ISSN 2007-9737

A Flexible Stochastic Method for Solving the MAP Problem in Topic Models

Tu Vu[1], Xuan Bui[1,2], Khoat Than[1], Ryutaro Ichise[3]

[1] Hanoi University of Science and Technology, Hanoi, Vietnam

[2] Thai Nguyen University of Information and Communication Technology, Vietnam

[3] National Institute of Informatics, Tokyo, Japan

{vutu201130, thanhxuan1581}@gmail.com, khoattq@soict.hust.edu.vn, ichise@nii.ac.jp

Abstract. The estimation of the posterior distribution is the core problem in topic models, unfortunately it is intractable. There are approximation and sampling methods proposed to solve it. However, most of them do not have any clear theoretical guarantee of neither quality nor rate of convergence. Online Maximum a Posteriori Estimation (OPE) is another approach with concise guarantee on quality and convergence rate, in which we cast the estimation of the posterior distribution into a non-convex optimization problem. In this paper, we propose a more general and flexible version of OPE, namely Generalized Online Maximum a Posteriori Estimation (G-OPE), which not only enhances the flexibility of OPE in different real-world situations but also preserves key advantage theoretical characteristics of OPE when comparing to the state-of-the-art methods. We employ G-OPE as inference a document within large text corpora. The experimental and theoretical results show that our new approach performs better than OPE and other state-of-the-art methods.

Keywords. Topic models, posterior inference, online MAP estimation, large-scale learning, non-convex optimization.

1 Introduction

Topic models are widely used in text processing and Latent Dirichlet Allocation (LDA) [3] is the core of a large family of probabilistic models. LDA provides an efficient tool to analyze hidden themes in data and helps us recover hidden structures/evolution in big text collections. The key problem in topic models is to compute the posterior distribution of a document given other parameters. The posterior inference problem in topic models is to infer the topic proportion of documents and topics which are distributions over vocabulary. Large datasets or streaming environments contain huge number of documents, hence the problem of estimating topic proportion for an individual document is especially important. The quality of learning for LDA is determined by the quality of the inference method being employed.

Unfortunately, solving directly a posterior distribution of a document is intractable [3]. There are two main approaches to tackle it.

One is approximating the intractable distribution by tractable distribution, for example Variational Bayes inference (VB) [3]. The other is a sampling method, which draws numerous the samples from target distribution then estimating the interesting quality from these samples. The well-known method is Collapsed Gibbs Sampling (CGS) [8]. There are also famous methods such as Collapsed

Variational Bayes (CVB) [15], CVB0 [2], Stochastic Variational Inference (SVI) [10], etc.

To our best knowledge, there are not any mathematical guarantees for quality and convergence rate in existing approaches. Therefore, in practice we do not have any ideas about how to stop the methods we are using but trying, observing and retrying again to reach the best solution.

Another way to solve the posterior distribution is to view it as an optimization problem. To infer about topic proportion of a document is to solve the maximum a posteriori of topic proportion given words in this document and all topics of corpus [16]. This optimization problem is usually non-convex and NP-hard [14]. There is very few theoretical contributions in non-convex optimization literature, especially in topic models. Online Maximum a Posteriori Estimation (OPE) [16] which is an online version of Frank-Wolfe algorithm [9] is a stochastic algorithm to solve such kind of non-convex problem.

OPE is theoretically guaranteed to converge to a local stationary point [16]. Although OPE is easy to implement and has fast convergence and mathematically guaranteed, it remains some problems. The weakness of OPE is that it is not well adaptive with different data sets because of the uniform distribution in its operation. We will exploit this crucial point to propose a new and more general algorithm based on OPE. When changing its operations, we have to retain the advantage of the original algorithms, that is theoretical guarantees.

Our main contribution is following:

— We propose new algorithm called Generalized Online Maximum a Posteriori Estimation (G-OPE) for solving posterior inference problem in topic models. G-OPE is more general and flexible than OPE, adapts better in different datasets and preserves the key advantages OPE.

— We employed G-OPE into the existing algorithm Online-OPE [16] to learn LDA in online settings and streaming environments.

— We conduct experiments to demonstrate that Online-GOPE outperforms existing methods to learn LDA.

Organization: The rest of this paper is organized as follows. In Section 2, we introduce an overview of posterior inference with LDA and main ideas of existing methods. In Section 3, our new algorithm G-OPE is proposed in details. In Section 4, we conduct experiments with two large datasets with state-of-the-art methods in two different measures. Finally Section 5 is our conclusion.

