values are better. Computational cost is shown as percentage relative to the full computational cost of a regular Gibbs sampler. Lower values indicate faster computation.

Note that the meaning of θ_d^{gLDA} in gLDA is slightly different from that in LDA. The gLDA model is nevertheless useful, because in many cases the parameters of interest are topics ϕ_k, whose meaning is not changed. If needed (e. g. for document classification), the original θ_d^{LDA} can easily be recovered as well. Gibbs updates in gLDA can be derived in a straightforward manner.

Experiments in section 3 show that the groups converge early in the sampling. After they have converged, it becomes possible to resample an entire group at a time. This corresponds to proper Gibbs sampling in a modified model where the g variables are fixed. Another possibility is to never stop sampling the g variables, but instead sample them less frequently than the z variables (say, only every fifth iteration). This is possible since Gibbs sampling allows an arbitrary schedule to be used, as long as all variables are sampled infinitely often. This latter possibility works slightly better and was used in all experiments reported below.

3 Results

The proposed variable grouping method relies on the assumption that the groups converge early in the sampling. To verify this assumption experimentally, we generated a synthetic dataset, where the ground truth topic assignment for each token is known. (This synthetic dataset was generated as in [4] and included phenomena like polysemy.) We then ran Gibbs sampling in gLDA and measured group convergence (using the known ground truth labels). The groups converged after about 200 iterations. For comparison, 2000 iterations were needed for convergence in [11]. Therefore, for 1800 iterations (90%), only the groups needed to be resampled, rather than all tokens.

Next, we evaluate the quality of models learned using group variables. We use hold-out set perplexity as performance measure. Perplexity immediately after initialization is also reported to provide scale. All methods were evaluated on the synthetic dataset described above, two text datasets (NIPS papers and Reuters articles from the UCI repository), and three image datasets (13 visual scene categories [5], 1000 color images from the Corel dataset, and images of 102 flower categories [9]). For variable grouping methods, the number of groups was chosen to give about 10 tokens per group.

The performance is reported in Table 1. As can be seen, the proposed methods achieve performance similar to standard LDA, but at a fraction of computational cost. Several groups learned by ad hoc LDA and gLDA are shown in Figure 1. As can be seen, groups consist of tokens that often belong to the same topics. Most groups contain multiple words, although often the same word is repeated multiple times. A simple word similarity measure (such as edit distance) would not be able to achieve such grouping.

A naive way to speed up sampling is to simply reduce the number of iterations. For example, a standard Gibbs sampler could be run for 10 times less time on the 102 flowers dataset, achieving the same time efficiency as gLDA. However, such a small number of iterations is insufficient for convergence. As a result, the perplexity increases (by 20–30% in our experiments, depending on the dataset). For some datasets, object categorization experiments were done, and the increase in perplexity was accompanied by the corresponding decrease in categorization performance. Note also that although the increase in perplexity of 20% may seem small, it requires many (often, hundreds or thousands) iterations for standard Gibbs sampler to compensate for it.

An additional application of variable grouping to a different topic model similar to the Nested Chinese Restaurant Process (NCRP) [4] was also implemented. The results were similar in that variable grouping allowed to obtain the same or better perplexity at a fraction of the computational cost of regular Gibbs sampling. The details are omitted due to space constraints.

Acknowledgments

The author would like to thank Kristin Branson and Marco Andreetto for their useful suggestions.

References

[1] A. Barbu and S.-C. Zhu. Graph partition by Swendsen-Wang cuts. In *ICCV*, 2003.

[2] E. Bart. Speeding up Gibbs sampling by variable grouping. Technical report, Palo Alto Research Center, Dec. 2009. http://www.vision.caltech.edu/~bart/Publications/2009/BartGroupingTR.pdf.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*, 2004.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[6] R. Gomes, M. Welling, and P. Perona. Memory bounded inference in topic models. In *ICML*, 2008.

[7] 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.

[8] P. Liang, M. Jordan, and B. Taskar. A permutation-augmented sampler for DP mixture models. In *ICML*, 2007.

[9] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[10] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

[11] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2005.

[12] R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.*, 58(2):86–88, Jan 1987.