ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1075

Momentum Online LDA for Large-scale Datasets

Jihong Ouyang and **You Lu** and **Ximing Li**[1]

Abstract. Modeling large-scale document collections is a significant direction in machine learning research. Online LDA uses stochastic gradient optimization technology to speed the convergence; however the large noise of stochastic gradients leads to slower convergence and worse performance. In this paper, we employ the momentum term to smooth out the noise of stochastic gradients, and propose an extension of Online LDA, namely Momentum Online LDA (MOLDA). We collect a large-scale corpus consisting of 2M documents to evaluate our model. Experimental results indicate that MOLDA achieves faster convergence and better performance than the state-of-the-art.

1 Introduction

Recently, Latent Dirichlet Allocation (LDA) [1] has been paid more and more attentions for modeling text document collections. Variational Inference (VI) [2] is commonly used to approximately compute the posterior of the LDA model. However, it is inefficient to infer large-scale datasets that contains hundreds of thousands of documents.

To address this problem, Hoffman [3] investigated on Stochastic Gradient (SG) optimization technology to accelerate the inference procedure, and proposed an online approach, i.e., Stochastic Variational Inference (SVI). At each iteration, SVI form the noisy natural gradient [4] only depending on few random subsamples in the large-scale dataset, rather than the entire dataset as in the VI approach. In order to further improve SVI, Ranganath [5] studied on how to set the learning rate of SVI adaptively, and Wang [6] used control variate to reduce the variance of noisy natural gradient.

In terms of SVI [3, 7], the number of global variational parameters is proportional to the size of vocabulary. Since the vocabulary commonly contains thousands of words, the dimension of global variational parameters is significantly high. However, there are only a few unique words occurred each time, resulting in two problems. First, sometimes the frequent words appearing in the subsample are rare in the entire corpus; second, different subsamples may contain totally different words. Both problems generate a large noise, which causes the stochastic gradients greatly deviate from the true gradients and slows down the algorithm. In this paper, we propose an extension of Online LDA, namely Momentum Online LDA (MOLDA), to deal with the two problems mentioned above. Momentum is the sum of weighted previous gradients. Intuitively, it covers almost all the words in the vocabulary. Therefore, momentum can smooth out the noisy gradients and speed the convergence. Experimental results showed that MOLDA speeded the convergence as well as provided a better predictive distribution.

2 Models

In this section, we first review the Online LDA approach, and then introduce the proposed MOLDA approach.

2.1 Online LDA

LDA [1] is one of the most successful probabilistic topic models. It assumes that each document is represented by K latent topics, where each topic is a distribution over words. Given a collection of observed words $w = w_{1:D}$ (i.e., D is the number of the documents), our goal is to estimate the posterior distribution $p(\beta,\theta,z|w,\alpha,\eta)$, where $\beta \triangleq \beta_{1:K}$, $\theta \triangleq \theta_{1:D}$ and $z = z_{1:D}$ denote topic-word distributions, document-topic distributions and topic assignments, respectively; α and η are Dirichlet hyper-parameters. Variational Inference (VI) [2] is commonly used to compute this posterior.

VI first posits a factored variational distribution $q(\beta,\theta,z)$, and then minimizes the KL divergence between this variational distribution and the posterior. This equals to maximizing the evidence lower bound with respect to the three variational parameters γ , ϕ and λ . λ is the global parameter. At each iteration, it is optimized as follows:

$$\lambda_k = \eta + \sum\nolimits_{d=1}^{D} \sum\nolimits_{n=1}^{N_d} \phi_{dn}^k w_{dn} \tag{1}$$

As shown in Eq.1, updating the global parameter λ needs to analyze the entire corpus at each time. Intuitively, it is inefficient in terms of large-scale datasets. Online LDA [3, 7] uses stochastic optimization to solve this problem. It scales well because each time, it only needs to subsample a subset of the corpus, and makes use of Eq.2 to optimize λ .

$$\lambda_k^{(t)} = \lambda_k^{(t-1)} + \rho_t(-\lambda_k^{(t-1)} + \eta + \frac{D}{S}\sum\nolimits_{s=1}^{S}\sum\nolimits_{n=1}^{N_s}\phi_{dn}^k w_{dn}) \tag{2}$$

where $-\lambda_k^{(t-1)} + \eta + \frac{D}{S}\sum\nolimits_{s=1}^{S}\sum\nolimits_{n=1}^{N}\phi_{dn}^k w_{dn}$ is the stochastic natural gradient [4] with respect to $\lambda_k^{(t-1)}$, S is the size of a mini-batch , and ρ_t is the learning rate, where $0 < \rho_t \le 1$.

2.2 MOLDA

Due to the high dimension of text modeling, the stochastic natural gradient used in Online LDA suffers a significantly large noise.

[1] College of Computer Science and Technology, Jilin University, email: luyou_0027@foxmail.com,

This leads to slower convergence and worse performance, especially for the large-scale datasets. A natural scheme to solve this problem is to enlarge the size of mini-batch. Unfortunately, if the size of mini-batch is too large, the algorithm will lose the advantage in efficiency of stochastic gradient optimization.

