

A stochastic algorithm for solving the posterior inference problem in topic models

Hoang Quang Trung^{1,2}, Xuan Bui³

¹Faculty of Electrical and Electronic Engineering, Phenikaa University, Yen Nghia, Ha Dong, Hanoi 12116, Vietnam

²Phenikaa Research and Technology Institute (PRATI), A&A Green Phoenix Group JSC,
No.167 Hoang Ngan, Trung Hoa, Cau Giay, Hanoi 11313, Vietnam

³Faculty of Computer Science and Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam

Article Info

Article history:

Received Jul 02, 2021

Revised Jun 04, 2022

Accepted Jun 19, 2022

Keywords:

Latent Dirichlet allocation

Posterior inference

Stochastic optimization

Topic models

ABSTRACT

Latent Dirichlet allocation (LDA) is an important probabilistic generative model and has usually used in many domains such as text mining, retrieving information, or natural language processing domains. The posterior inference is the important problem in deciding the quality of the LDA model, but it is usually non-deterministic polynomial (NP)-hard and often intractable, especially in the worst case. For individual texts, some proposed methods such as variational Bayesian (VB), collapsed variational Bayesian (CVB), collapsed Gibb's sampling (CGS), and online maximum a posteriori estimation (OPE) to avoid solving this problem directly, but they usually do not have any guarantee of convergence rate or quality of learned models excepting variants of OPE. Based on OPE and using the Bernoulli distribution combined, we design an algorithm namely general online maximum a posteriori estimation using two stochastic bounds (GOPE2) for solving the posterior inference problem in LDA model. It also is the NP-hard non-convex optimization problem. Via proof of theory and experimental results on the large datasets, we realize that GOPE2 is performed to develop the efficient method for learning topic models from big text collections especially massive/streaming texts, and more efficient than previous methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Hoang Quang Trung

Faculty of Electrical and Electronic Engineering, Phenikaa University

Yen Nghia, Ha Dong, Hanoi 12116, Vietnam

Email: trung.hoangquang@phenikaa-uni.edu.vn

1. INTRODUCTION

In data mining, one of the most general and powerful techniques is the topic modeling [1]-[3]. In recently, there are much published research in the field of topic modeling and applied in various fields such as medical and linguistic science. Latent Dirichlet allocation (LDA) [4] is the popular methods for topic modeling [5]-[7], LDA has found successful applications in text modeling [8], bioinformatic [9], [10], biology [11], history [12], [13], politics [14]-[16], and psychology [17], to name a few. Recently, there are much research related to corona virus disease 2019 (COVID-19) pandemic that also use LDA model in data analysis. These show the important role and advantage of the topic models in text mining [18]-[20]. We find out that the quality of the LDA model is highly dependent on the inference methods [4]. In recent years, many posterior inference methods have obtained more attention from scientists such as variational Bayesian (VB) [4], collapsed variational Bayesian (CVB) [21], collapsed Gibb's sampling (CGS) [12], [22], and online maximum a posteriori estimation (OPE) [23]. Those methods enable us to easily work with big data [12], [24]. Except variants of OPE, most of the methods do not have a guarantee of convergence rate or model quality in theory.

We realize that in topic models, the posterior inference problem is in fact non-convex optimization problem. It also belongs to class of non-deterministic polynomial (NP)-hard problem [25]. We also find out that OPE has the convergence rate is $O(1/T)$ where T is the number of iterations. OPE overcomes the best rate of existing stochastic algorithms for solving the non-convex problems [23], [26], [27]. To the best of my knowledge, solving the posterior inference problem usually leads to a non-convex optimization problem. The big question is how efficiently an optimization algorithm can try to escape saddle points? we carefully consider the optimization algorithms applied to the posterior inference problem. It is the basis for us to propose the general online maximum a posteriori estimation using two stochastic bounds (GOPE2) algorithm. In this paper, we propose the GOPE2 algorithm based on a stochastic optimization approach for solving the posterior inference problem. Using the Bernoulli distribution and two stochastic bounds of the true non-convex objective function, we have shown that GOPE2 achieves even better than previous algorithms. It also keeps the good properties of OPE and continues to do better than OPE. Stochastic bounds replacing true objective function reduces the possibility of getting stuck at a local stationary point or escaping saddle points. This is an effective approach to get rid of saddle points while existing methods are unsuitable especially in high-dimensional non-convex optimization. We use GOPE2 as the core algorithm for doing inference, we obtain online-GOPE2 which is an efficient method for learning LDA from large text collections, especially short-text documents. Based on our experiments on large datasets, we show that our method can reach state-of-the-art performance in both qualities of learned model and predictiveness.