Notation: Throughout the paper, we use the following conventions and notations. Bold faces denote vectors or matrices. x_i denotes the i^{th} element of vector $\boldsymbol{x}$, and A_{ij} denotes the element at row i and column j of matrix $\boldsymbol{A}$. 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 $\boldsymbol{V}$ dimensions (dictionary size). Each document $\boldsymbol{d}$ will be represented as frequency vector, $\boldsymbol{d} = (d_1, .., d_V)^T$, where d_j represents the frequency of term j in $\boldsymbol{d}$. Denote n_d as the length of $\boldsymbol{d}$, i.e., $n_d = \sum_j d_j$. The inner product of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is denoted as $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$. $\boldsymbol{I}(x)$ is the indicator function which returns 1 if x is true, and 0 otherwise and $\boldsymbol{E}(X)$ is expectation of random variable $\boldsymbol{X}$.

2 Related Work

LDA [3] is the basic and famous model in topic modeling. It models each document as a probability distribution $\boldsymbol{\theta}_d$ over topics, and each topic $\boldsymbol{\beta}_k$ as a probability distribution over words. In Fig.1, K is number of topics, M is number of documents in corpus, N is number of words in each documents.

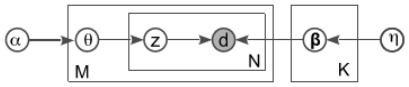

Fig. 1. Latent Dirichlet Allocation

Note that $\boldsymbol{\theta}_d \in \Delta_K,\ \boldsymbol{\beta}_k \in \Delta_V$. The generative process for each document $\boldsymbol{d}$ is as follows:

1. Draw a topic distribution $\boldsymbol{\theta}_d | \alpha \sim Dirichlet(\alpha)$

2. For the n^{th} word of $\boldsymbol{d}$:
   – draw topic index $z_{dn}|\boldsymbol{\theta}_d \sim Multinomial(\boldsymbol{\theta}_d)$
   – draw word $w_{dn}|z_{dn}, \boldsymbol{\beta} \sim Multinomial(\boldsymbol{\beta}_{z_{dn}})$

The most important problem we need to solve in order to use LDA is to compute the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \alpha, \boldsymbol{\beta})$ of hidden variables in a given document $\boldsymbol{d}$. However, it is intractable. There are many ways to handle it. Variational Bayesian Inference [3] approximates $p(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{d}|\boldsymbol{\beta}, \alpha)$ by obtaining a lower bound on the likelihood which is adjustable by variational distributions. CVB and CVB0 deal with $p(\boldsymbol{z}_d, d|\boldsymbol{\beta}, \alpha)$, CGS draws samples from $p(\boldsymbol{z}_d, \boldsymbol{w}|\boldsymbol{\beta}, \alpha)$ to estimate it. Eventually, all methods try to estimate the topic proportion $\boldsymbol{\theta}_d$.

In this paper, we infer topic proportion for a document directly by solving the Maximum a Posteriori Estimation (MAP) of $\boldsymbol{\theta}_d$ given all words of this document and parameters of the model. The MAP estimation of topic mixture for a given document $\boldsymbol{d}$:

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

using Bayes' rule, we have:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \overline{\Delta}_K} \Pr(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\beta}) \Pr(\boldsymbol{\theta}|\alpha). \tag{2}$$

Under the assumption about the generative process, problem (2) is equivalent to the following:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \overline{\Delta}_K} \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k. \tag{3}$$

Within convex/concave optimization, problem (3) is relatively well-studied. In the case of $\alpha \geq 1$, it can easily be shown that the problem (3) is concave, and therefore it can be solved in polynomial time.

Unfortunately, in practice of LDA, the parameter α is often small, says $\alpha < 1$, causing problem (3) to be non-concave. Sontag et al. in [14] has showed that problem (3) is NP-hard in the worst case when parameter $\alpha < 1$. Consider problem (3) as a non-convex optimization problem, the gradient-based methods such as Gradient Descent (GD) and its variants are ineffective because of the existence of saddle points and flat regions, hence we need an effective random method to avoid them. OPE [16] is an efficient iterative algorithm for solving problem (3). It is a good solution in escaping saddle points and flat regions.

