A Fast Algorithm for Posterior Inference with Latent Dirichlet Allocation

Bui Thi-Thanh-Xuan[1,2(✉)], Vu Van-Tu[1], Atsuhiro Takasu[3], and Khoat Than[1]

[1] Hanoi University of Science and Technology, Hanoi, Vietnam
thanhxuan1581@gmail.com, vutu201130@gmail.com, khoattq@soict.hust.edu.vn
[2] University of Information and Communication Technology, Thai Nguyen, Vietnam
[3] National Institute of Informatics, Tokyo, Japan
takasu@nii.ac.jp

Abstract. Latent Dirichlet Allocation (LDA) [1], among various forms of topic models, is an important probabilistic generative model for analyzing large collections of text corpora. The problem of posterior inference for individual texts is very important in streaming environments, but is often intractable in the worst case. To avoid directly solving this problem which is NP-hard, some proposed existing methods for posterior inference are approximate but do not have any guarantee on neither quality nor convergence rate. Based on the idea of Online Frank-Wolfe algorithm by Hazan [2] and improvement of Online Maximum a Posteriori Estimation algorithm (OPE) by Than [3,4], we propose a new effective algorithm (so-called NewOPE) solving posterior inference in topic models by combining Bernoulli distribution, stochastic bounds, and approximation function. Our algorithm has more attractive properties than existing inference approaches, including theoretical guarantees on quality and fast convergence rate. It not only maintains the key advantages of OPE but often outperforms OPE and existing algorithms before. Our new algorithm has been employed to develop two effective methods for learning topic models from massive/streaming text collections. Experimental results show that our approach is more efficient and robust than the state-of-the-art methods.

Keywords: Topic models · OPE · Stochastic inference · NewOPE
MAP estimation

1 Introduction

Latent Dirichlet analysis is a flexible latent variable framework for modeling high-dimensional sparse count data. Latent Dirichlet Allocation (LDA) [1] has found successful applications in a wide range of areas including text modeling [5], bioinformatics [6], history [7,8], politics [5,9], psychology [10]. Estimation of posterior distributions for individual documents is one of the core issues in LDA.

N. T. Nguyen et al. (Eds.): ACIIDS 2018, LNAI 10752, pp. 137–146, 2018.
https://doi.org/10.1007/978-3-319-75420-8_13

Recently, this estimation problem is considered by many researchers, and many methods such as Variational Bayes (VB) [1], Collapsed Variational Bayes (CVB) [11], CVB0 [12], Collapsed Gibbs Sampling (CGS) [7,13], and OPE [3] have been proposed and applied. We also see that the quality of LDA in practice is determined by the quality of the inference method being employed. However, almost mentioned methods do not have a theoretical guarantee of quality or convergence rate, except OPE algorithm.

Our first contribution is that we propose a new algorithm which is called NewOPE, for doing posterior inference of topic mixture in LDA. The posterior inference problem is in fact non-convex and is NP-hard [14]. Than and Doan [4] proved that OPE converges at a rate of $\mathcal{O}(1/T)$, which surpasses the best rate of existing stochastic algorithms for non-convex problems [15,16], where T is the number of iterations. One main drawback of OPE is that there is no guarantee for OPE escaping from saddle points of the inference problems [17][1].

Similar to OPE, NewOPE is stochastic in nature and theoretically converges to a local maximal/stationary point of the inference problem. We approximate the objective function by a combination of the upper and lower bounds using Bernoulli distribution. The usage of both bounds is stochastic in nature and helps us reduce the possibility of getting stuck at a local stationary point. Thus, it is an efficient approach for escaping saddle points in non-convex optimization. So, our new variant is appropriate and effective more than the original OPE. Existing methods become less relevant in high dimensional non-convex optimization.

Using NewOPE as core routines to do inference, our second contribution is that we obtain two effective methods for learning LDA (so-called ML-NewOPE, Online-NewOPE) from massive/streaming text collections. From extensive experiments on two large corpora New York Time and Pubmed, we find that our methods can reach state-of-the-art performance in both predictiveness and model quality.