2. RELATED WORK

LDA [4] is a generative model for discrete data and modeling text. In LDA, a corpus is composed from K topics $\beta = (\beta_1, \dots, \beta_K)$, each of which is a sample from Dirichlet (η) which is V -dimensional Dirichlet distribution. LDA model assumes that each document d is a mixture of topics and arises from the following generative process:

- a) Draw $\theta_d \mid \alpha \sim \text{Dirichlet}(\alpha)$
- b) For the n^{th} word of d :
 - Draw topic index $z_{dn} \mid \theta_d \sim \text{multinomial}(\theta_d)$
 - Draw word $w_{dn} \mid z_{dn}, \beta \sim \text{multinomial}(\beta_{z_{dn}})$

For each document, both θ_d and z_d are unobserved variables and are local, $\theta_d \in \Delta_K, \beta_k \in \Delta_V, \forall k$. We find out that each topic mixture $\theta_d = (\theta_{d1}, \dots, \theta_{dK})$ represents the contributions of topics to document d , while β_{kj} shows the contribution of term j to topic k . LDA model described in Figure 1.

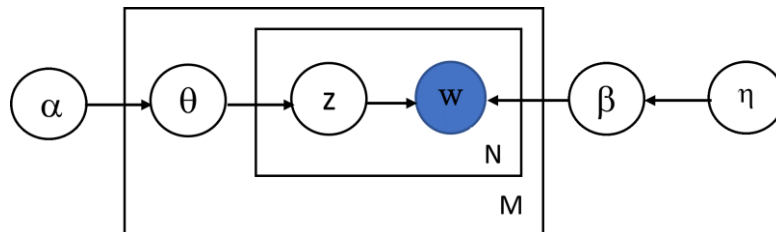


Figure 1. The graphic model for latent Dirichlet allocation

According to Teh *et al.* [28], given a corpus $\mathcal{C} = \{d_1, \dots, d_M\}$, the Bayesian inference (or learning) is to estimate the posterior distribution $P(z, \theta, \beta \mid \mathcal{C}, \alpha, \eta)$ over the latent topic indices $z = \{z_1, \dots, z_d\}$, topic mixtures $\theta = \{\theta_1, \dots, \theta_M\}$, and topics $\beta = (\beta_1, \dots, \beta_K)$. Given a model $\{\beta, \alpha\}$, the problem of posterior inference for each document d is to estimate the full joint distribution $P(z_d, \theta_d, d \mid \beta, \alpha)$. There are many research show that this distribution is intractable by direct estimation. Existing methods are usually sampling-based or optimization-based approaches, such as VB, CVB, and CGS. VB or CVB estimates the distribution by maximizing a lower bound of the likelihood $P(d \mid \beta, \alpha)$, CGS estimate $P(z_d \mid d, \beta, \alpha)$.

We consider the maximum a posteriori (MAP) estimation of topic mixture for a given document d :

$$\theta^* = \operatorname{argmax}_{\theta \in \Delta_K} P(\theta, d \mid \beta, \alpha) \tag{1}$$

Problem (1) is equivalent to the (2).

$$\theta^* = \operatorname{argmax}_{\theta \in \Delta_K} \left(\sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k \right) \tag{2}$$

We find out that:

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

Is non-concave when hyper-parameter $\alpha < 1$, then (2) is the non-concave optimization problem. Denote:

$$g(\theta) := \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}, \quad h(\theta) := (1 - \alpha) \sum_{k=1}^K \log \theta_k \quad (3)$$

And see that $g(\theta)$ and $h(\theta)$ are concave, then: $f(\theta) = g(\theta) - h(\theta)$ as the different concave (DC) function. We find that the problem (2) can be formulated as a DC optimization as:

$$\theta^* = \operatorname{argmax}_{\theta \in \Delta_K} [g(\theta) - h(\theta)] \quad (4)$$