SG method with momentum can smooth out the noise of stochastic gradient, so that obtains a faster convergence [8, 9, 10]; further, the momentum is efficient to be computed. Hence, we use the momentum term to extend Online LDA, and propose a Momentum Online LDA (MOLDA). The main contribution of MOLDA is that of improving the updating rule of the global parameter λ. In contrast to Eq. 2, MOLDA optimizes λ as follows:

$$\lambda_k^{(t)} = \lambda_k^{(t-1)} + \rho_t g(\lambda_k^{(t-1)}, w^{(t)}) + \sigma(\lambda_k^{(t-1)} - \lambda_k^{(t-2)}) \qquad (3)$$

where $\lambda_k^{(t-1)} - \lambda_k^{(t-2)}$ is the momentum term; σ is the constant momentum parameter: $0 \le \sigma < 1$; $g(\lambda_k^{(t-1)}, w^{(t)})$ is the stochastic natural gradient.

Expanding out the right hand side of Eq.3, we can obtain:

$$\lambda^{(t)} = \lambda^{(t-1)} + \sum\nolimits_{i=1}^{t} \rho_i \sigma^{t-i} g(\lambda^{(i-1)}, w^{(i)}) \qquad (4)$$

Note that the term $\sum\nolimits_{i=1}^{t} \rho_i \sigma^{t-i} g(\lambda^{(i-1)}, w^{(i)})$ is the sum of weighted stochastic natural gradients over iteration 1 to t, therefore, it contains almost all the words in the vocabulary. Intuitively, this performs a positive influence to the large noise in the high dimension space.

3 Experiment

In this section, we have conducted some experiments to evaluate the proposed approach.

Experiment settings. We randomly downloaded 2M documents from English version of Wikipedia, and removed all words not in a pruned vocabulary[2] of 7,700 words. We randomly selected 2000 documents from the Wikipedia collection as the held-out data. The popular metric, i.e., *per-word log likelihood* (hereafter referred to as *likelihood*) [3, 6, 7], is chosen to evaluate the text modeling performance. Note that the higher *likelihood,* the better the performance is.

Results. The state-of-the-art Online LDA[2] is chosen as baseline algorithm. Both MOLDA and Online LDA set the same parameters. Following the suggestions in [3], we set the number of topics $K = 100$, hyper-parameters $\alpha = 0.1$ and $\eta = 0.01$. We fixed the mini-batch size as 500, and set the learning rate at iteration t as: $\rho_t = (t + \tau)^{-\kappa}$, where the delay $\tau = 1024$, and the forgetting rate $\kappa = 1$.

Since the variational objective is a non-convex function and sensitive to the initial value of λ, we run Online LDA at the beginning of 20 iterations to initialize λ.

Figure 1 shows the results. Obviously, MOLDA converges faster and achieves a better predictive distribution than Online LDA. On one hand, MOLDA converges after about 900 iterations, while Online LDA needs about 1200 iteration. On the other hand, the final *likelihood* of MOLDA is higher than Online LDA about 0.02. We are thus concluding that our improvement promotes the performance of Online LDA

In addition, we also evaluate various value of parameter σ, such as 0.2, 0.25, 0.3, 0.35, and 0.4. All performs similar trends. Due to the space limit, we omit some of them.

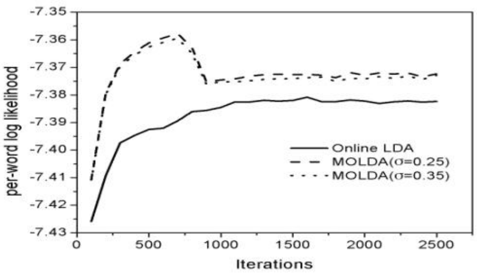

Figure 1. Experiments results of Online LDA and MOLDA

4 Conclusions

In order to efficiently modeling large-scale documents, we developed a MOLDA model. The proposed model uses a momentum term to smooth out the noise of stochastic natural gradient in Online LDA. We collect a Wikipedia collection consisting of 2M documents. The experiment results on this collection show that MOLDA achieves competitive performance with the state-of-the-art.

This short paper presents our preliminary results. In the future, we plan to perfect this work as: (1) conducting extensive experiments to further evaluate MOLDA; (2) studying on method which can tune the momentum parameter adaptively.

ACKNOWLEDGEMENTS

This work was supported by National Nature Science Foundation of China (NSFC) under the Grant No. 61170092, 61133011, and 61103091

REFERENCES

[1] D.M. Blei, A.Y. NG, and M.I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, **3**, 993-1022, (2003).

[2] C. Bishop, Pattern Recognition and Machine Learning. Springer New York, (2006).

[3] M.D. Hoffman, D.M Blei, C. Wang, and J. Paisley, Stochastic Variational Inference, Journal of Machine Learning Research, **14**(1), 1303-1347, (2013).

[4] S. Amari, Natural gradient works efficiently in learning, Neural computation, **10**(2), 251-276, (1998).

[5] R. Ranganath, C. Wang, D.M. Blei, and E.P. Xing, An adaptive learning rate for stochastic variational inference. In International Conference on Machine learning, 880-887, (2013).

[6] C. Wang, X. Chen, A. Smola, and E.P. Xing, Variance reduction for stochastic gradient optimization, In Neural Information Processing Systems, (2013).

[7] M.D. Hoffman, D. Blei, and F. Bach. Online inference for latent Dirichlet allocation, In Neural Information Processing Systems, (2010).

[8] T. Leen and G. Orr, Optimal Stochastic Search and Adaptive Momentum, In Neural Information Processing Systems, (1994).

[9] N. Qian, On the Momentum Term in Gradient Descent Learning Algorithms, *Neural* Networks, **12**,145-151, (1999).

[10] P. Tseng, An incremental gradient (-projection) method with momentum term and adaptive stepsize rule, SIAM Journal on Optimization, **8**(2), 506-531, 1998.

[2] http://www.cs.princeton.edu/~mdhoffma/.