Organization: The rest of this paper is organized as follows. We introduce an overview of posterior inference with LDA in Sect. 2. In Sect. 3, a new algorithm NewOPE for posterior inference is proposed in detail and is applied to online learning LDA. In Sect. 4, we give some results test with large datasets. Finally, we conclude the paper in Sect. 5.

Notation: Throughout the paper, we use the following conventions and notations. The unit simplex in the n-dimensional Euclidean space is denoted as $\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \geq 0, \sum_{k=1}^{n} x_k = 1\}$, and its interior is denoted as $\overline{\Delta}_n$. We will work with text collections with V dimensions (dictionary size). Each document $\boldsymbol{d}$ will be represented as a frequency vector, $\boldsymbol{d} = (d_1, \ldots, d_V)^T$ where d_j represents the frequency of term j in $\mathbf{d}$. Denote n_d as the length of $\boldsymbol{d}$, i.e., $n_d = \sum_j d_j$.

[1] A saddle point is critical point but not always a (local) maximal point. Further, the inference might have exponentially large number of saddle points.

2 Related Work

LDA [1] assumes that a corpus is composed from K topics $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$. Each document $\boldsymbol{d}$ is a mixture of those topics and is assumed to arise from the following generative process. For the n^{th} word of $\boldsymbol{d}$:

- draw topic index $z_{dn}|\boldsymbol{\theta}_d \sim Multinomial(\boldsymbol{\theta}_d)$
- draw word $w_{dn}|z_{dn}, \boldsymbol{\beta} \sim Multinomial(\boldsymbol{\beta}_{z_{dn}})$.

We consider the MAP estimation of topic mixture for a given document d

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Delta_K} \Pr(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\beta}) Pr(\boldsymbol{\theta}|\alpha) \tag{1}$$

For a given document $\boldsymbol{d}$, the probability that a term j appears in $\boldsymbol{d}$ can be expressed as

$\Pr(w = j|d) = \sum_{k=1}^{K} \Pr(w = j|z = k).\Pr(z = k|\boldsymbol{d})$ or $\Pr(w = j|d) = \sum_{k=1}^{K} \beta_{kj}\theta_k$.

Hence the log likelihood of $\boldsymbol{d}$ is

$\log \Pr(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\beta}) = \log \prod_j \Pr(w = j|\boldsymbol{d})^{d_j} = \sum_j d_j \log \Pr(w = j|\boldsymbol{d})$

$= \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj}$

Remember that the density of the K-dimensional Dirichlet distribution with the parameter α is $P(\boldsymbol{\theta}|\alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha-1}$. Therefore, problem (1) is equivalent to the following:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Delta_K} \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^{K} \log \theta_k \tag{2}$$

where $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^{K} \log \theta_k$ is the objective function of (2).

In the case of $\alpha \geq 1$, one can easily show that the problem (2) is concave, therefore it can be solved in polynomial time. Unfortunately, in practice of LDA, the parameter α is often small, says $\alpha < 1$, causing (2) to be non-concave. That is the reason for why (2) is intractable in the worst case. Sontag and Roy [14] showed that problem (2) is NP-hard in the worst case when $\alpha < 1$. Many "batch" posterior inference algorithms have been proposed, including variational Bayes (VB), collapsed variational Bayesian inference (CVB), CVB0, and collapsed Gibbs sampling (CGS). VB, CVB, and CVB0 try to estimate the distribution by maximizing a lower bound of the likelihood $p(d|\beta, \alpha)$, whereas CGS [7,13] tries to estimate $p(z|d, \beta, \alpha)$. However, those "batch" algorithms are not practical for large scale data analysis because they often require many sweeps through all documents in the corpus.

Based on Online Frank-Wolfe algorithm which is proposed by Hazan [2] solving effective convex optimization using stochastic approximation, Than and Doan [3,4] proposed Online Maximum a Posteriori Estimation (OPE) algorithm for doing inference of topic mixtures for documents. Details of OPE are presented in Algorithm 1.