There has been active research in the non-convex optimization. Some popular techniques such as branch and bound, cutting planes algorithm or DC algorithm (DCA) [29] for solving a DC optimization, but they are not suitable when applying in posterior inference (2) in probabilistic topic models. Note that CGS, CVB and VB are inference methods for probabilistic topic models. CGS, CVB and VB are popularly used in topic modeling, but we have not seen any theoretical analysis about how fast they do inference for individual documents. In addition, other candidates include concave-convex procedure (CCCP) [30], online frank-wolfe (OFW) [31], stochastic majorization-minimization (SMM) [32]. However, they are not sure about the convergence speed of the method and the quality of the model. In practice, the posterior inference in topic models is usually non-convex. Applying online-FW for solving a convex problem in [31], a new algorithm for MAP inference in LDA namely OFW have proposed by using a stochastic sequence combining with uniform distribution and show that convergence rate of OFW is $O\left(\frac{1}{\sqrt{t}}\right)$. Via doing many experiments with large datasets, OFW is a good approach for MAP problem and usually better than previous methods such as CGS, CVB and VB. Changing the learning rate and considering about theoretical aspect carefully, OPE algorithm has proposed. OPE approximates the true objective function $f(\theta)$ by a stochastic sequence $F_t(\theta)$ made up from the uniform distribution, thus (2) is easy for solving. OPE is better than previous methods, but we can explore a better new algorithm based on stochastic optimization for solving (2). Finding out the limitations of OPE, we improve OPE and obtain the GOPE2 algorithm applying for problem (2). Details of GOPE2 is presented in section 3.

3. PROPOSED METHOD

Finding out that OPE is a stochastic algorithm better than others for solving posterior inference. It also is quite simple and easily apply, so we improve OPE by randomization to obtain a better variant. We find out that the Bernoulli distribution is a discrete probability distribution of a random variable having two possible outcomes respectively with probabilities p and $1 - p$ and a special case of the Bernoulli one when probability $p = \frac{1}{2}$ is called the uniform distribution. We use Bernoulli distribution to construct the approximation functions to easily maximize and approximate well for objective function $f(\theta)$ we consider the problem.

$$\theta^* = \operatorname{argmax}_{\theta \in \Delta_K} \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k \quad (5)$$

We see that:

$$g_1(\theta) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} < 0, \quad g_2(\theta) = (\alpha - 1) \sum_{k=1}^K \log \theta_k > 0 \text{ then } g_1(\theta) < f(\theta) < g_2(\theta).$$

Pick f_h as a Bernoulli random sample from $\{g_1(\theta), g_2(\theta)\}$, where:

$$P(f_h = g_1) = p, \quad P(f_h = g_2) = 1 - p \text{ and make the approximation } F_t(\theta) = \frac{1}{t} \sum_{h=1}^t f_h.$$

The stochastic approximation $F_t(\theta)$ is easier to maximize and do differential than $f(\theta)$. We also see that $g_1(\theta) < 0$, $g_2(\theta) > 0$. Hence, if we choose $f_1 := g_1$ then $F_1(\theta) < f(\theta)$, which leads $F_t(\theta)$ is a lower bound of $f(\theta)$. In contrast, if we choose $f_1 := g_2$ then $F_1(\theta) > f(\theta)$, and $F_t(\theta)$ is an upper bound of $f(\theta)$. Using two stochastic approximation from above and below of $f(\theta)$ is better than one, we hope that will make the new algorithm has a faster converge rate. We use $\{L_t\}$ is an approximate sequences of $f(\theta)$ and begins with $g_1(\theta)$, another called $\{U_t\}$ begins with $g_2(\theta)$. We set $f_1^\ell := g_1(\theta)$. Pick f_t^ℓ as a Bernoulli random sample with probability p from $\{g_1(\theta), g_2(\theta)\}$ where $P(f_t^\ell = g_1(\theta)) = p, P(f_t^\ell = g_2(\theta)) = 1 - p$. We have $L_t := \frac{1}{t} \sum_{h=1}^t f_h^\ell, \forall t = 1, 2, \dots$. The sequence $\{L_t\}$ is a lower bound of the true objective $f(\theta)$.