Algorithm 1 OPE: Online Maximum a Posteriori Estimation

Input: document $\boldsymbol{d}$ 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 θ_1 arbitrary in Δ̄_K = {x ∈ R^K : Σ_{k=1}^K x_k = 1, x ≥ ε > 0}
for t = 1, 2, ...T do
    Pick f_t uniformly from
    {Σ_j d_j log Σ_{k=1}^K θ_j β_kj; (α − 1) Σ_{k=1}^K log θ_k}
    F_t := (2/t) Σ_{h=1}^t f_h
    e_t := arg max_{x ∈ Δ_K} ⟨F_t'(θ_t), x⟩
    θ_{t+1} := θ_t + (e_t − θ_t)/t
end for
```

In the literature of iterative optimization algorithms, in each iteration, they try to build a tractable function that approximates true objective function, then optimize approximating function to reach the next point. The various algorithms have different techniques to build their own approximation. For example, using Jensen's inequality, Expectation-Maximization (EM) [5] or Variational Inference (VI) [3] calculate the Evidence Lower Bound (ELBO) then maximize it. Gradient Descent constructs its quadratic approximation in each step and minimizes the quadratic. OPE solves the problem (3) by constructing an approximate sequence by stochastic way and solve it by Frank-Wolfe update formula [7].

Details of OPE is in Algorithm 1. The idea of OPE is quite simple. At each iteration t, it draws a sample function $f_t(\boldsymbol{\theta})$ and builds the approximation $F_t(\boldsymbol{\theta})$ which is the average of all previous sample function. The most interesting idea behind OPE is that the objective function is the sum of a likelihood and a prior. In each step, it builds an approximate function $F_t(\boldsymbol{\theta})$ by choosing either likelihood or prior with equal probabilities $\{0.5,\ 0.5\}$. That means when inferring about the topic proportion of a document, we use either the evidence of the document (likelihood) or knowledge we have known before (prior). This behavior is very natural

to human. However, OPE considers likelihood and prior with the same contributions by using uniform distribution.

In fact, when humans deal with a new sample, one can rely on more likelihood if we have observed enough evidences, or rely on more prior knowledge if we have been lack of evidences. This simple idea leads us to build a more general and flexible version of OPE by using Bernoulli distribution instead of uniform distribution.

3 Generalized Online Maximum a Posteriori Estimation

In this section, we introduce our new algorithm, namely Generalized Online Maximum a Posteriori Estimation (G-OPE) based on OPE. OPE operates by choosing the likelihood or prior at each step t, then builds the approximation $F_t(\boldsymbol{\theta})$ which is the average of all parts draw from previous steps and current step. In G-OPE, in order to introduce the Bernoulli distribution into the sampling step, we need to modify the likelihood and prior so that the approximation function $F_t(\boldsymbol{\theta}) \rightarrow f(\boldsymbol{\theta})$ as $t \rightarrow \infty$. Denote:

$$g_1(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj},$$

$$g_2(\boldsymbol{\theta}) = (\alpha - 1) \sum_{k=1}^{K} \log \theta_k,$$

then the true objective function $f(\boldsymbol{\theta})$ includes two components:

$$f(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta}),$$

where $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ are the log likelihood and prior respectively.

Denote:

$$G_1(\boldsymbol{\theta}) := \frac{g_1(\boldsymbol{\theta})}{p} \ , \ G_2(\boldsymbol{\theta}) := \frac{g_2(\boldsymbol{\theta})}{1-p},$$

where $G_1(\boldsymbol{\theta})$ and $G_2(\boldsymbol{\theta})$ are the adjusted likelihood and prior respectively.

G-OPE is detailed in Algorithm 2. In Algorithm 2, $f(\boldsymbol{\theta})$ is the true objective function we need to maximize. At t^{th} iteration, we draw sample function $f_t(\boldsymbol{\theta})$ from set of adjusted likelihood $G_1(\boldsymbol{\theta})$ and prior $G_2(\boldsymbol{\theta})$, then we build the approximate function $F_t(\boldsymbol{\theta})$. Because G-OPE is stochastic, in theory we consider $T \rightarrow \infty$, where T is number of iterations for whole algorithm.