Algorithm 1. OPE: Online maximum a posteriori estimation

Input: document **d** and model $\{\boldsymbol{\beta}, \alpha\}$
Output: $\boldsymbol{\theta}^*$ that maximizes $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$
  Initialize $\boldsymbol{\theta_1}$ arbitrary in Δ_K
  for $t = 1, 2, ...\infty$ **do**
    Pick f_t uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$
    $F_t := \frac{2}{t} \sum_{h=1}^t f_h$
    $\boldsymbol{e}_t := \arg\max_{\boldsymbol{x} \in \Delta_K} < F_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x} >$
    $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}$
  end for

The main idea of OPE is to construct a stochastic sequence $F_t(\theta)$ that approximates for $f(\theta)$ by using uniform distribution, so that the (2) becomes easy to solve.

3 Proposed Method

We continue to consider the MAP problem (2) with the objective function

$$f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$$

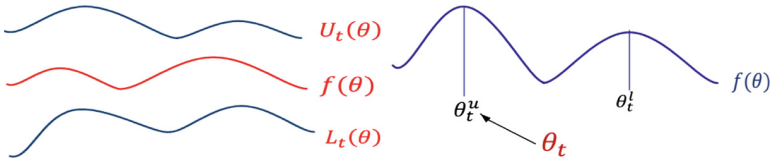
We find out that OPE is a good algorithm for posterior inference, then we proceed to improve OPE by randomization to get better algorithm. To avoid solving directly non-convex optimization problem (2), we need to construct an approximation function which is easy to maximize and that approximates well for $f(\theta)$. We use Bernoulli distribution to construct the approximation of the objective function $f(\theta)$. Pick f_h has Bernoulli distribution with probability p from $\{g_1(\boldsymbol{\theta}); \ g_2(\boldsymbol{\theta})\}$ where $\Pr(f_h = g_1(\theta)) = p, \quad \Pr(f_h = g_2(\theta)) = 1 - p$ and approximation $F_t(\theta)$ as a form $F_t(\theta) = \frac{2}{t} \sum_{h=1}^t f_h$. We see that $F_t(\theta)$ is a stochastic approximation which is easy to maximize and differentiable. We also see that

$$g_1(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} < 0; \quad g_2(\boldsymbol{\theta}) = (\alpha - 1) \sum_{k=1}^K \log \theta_k > 0$$

Hence, if we choose $f_1 = g_1$ then $F_1(\theta) < f(\theta)$, which leads $F_t(\boldsymbol{\theta})$ is a lower bound for $f(\boldsymbol{\theta})$. In contrast, if we choose $f_1 = g_2$ then $F_1(\theta) > f(\theta)$, and $F_t(\boldsymbol{\theta})$ is a upper bound for $f(\boldsymbol{\theta})$.

Although, OPE is a good candidate for solving a posterior inference in topic models, but we want to enhance OPE in different ways. It makes sense that two stochastic approximating sequences from above and below are better than one. So we construct two sequences that both approximating to $f(\theta)$, one begins with g_1 called the sequence $\{L_t\}$, another begins with g_2 called the sequence

$\{U_t\}$: Setting $f_1^l := g_1(\boldsymbol{\theta})$, $f_1^u := g_2(\boldsymbol{\theta})$. Pick f_t^l, f_t^u as Bernoulli distribution from $\{g_1(\boldsymbol{\theta});\ g_2(\boldsymbol{\theta})\}$ with probability p, we have $L_t := \frac{2}{t}\sum_{h=1}^{t} f_h^l$ and $U_t := \frac{2}{t}\sum_{h=1}^{t} f_h^u$.



(a) Upper bound and lower bound of objective function

(b) Choose the best point

Fig. 1. The ideas of NewOPE

The idea of NewOPE is based on the greedy approach. At each iteration we always compare two values of $\{f(\boldsymbol{\theta}_t^u), f(\boldsymbol{\theta}_t^l)\}$ and take the point that makes the value of f highest as possible. NewOPE works differently from OPE. OPE just constructs one sequence $\{\boldsymbol{\theta}_t\}$ while NewOPE creates three sequences $\{\boldsymbol{\theta}_t^u\}$, $\{\boldsymbol{\theta}_t^l\}$ and $\{\boldsymbol{\theta}_t\}$ depending on each other. The structure of sequence $\{\boldsymbol{\theta}_t\}$ really changes, but OPE's good properties are remained in NewOPE (Fig. 1). Comparing with other inference approaches (including VB, CVB, CVB0 and CGS), NewOPE has many preferable properties as summarized