By using an iterative approach, from a random sequence $\{L_t\}$, we obtain a numerical sequence $\{\theta_t^\ell\}$ ($t = 1, 2, \dots$) as:

$$e_t^\ell := \operatorname{argmax}_{x \in \Delta_K} \langle L'_t(\theta_t), x \rangle, \theta_{t+1}^\ell := \theta_t + \frac{e_t^\ell - \theta_t}{t} \tag{6}$$

Similarly, we set $f_1^u := g_2(\theta)$. Pick f_t^u as a Bernoulli random sample with probability p from $\{g_1(\theta), g_2(\theta)\}$, where $P(f_t^u = g_1(\theta)) = p$, $P(f_t^u = g_2(\theta)) = 1 - p$, $\forall t = 2, 3, \dots$, we have:

$$U_t := \frac{1}{t} \sum_{h=1}^t f_h^u, \forall t = 1, 2, \dots$$

The sequence $\{U_t\}$ is an upper bound of the true objective $f(\theta)$. From sequence $\{U_t\}$, we also obtain the numerical sequence $\{\theta_t^u\}$ ($t = 1, 2, \dots$) as:

$$e_t^u := \operatorname{argmax}_{x \in \Delta_K} \langle U'_t(\theta_t), x \rangle, \theta_{t+1}^u := \theta_t + \frac{e_t^u - \theta_t}{t} \tag{7}$$

We combine two approximating sequences $\{U_t\}$ and $\{L_t\}$. Based on two sequences $\{\theta_t^\ell\}$ and $\{\theta_t^u\}$, we construct an approximate solution sequence $\{\theta_t\}$ as:

$$\theta_t := \theta_t^u \text{ with probability } \frac{e^{f(\theta_t^u)}}{e^{f(\theta_t^u)} + e^{f(\theta_t^\ell)}} \text{ and } \theta_t := \theta_t^\ell \text{ with probability } \frac{e^{f(\theta_t^\ell)}}{e^{f(\theta_t^u)} + e^{f(\theta_t^\ell)}} \tag{8}$$

Then, we obtain GOPE2 algorithm. The effectiveness of GOPE2 depends on choosing the θ_t differently at each iteration. Details of GOPE2 are presented in Algorithm 1.

Algorithm 1. GOPE2 algorithm for the posterior inference

Input: document d , Bernoulli parameter $p \in (0, 1)$ and model $\{\beta, \alpha\}$

Output: θ that maximizes $f(\theta) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$

Initialize θ_1 arbitrarily in Δ_K

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

$f_1^u := g_1(\theta)$; $f_1^\ell := g_2(\theta)$

for $t = 2, 3, \dots, T$

Pick f_t^u randomly from $\{g_1(\theta), g_2(\theta)\}$ according to the Bernoulli distribution where:

$$P(f_t^u = g_1(\theta)) = p, P(f_t^u = g_2(\theta)) = 1 - p$$

$$U_t := \frac{1}{t} \sum_{h=1}^t f_h^u$$

$$e_t^u := \operatorname{argmax}_{x \in \Delta_K} \langle U'_t(\theta_t), x \rangle$$

$$\theta_t^u := \theta_{t-1} + \frac{e_t^u - \theta_{t-1}}{t}$$

Pick f_t^ℓ randomly from $\{g_1(\theta), g_2(\theta)\}$ according to the Bernoulli distribution where:

$$P(f_t^\ell = g_1(\theta)) = p, P(f_t^\ell = g_2(\theta)) = 1 - p$$

$$L_t := \frac{1}{t} \sum_{h=1}^t f_h^\ell$$

$$e_t^\ell := \operatorname{argmax}_{x \in \Delta_K} \langle L'_t(\theta_t), x \rangle$$

$$\theta_t^\ell := \theta_{t-1} + \frac{e_t^\ell - \theta_{t-1}}{t}$$

$$\theta_t := \theta_t^u \text{ with probability } q \text{ and } \theta_t := \theta_t^\ell \text{ with probability } 1 - q, \text{ where } q = \frac{e^{f(\theta_t^u)}}{e^{f(\theta_t^u)} + e^{f(\theta_t^\ell)}}$$

End for

The interweaving two-bounds of the objective function combine with Bernoulli distribution makes GOPE2 behave very differently from OPE. GOPE2 creates three numerical sequences $\{\theta_t^u\}$, $\{\theta_t^\ell\}$, and $\{\theta_t\}$ where $\{\theta_t\}$ depends on $\{\theta_t^u\}$ and $\{\theta_t^\ell\}$ at each iteration. The sequence $\{\theta_t\}$ really changes on structure, but the good properties of OPE are remained. There are many nice properties of GOPE2 that other algorithms do not have. Based on the online-OPE [23] for learning LDA, replacing OPE by GOPE2, we design the online-GOPE2 algorithm to learn LDA from large corpora. This algorithm employs GOPE2 to do maximum a posteriori estimation for individual documents to infer global variables such as topics.