We use Bernoulli distribution with parameter p to replace for uniform distribution in OPE. At t^{th} iteration, we pick $f_t(\boldsymbol{\theta})$ as Bernoulli random variable with probability p from $\{G_1(\boldsymbol{\theta}) \ , \ G_2(\boldsymbol{\theta})\}$ where:

$$\Pr(f_t(\boldsymbol{\theta}) = G_1(\boldsymbol{\theta})) = p,$$
$$\Pr(f_t(\boldsymbol{\theta}) = G_2(\boldsymbol{\theta})) = 1 - p.$$

In statistic theory, as t increases (at least 20) and it is better to choose p not close to 0 or 1. Consider t independent Bernoulli trials with probabilities:

$$\{\Pr(f_h = G_1) = p \, , \, \Pr(f_h = G_2) = 1-p\} \ \forall h = 1, .., t,$$

we build a stochastic approximate sequence:

$$F_t := \frac{1}{t} \sum_{h=1}^{t} f_h, \ \forall t = 1, 2, ..., T.$$

We find out that $F_t(\boldsymbol{\theta})$ is the average of all sample functions drawn until current step.

So it is guaranteed to converge to $f(\boldsymbol{\theta})$ as $t \rightarrow \infty$, which will be shown in Theorem 1. The Bernoulli parameter p controls how much likelihood part and prior part contribute to the objective function $f(\boldsymbol{\theta})$. We can utilize this point to choose the most suitable p in each circumstance. OPE is a special case of G-OPE when Bernoulli parameter p is chosen equal to 0.5. So OPE is not flexible in many datasets. G-OPE adapts well with different datasets, we will show it in the experiment section. In the rest of this section, we will show that G-OPE preserves the key advantage of OPE which is the guarantee of the quality and convergence rate. This character is unknown for the existing methods in posterior estimation in topic models.

Theorem 1 (Convergence of G-OPE algorithm) *Consider the objective function $f(\boldsymbol{\theta})$ in Eq.3, given fixed $\boldsymbol{d}, \boldsymbol{\beta}, \alpha, p$. For G-OPE, with probability one, the followings hold:*

1. For any $\boldsymbol{\theta} \in \Delta_K$, $F_t(\boldsymbol{\theta})$ converges to $f(\boldsymbol{\theta})$ as $t \rightarrow +\infty$.

2. $\boldsymbol{\theta}_t$ converges to a local maximal/stationary point of $f(\boldsymbol{\theta})$.

Algorithm 2 G-OPE: Generalized Online maximum a Posteriori Estimation

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

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

$g_1(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$

$g_2(\boldsymbol{\theta}) = (\alpha - 1) \sum_{k=1}^K \log \theta_k$

$G_1(\boldsymbol{\theta}) := \dfrac{g_1(\boldsymbol{\theta})}{p}$, $G_2(\boldsymbol{\theta}) := \dfrac{g_2(\boldsymbol{\theta})}{1-p}$

for $t = 1, 2, ..., T$ **do**

Pick f_t as Bernoulli distribution from $\{G_1(\boldsymbol{\theta}),\ G_2(\boldsymbol{\theta})\}$ where

$$\Pr(f_t(\boldsymbol{\theta}) = G_1(\boldsymbol{\theta})) = p\ ,$$
$$\Pr(f_t(\boldsymbol{\theta}) = G_2(\boldsymbol{\theta})) = 1 - p$$

$F_t(\boldsymbol{\theta}) := \frac{1}{t} \sum_{h=1}^t f_h(\boldsymbol{\theta})$

$\boldsymbol{e}_t := \arg\max_{\boldsymbol{x} \in \Delta_K} \langle F_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x} \rangle$

$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}$

end for

Proof: Before the proof, we remind some notations: $B(n, p)$ is binomial distribution with parameters n and p (Bernoulli distribution is a special case of the binomial distribution with $n = 1$), $N(\mu,\ \sigma^2)$ is normal distribution. $E(X)$ and $D(X)$ are expectation and variance of random variable X respectively.

We find out that problem 3 is the constrained optimization problem with the objective function $f(\boldsymbol{\theta})$ is non-convex. The criterion used for the convergence analysis is importance in non-convex optimization. For unconstrained problems, the gradient norm $\|\nabla f(\boldsymbol{\theta})\|$ is typically used to measure convergence, because $\|\nabla f(\boldsymbol{\theta})\| \to 0$ captures convergence to a stationary point. However, this criterion can not be used for constrained problems. Instead, we use the "Frank-Wolfe gap" criterion in [13].