Algorithm 2. NewOPE

Input: document $\mathbf{d}$ and model $\{\boldsymbol{\beta}, \alpha\}$
Output: $\boldsymbol{\theta}$ that maximizes $f(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta})$

Initialize $\boldsymbol{\theta_1}$ arbitrary in Δ_K
$f_1^l := g_1(\boldsymbol{\theta})$, $f_1^u := g_2(\boldsymbol{\theta})$
for $t = 2, 3, ...\infty$ **do**
  Pick f_t^l as Bernoulli distribution from $\{g_1(\boldsymbol{\theta})\ ;\ g_2(\boldsymbol{\theta})\}$ with probability p
  $L_t := \frac{2}{t}\sum_{h=1}^{t} f_h^l$
  $\boldsymbol{e}_t^l := \arg\max_{\boldsymbol{x} \in \Delta_K} < L_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x} >$
  $\boldsymbol{\theta}_{t+1}^l := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t^l - \boldsymbol{\theta}_t}{t}$

  Pick f_t^u as Bernoulli distribution from $\{g_1(\boldsymbol{\theta})\ ;\ g_2(\boldsymbol{\theta})\}$ with probability p
  $U_t := \frac{2}{t}\sum_{h=1}^{t} f_h^u$
  $\boldsymbol{e}_t^u := \arg\max_{\boldsymbol{x} \in \Delta_K} < U_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x} >$
  $\boldsymbol{\theta}_{t+1}^u := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}$

  $\boldsymbol{\theta}_{t+1} := \arg\max_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{t+1}^u, \boldsymbol{\theta}_{t+1}^l\}} f(\boldsymbol{\theta})$
end for

- NewOPE explicitly has a theoretical guarantee on fast convergence rate. This is the most notable property of NewOPE, for which existing inference methods often do not have, except OPE.
- Unlike CVB and CVB0 [12], NewOPE does not change the global variables when doing inference for individual documents. So, NewOPE is more beneficial than CVB and CVB0.

Our algorithm not only retains the good characteristics of OPE but also makes it better and more effective when solving the problem of inference with LDA. We focus on overcoming the disadvantages of ineffective batch methods, exploiting stochastic approximation, and putting probability distributions into our new algorithm.

We have seen many attractive properties of NewOPE that other methods do not have. We further show in this section the simplicity of using NewOPE for designing fast learning algorithms for topic models. More specifically, we obtain two algorithms: Online-NewOPE which learns LDA from large corpora in an online fashion, and ML-NewOPE which enables us to learn LDA from either large corpora or data streams based on ML-OPE and Online-OPE in [4].

4 Empirical Evaluation

This section is devoted to investigating practical behaviors of NewOPE, and how useful it is when NewOPE is employed to design new algorithms for learning topic models at large scales. To this end, we take the following methods, datasets, and performance measures into investigation.

Inference methods: Variational Bayes (VB) [1], Collapsed variational Bayes (CVB0) [12], Collapsed Gibbs sampling (CGS) [7], Online MAP estimation (OPE) [4], New Online MAP estimation (NewOPE).

Large-scale learning methods: ML-NewOPE, Online-NewOPE, ML-OPE, Online-OPE [4], Online-CGS [7], Online-CVB0 [12], Online-VB [18].