4. EXPERIMENTAL RESULTS

In this section, we devote investigating GOPE2's behavior and show how useful it is when GOPE2 is used as a fast inference method to design a new algorithm for large-scale learning of topic models. We compare GOPE2 with other inference methods such as CGS, CVB, VB and OPE. Applying these inference methods to construct methods learning LDA such as online-CGS [12], online-CVB [21], online-VB [24] and online-OPE. We evaluate the GOPE2 algorithm indirectly via efficiency of online-GOPE2.

– Datasets

In our experiments we use two long-text large datasets: PubMed and New York Times datasets 1. We also use three short-text large datasets: Tweets from Twitter, NYT-titles from the New York Times where each document is the title of an article, Yahoo questions crawled from answers.yahoo.com. Details of these datasets are presented in Table 1.

– Parameter settings

We set $K = 100$ as the number of topics, the hyper-parameter $\alpha = \frac{1}{K}$ and $\eta = \frac{1}{K}$ as the topic Dirichlet parameter are commonly used in topic models. We also choose $T = 50$ as the number of iterations. We set $\kappa = 0.9, \tau = 1$ which are adapted best for inference methods. Performance measures: log predictive probability (LPP) [12] measures the predictability and generalization of a model to new data. Normalized pointwise mutual information (NPMI) [33] evaluates the semantic quality of an individual topic. From extensive experiments, NPMI agrees well with human evaluation on the interpretability of topic models. In this paper, we used LPP and NPMI to evaluate the learning methods. Choosing the Bernoulli parameter $p \in \{0.30, 0.35, \dots, 0.70\}$ and mini-batch size $|C_t| = 25,000$ on two long-text datasets, our experimental results are presented in Figure 2.

In Figure 2, we find out that the effectiveness of online-GOPE2 depends on the value of selected probability p and datasets. We also see that on the same measure, the results performed on the New York Times dataset are not too different as on the PubMed dataset and on the same dataset, the experimental results on the NPMI are different more than on LPP. We also see that our method usually is better than others when $p \approx 0.7$. Dividing the data into smaller mini-batches, $|C_t| = 5,000$ and the parameter Bernoulli $p \in \{0.1, 0.2, \dots, 0.9\}$ more extensive. Results of online-GOPE2 on two long-text datasets are presented in Figure 3.

In Figure 3, we find out that online-GOPE2 results depend much on the choose of parameter Bernoulli p and mini-batch size $|C_t|$. Through the experimental results, with the mini-batch size as 5,000, it gives results better than 25,000. We find out that when the mini-batch size decreases the value of measures increases, so the model learned better. Next, we compare online-GOPE2 with other learning such as online-CGS, online-CVB, and online-VB. These experimental results on long-text datasets: Pubmed and New York Times are showed in Figure 4, we see that online-GOPE2 is better than online-CGS, online-CVB, online-VB, and online-OPE on two datasets in LPP and NPMI measures.

Table 1. Five datasets in our experiments

Datasets	Corpus size	Average length per doc	Vocabulary size
PubMed	330,000	65.12	141,044
New York Times	300,000	325.13	102,661
Twitter tweets	1,457,687	10.14	89,474
NYT-titles	1,664,127	5.15	55,488
Yahoo questions	517,770	4.73	24,420

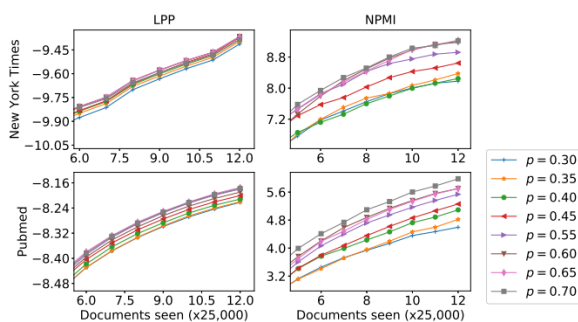


Figure 2. Predictiveness (LPP) and semantic quality (NPMI) of the learned models by online-GOPE2 with mini-batch size $|C_t| = 25,000$ on long-text datasets

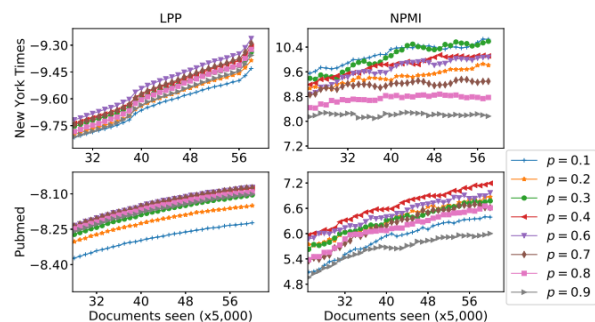


Figure 3. LPP and NPMI of the models learned by online-GOPE2 with mini-batch size $|C_t| = 5,000$ on long-text datasets

We also find out that LDA usually do not well on short texts. We provide additional evidence of GOPE2’s effectiveness by investigating the effectiveness of the learned model with short texts. We do experiments on three short-text datasets: Yahoo, Twitter and NYT-titles. Experimental results of online-GOPE2 on three short-text datasets are presented in Figure 5 and Figure 6.

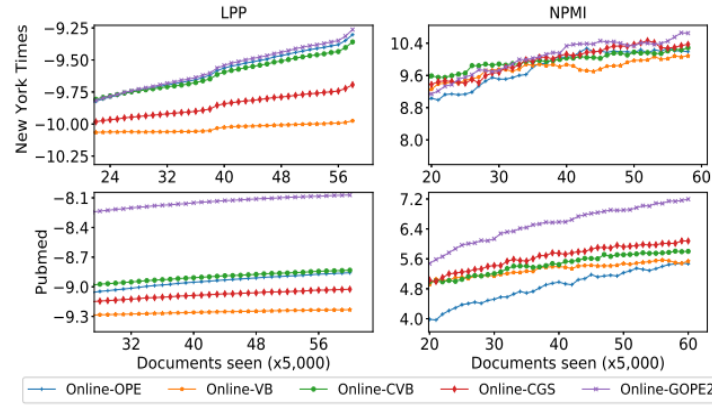


Figure 4. Performance of different learning methods on long-text datasets. Online-GOPE2 often surpasses all other methods

In Figure 6 and Table 2, we see that GOPE2 usually gives better results with parameter Bernoulli p chosen small on short-text datasets. Through Figure 6 we also see the model is over-fitting when learning by VB and CVB methods. The evidence is that the LPP and NPMI measures of the model by online-VB and online-CVB are reduced on three short-text datasets. Whereas, this do not happen for the GOPE2 method and variants. We do experiments with different mini-batch size and datasets, we show that our improvements usually give better than previous methods. GOPE2 gives better results than other methods because of the following reasons.

- Bernoulli distribution is more general than uniform. Bernoulli parameter p plays a role of the regularization parameter, then it makes our model avoid the over-fitting. This explains the contribution of prior/likelihood to solving the inference problem.
- Applying the squeeze theorem when constructing lower bound $\{L_t\}$ and upper bound $\{U_t\}$ of true objective function $f(\theta)$.

Table 2. Experimental results of some learning methods on short-text datasets

Datasets	Measures	Online-GOPE2	Online-OPE	Online-VB	Online-CVB	Online-CGS
NYT-titles	LPP	-8.4635	-8.6031	-9.6374	-9.5583	-8.4963
Twitter	LPP	-6.4297	-6.6943	-7.6152	-7.1264	-6.8151
Yahoo	LPP	-7.7222	-7.8505	-8.9342	-8.8417	-7.8501
NYT-titles	NPMI	4.1256	3.6037	0.6381	1.0348	4.6319
Twitter	NPMI	9.8042	9.3677	5.4541	6.0338	8.0621
Yahoo	NPMI	5.0205	4.4785	1.4634	1.9191	5.0181

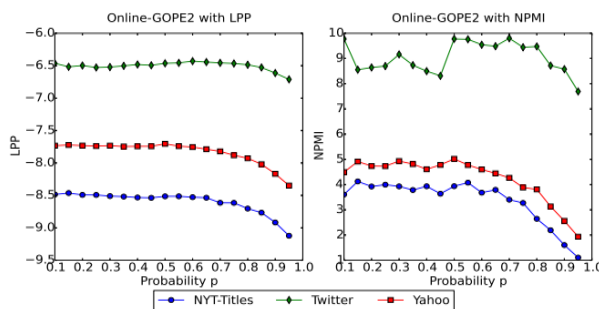


Figure 5. LPP and NPMI of the models learned by online-GOPE2 with Bernoulli parameter p and $|C_t| = 5,000$ on short-text datasets