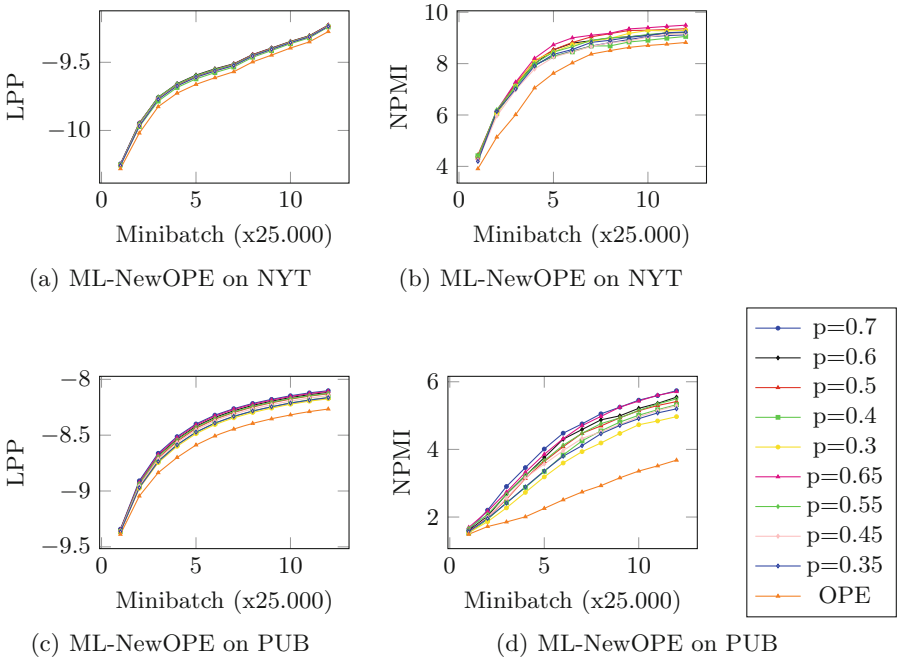
To avoid randomness, the learning methods for each dataset is run five times and reported its average results.

- Model parameters: The number of topics $K = 100$, the hyper-parameters $\alpha = \frac{1}{K}$ and $\eta = \frac{1}{K}$. These parameters are commonly used in topic models.
- Inference parameters: The number of iterations was chosen as $T = 50$.
- Learning parameters: $\kappa = 0.9$, $\tau = 1$ adapted best for existing inference methods.

Datasets: We used the two large corpora: Dataset PubMed consists of 330,000 articles from the PubMed central and dataset New York Times (NYT) consists of 300,000 news[2].

[2] The data sets were taken fromhttp://archive.ics.uci.edu/ml/datasets.

Measures: We used two measures *Log Predictive Probability* (LPP) [7] and *NPMI* [19]. Predictive probability measures the predictiveness and generalization of a model to new data, while NPMI evaluates semantics quality of an individual topic.