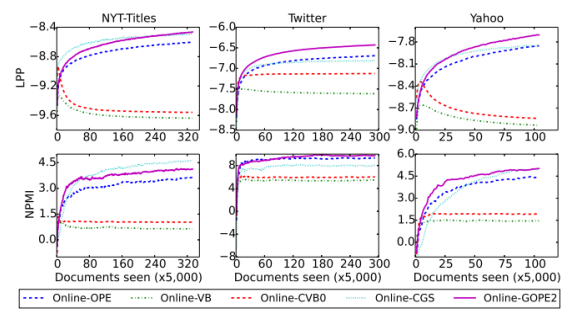


Figure 6. Performance of learning methods on short-text datasets. online-GOPE2 often surpasses other methods

5. CONCLUSION

The posterior inference for individual texts is very important in topic models. It directly determines the quality of the learned models. In this paper, we have proposed GOPE2, a stochastic optimization, helping the posterior inference problem can be solved well by using Bernoulli distribution and two stochastic approximations. In addition, the parameter Bernoulli p is seen as the regularization parameter that helps the model to be more efficient and avoid overfitting. Using GOPE2, we have online-GOPE2, an efficient method for learning LDA from data streams or large corpora. The experimental results show that GOPE2 is usually better than compared methods such as CGS, CVB, and VB. Thus, online-GOPE2 is a good candidate to help us deal with big data.




REFERENCES

- [1] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020, doi: 10.1162/tacl_a_00325.
- [2] M. Belford and D. Greene, "Ensemble topic modeling using weighted term co-associations," *Expert Systems with Applications*, vol. 161, 2020, doi: 10.1016/j.eswa.2020.113709.
- [3] K. E. Ihou and N. Bouguila, "Stochastic topic models for large scale and nonstationary data," *Engineering Applications of Artificial Intelligence*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103364.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com>
- [5] S. Ding, Z. Li, X. Liu, H. Huang, and S. Yang, "Diabetic complication prediction using a similarity-enhanced latent dirichlet allocation model," *Information Sciences*, vol. 499, pp. 12–24, 2019, doi: 10.1016/j.ins.2019.05.037.
- [6] A. Osmani, J. B. Mohasefi, and F. S. Gharehchopogh, "Enriched latent dirichlet allocation for sentiment analysis," *Expert Systems*, vol. 37, no. 4, 2020, doi: 10.1111/exsy.12527.
- [7] L. Pan *et al.*, "Latent dirichlet allocation based generative adversarial networks," *Neural Networks*, vol. 132, pp. 461–476, 2020, doi: 10.1016/j.neunet.2020.08.012.
- [8] L. Zheng, Z. Caiming, and C. Caixian, "MMDF-LDA: An improved multi-modal latent dirichlet allocation model for social image annotation," *Expert Systems with Applications*, vol. 104, pp. 168–184, 2018, doi: 10.1016/j.eswa.2018.03.014.
- [9] D. Falush, M. Stephens, and J. K. Pritchard, "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies," *Genetics*, vol. 164, no. 4, pp. 1567–1587, 2003, doi: 10.1093/genetics/164.4.1567.
- [10] D. Backenroth *et al.*, "FUN-LDA: A latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications," *The American Journal of Human Genetics*, vol. 102, no. 5, pp. 920–942, 2018, doi: 10.1016/j.ajhg.2018.03.026.
- [11] D. Valle, P. Albuquerque, Q. Zhao, A. Barberan, and R. J. Fletcher Jr, "Extending the latent Dirichlet allocation model to presence/absence data: A case study on north american breeding birds and biogeographical shifts expected from climate change," *Global change biology*, vol. 24, no. 11, pp. 5560–5572, 2018, doi: 10.1111/gcb.14412.
- [12] D. Mimno, M. D. Hoffman, and D. M. Blei, "Sparse stochastic inference for latent dirichlet allocation," in *Proc. of the 29th International Conference on Machine Learning (ICML-12)*. ACM, 2012, pp.1599–1606. [Online]. Available: <https://icml.cc/Conferences/2012/papers/784.pdf>
- [13] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 937–946, doi: 10.1145/1557019.1557121.
- [14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012, doi: 10.1145/2133806.2133826.
- [15] J. Grimmer, "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, 2010. doi: 10.1093/pan/mpp034.
- [16] Z. Chen, Y. Zhang, C. Wu, and B. Ran, "Understanding individualization driving states via latent dirichlet allocation model," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 41–53, 2019, doi: 10.1109/MITS.2019.2903525.
- [17] H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, M. E. P. Seligman, and L. H. Ungar, "Toward personality insights from language exploration in social media," in *AAAI Spring Symposium: Analyzing Microtext*, 2013, pp. 72–79. [Online]. Available: <https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5764/5915>
- [18] X. Cheng, Q. Cao, and S. S. Liao, "An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation," *Journal of Information Science*, vol. 48, no. 3, pp. 1–17, 2020, doi: 10.1177/0165551520954674.
- [19] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *PLoS one*, vol. 15, no. 9, 2020, doi: 10.1371/journal.pone.0239441.
- [20] A. Ålgå, O. Eriksson, and M. Nordberg, "Analysis of scientific publications during the early phase of the covid-19 pandemic: Topic modeling study," *Journal of medical Internet research*, vol. 22, no. 11, 2020, doi: 10.2196/21559.
- [21] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proc. of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 27–34. [Online]. Available: http://auai.org/uai2009/papers/UAI2009_0243_1a80458f5db72411c0c1e392f7dbbc48.pdf
- [22] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. of the National academy of Sciences*, 2004, vol. 101, pp. 5228–5235, doi: 10.1073/pnas.0307752101.
- [23] X. Bui, N. Duong, and T. Hoang, "GS-OPT: A new fast stochastic algorithm for solving the non-convex optimization problem," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 183–192, 2020, doi: 10.11591/ijai.v9.i2.pp183-192.
- [24] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, 2013. [Online]. Available: <https://www.jmlr.org/papers/volume14/hoffman13a/hoffman13a.pdf>
- [25] D. Sontag and D. M. Roy, "Complexity of inference in latent dirichlet allocation," in *Proc of the Neural Information Processing System (NIPS)*, 2011, pp. 1–9. [Online]. Available: <https://proceedings.neurips.cc/paper/2011/file/3871bd64012152bfb53fd04b401193f-Paper.pdf>
- [26] S. Ghadimi and G. Lan, "Stochastic first-and zero-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013, doi: 10.1137/120880811.