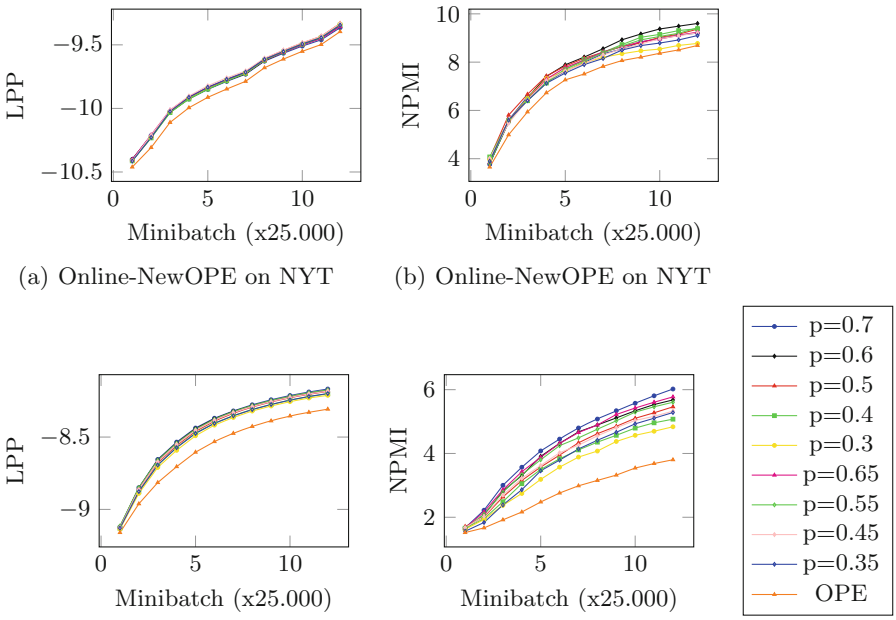


(a) ML-NewOPE on NYT (b) ML-NewOPE on NYT

(c) ML-NewOPE on PUB (d) ML-NewOPE on PUB

Fig. 2. ML-NewOPE with different parameter p

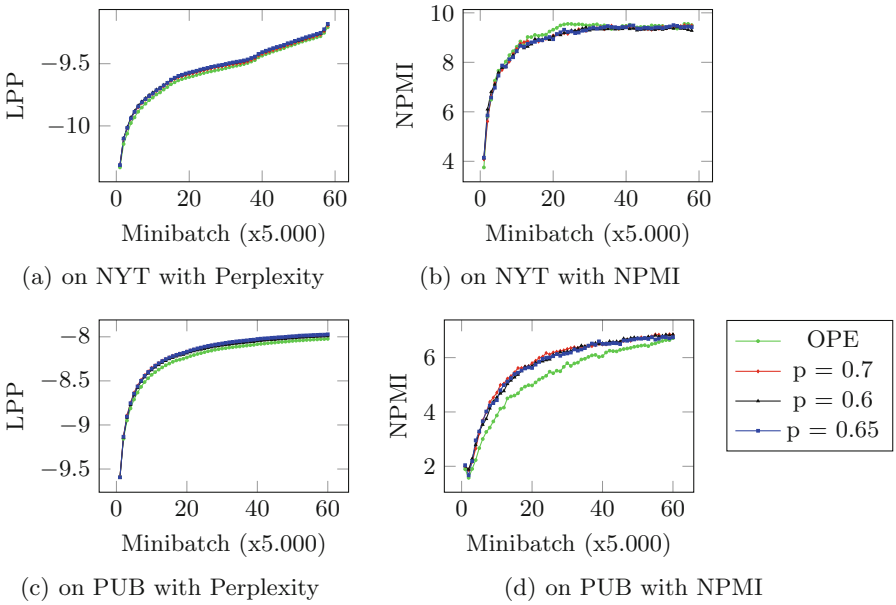
In Figs. 2 and 3, we find that the NewOPE algorithm which uses Bernoulli distribution instead of uniform better than OPE, especially when probability p is greater than 0.5. From Figs. 4 and 5, NewOPE is better than VB, CVB and CGS on two datasets and with two measures Predictive Probability and NPMI, especially when the probability p is greater than 0.5 such as 0.6, 0.65 or 0.7 in our experiments. This explains the contribution of prior/likelihood to solving the inference problem. In terms of semantic quality measured by NPMI, Fig. 5 shows the results from our experiments. It is easy to observe that Online-NewOPE often learns models with a good semantic and NewOPE makes ML-NewOPE and Online-NewOPE work more efficient. NewOPE demonstrates our idea of using two stochastic sequences $\{U_t(\theta), L_t(\theta)\}$ to approximate an objective function $f(\boldsymbol{\theta})$. The idea of increasing the randomness and greedy of the algorithm is exploited here. Firstly, two stochastic sequences of function $U_t(\theta), L_t(\theta)$ are used to raise our participants and information relevant to objective function $f(\theta)$. Hence, at the next iteration, we have more choices in $\boldsymbol{\theta}_t$. Secondly, choosing $\boldsymbol{\theta}_t$