- [27] C. D. Dang and G. Lan, "Stochastic block mirror descent methods for nonsmooth and stochastic optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 856–881, 2015, doi: 10.1137/130936361.
- [28] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Advances in neural information processing systems*, 2006, pp. 1353–1360. [Online]. Available: <https://proceedings.neurips.cc/paper/2006/file/532b7cbe070a3579f424988a040752f2-Paper.pdf>
- [29] L. T. H. An and P. D. Tao, "The DC (difference of convex functions) programming and DCA revisited with dc models of real world nonconvex optimization problems," *Annals of Operations Research*, vol. 133, pp. 23–46, 2005, doi: 10.1007/s10479-004-5022-1.
- [30] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, 2003, doi: 10.1162/08997660360581958.
- [31] E. Hazan and S. Kale, "Projection-free online learning," in *Proc. of the 29th International Conference on Machine Learning*, 2012. [Online]. Available: <https://icml.cc/2012/papers/292.pdf>
- [32] J. Mairal, "Stochastic majorization-minimization algorithms for large-scale optimization," in *Neural Information Processing System (NIPS)*, 2013, pp. 1–9. [Online]. Available: <https://proceedings.neurips.cc/paper/2013/file/4da04049a062f5adfe81b67dd755cecc-Paper.pdf>
- [33] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539. [Online]. Available: <https://aclanthology.org/E14-1056.pdf>

BIOGRAPHIES OF AUTHORS



Hoang Quang Trung    received B.Sc. and M.Sc. degrees in Electronics and Telecommunications from the Vietnam National University respectively in 2003 and 2010. He also received the PhD degree in Electronic Engineering from the Le Quy Don Technique University, Vietnam, in 2017. He is currently a lecturer at the Phenikaa University, Hanoi, Vietnam. His research interests include signal and data processing for telecommunication systems, internet of things (IoT). He can be contacted at email: trung.hoangquang@phenikaa-uni.edu.vn.



Xuan Bui    received B.Sc. in applied mathematics and informatics and M.Sc. in information technology in 2003 and 2007. She also received the PhD degree in information systems from Hanoi University of Science and Technology, Vietnam, in 2020. She is currently a lecturer at the Thuyloi University, Hanoi, Vietnam. Her current research interests include optimization in machine learning, stochastic optimization, deep learning. She can be contacted at email: xuanbtt@tlu.edu.vn.