(a) Online-NewOPE on NYT (b) Online-NewOPE on NYT

(c) Online-NewOPE on PUB (d) Online-NewOPE on PUB

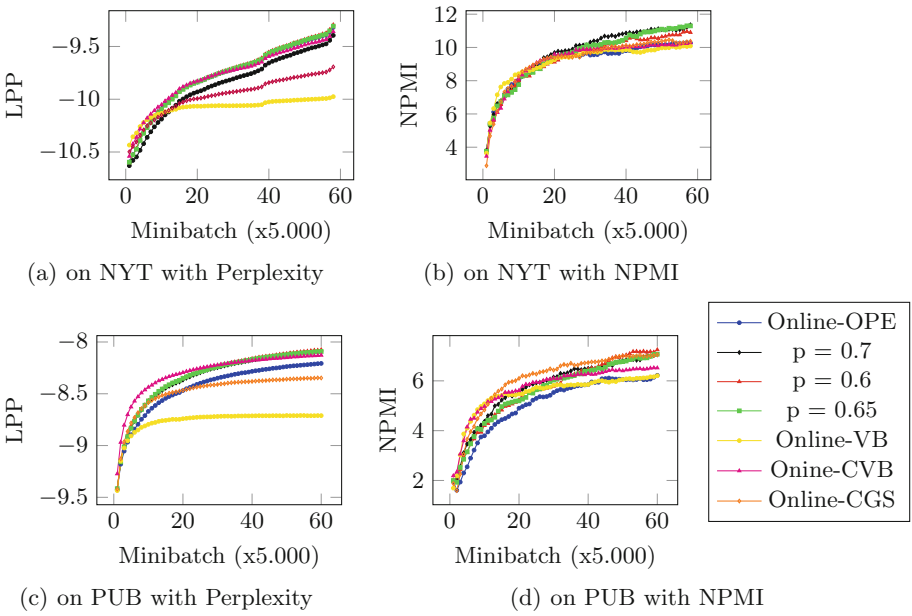
Fig. 3. Online-NewOPE with different parameter p



(a) on NYT with Perplexity (b) on NYT with NPMI

(c) on PUB with Perplexity (d) on PUB with NPMI

Fig. 4. ML-NewOPE compares with ML-OPE



(a) on NYT with Perplexity (b) on NYT with NPMI

(c) on PUB with Perplexity (d) on PUB with NPMI

Fig. 5. Online-NewOPE compares with Online-OPE, Online-VB, Online-CVB0 and Online-CGS

from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$ that makes the value of $f(\boldsymbol{\theta})$ higher in each iteration comes from idea of greedy algorithms. There are many ways of choices here, but we design a best way to create $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$. This approach is simple and there is no need for extra parameter.

5 Conclusion

We have discussed how posterior inference for individual texts in topic models can be done efficiently. Using Bernoulli distribution and stochastic approximation, we now provide NewOPE algorithm to deal well with this problem. By exploiting NewOPE carefully, we have arrived at two efficient methods for learning LDA from data streams or large corpora. As a result, they are good candidates to help us deal with text streams and big data.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
2. Hazan, E., Kale, S.: Projection-free online learning. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 26 June – 1 July 2012, Edinburgh, Scotland, UK (2012)

3. Than, K., Doan, T.: Dual online inference for latent Dirichlet allocation. In: ACML (2014)
4. Than, K., Doan, T.: Guaranteed inference in topic models. arXiv preprint arXiv:1512.03308 (2015)
5. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
6. Falush, D., Stephens, M., Pritchard, J.K.: Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**(4), 1567 (2003)
7. Hoffman, M., Blei, D.M., Mimno, D.M.: Sparse stochastic inference for latent Dirichlet allocation. In: Proceedings of the 29th International Conference on Machine Learning (ICML 2012), pp. 1599–1606. ACM (2012)
8. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 937–946. ACM (2009)
9. Grimmer, J.: A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Polit. Anal. **18**(1), 1–35 (2010)
10. Schwartz, H.A., Eichstaedt, J.C., Dziurzynski, L., Kern, M.L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M.E., Ungar, L.H.: Toward personality insights from language exploration in social media. In: AAAI Spring Symposium: Analyzing Microtext (2013)
11. Teh, Y.W., Kurihara, K., Welling, M.: Collapsed variational inference for HDP. In: Advances in Neural Information Processing Systems, pp. 1481–1488 (2007)
12. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 27–34. AUAI Press (2009)
13. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
14. Sontag, D., Roy, D.: Complexity of inference in latent Dirichlet allocation. In: Neural Information Processing System (NIPS) (2011)
15. Dang, C.D., Lan, G.: Stochastic block mirror descent methods for nonsmooth and stochastic optimization. SIAM J. Optim. **25**(2), 856–881 (2015)
16. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. **155**, 267–305 (2016)
17. Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in Neural Information Processing Systems, pp. 2933–2941 (2014)
18. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.W.: Stochastic variational inference. J. Mach. Learn. Res. **14**(1), 1303–1347 (2013)
19. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: EACL, pp. 530–539 (2014)