Denoted:

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

and

$$G_1(\boldsymbol{\theta}) := \frac{g_1(\boldsymbol{\theta})}{p}\ ,\ G_2(\boldsymbol{\theta}) := \frac{g_2(\boldsymbol{\theta})}{1-p},$$

so, $f(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta}) = p.G_1(\boldsymbol{\theta}) + (1-p)G_2(\boldsymbol{\theta})$.

Pick f_t follows the Bernoulli distribution from $\{G_1(\boldsymbol{\theta}), G_2(\boldsymbol{\theta})\}$ where:

$$\Pr(f_t = G_1(\boldsymbol{\theta})) = p\ ,\ \Pr(f_t = G_2(\boldsymbol{\theta})) = 1 - p.$$

Let a_t and b_t be the number of times that we have already picked $G_1(\boldsymbol{\theta})$ and $G_2(\boldsymbol{\theta})$ respectively after t iterations.

We find that $a_t + b_t = t$ or $b_t = t - a_t$. We have $a_t \sim B(t, p)$ and $E(a_t) = t.p\ ,\ D(a_t) = t.p.(1-p)$.

We have:

$$F_t = \frac{1}{t}(a_t G_{_1} + b_t G_2),$$
$$F_t - f = \frac{a_t - t.p}{t}(G_1 - G_2) = \frac{S_t}{t}(G_1 - G_2), \quad (4)$$
$$F_t' - f' = \frac{a_t - t.p}{t}(G_1' - G_2') = \frac{S_t}{t}(G_1' - G_2'),$$

where $S_t = a_t - t.p$.

We have:

$$E(S_t) = 0\ ,\ D(S_t) = tp(1-p),$$

then $S_t \to N(0,\ tp(1-p))$ when $t \to \infty$.

So $S_t/t \to 0$ as $t \to \infty$ with probability one. From (4), we conclude that the $F_t \to f$ as $t \to +\infty$ with probability one.

Consider:

$$\langle F_t'(\boldsymbol{\theta_t}), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t} \rangle =$$
$$= \langle F_t'(\boldsymbol{\theta}_t) - f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t} \rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t} \rangle$$
$$= \frac{S_t}{t^2} \langle G_1^{'}(\boldsymbol{\theta}_t) - G_2^{'}(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t \rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t} \rangle.$$

Note that $g_1(\boldsymbol{\theta}),\ g_2(\boldsymbol{\theta})$ are Lipschitz continuous on $\overline{\Delta}_K$. Hence there exists a constant L such that:

$$\langle f'(z), y-z\rangle \leq f(y)-f(z)+L\|y-z\|^2\ \ \forall y,z \in \overline{\Delta}_K.$$

We have:

$$\begin{aligned}\langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t-\boldsymbol{\theta}_t}{t}\rangle &= \langle f'(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\rangle \\ &\leq f(\boldsymbol{\theta}_{t+1})-f(\boldsymbol{\theta}_t)+L\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|^2 \\ &= f(\boldsymbol{\theta}_{t+1})-f(\boldsymbol{\theta}_t)+\frac{L}{t^2}\|\boldsymbol{e}_t-\boldsymbol{\theta}_t\|^2.\end{aligned}$$

Since $\boldsymbol{e}_t$ and $\boldsymbol{\theta}_t$ belong to $\overline{\Delta}_K$ then $|\langle G_1'(\boldsymbol{\theta}_t) - G_2'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle|$ and $\|\boldsymbol{e}_t-\boldsymbol{\theta}_t\|^2$ are bounded above for any t.

Therefore, there exits a constant $c_1 > 0$ such that:

$$\langle F_t'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t-\boldsymbol{\theta}_t}{t}\rangle \leq c_1\frac{|S_t|}{t^2}+f(\boldsymbol{\theta}_{t+1})-f(\boldsymbol{\theta}_t)+\frac{c_1L}{t^2}. \tag{5}$$

Summing both sides of (5) for all t, we have:

$$\begin{aligned}&\sum_{h=1}^{t}\frac{1}{h}\langle F_h'(\boldsymbol{\theta}_h), \boldsymbol{e}_h-\boldsymbol{\theta}_h\rangle \leq \\ &\leq \sum_{h=1}^{t} c_1\frac{|S_h|}{h^2}+f(\boldsymbol{\theta}_{t+1})-f(\boldsymbol{\theta}_1)+\sum_{h=1}^{t}\frac{c_1L}{h^2}.\end{aligned} \tag{6}$$

As $t \to +\infty$, $f(\boldsymbol{\theta}_t) \to f(\boldsymbol{\theta}^*)$ due to the continuity of $f(\boldsymbol{\theta})$. As a result, (6) implies:

$$\begin{aligned}&\sum_{h=1}^{+\infty}\frac{1}{h}\langle F_h'(\boldsymbol{\theta}_h), \boldsymbol{e}_h-\boldsymbol{\theta}_h\rangle \leq \\ &\leq \sum_{h=1}^{+\infty} c_1\frac{|S_h|}{h^2}+f(\boldsymbol{\theta}^*)-f(\boldsymbol{\theta}_1)+\sum_{h=1}^{+\infty}\frac{c_1L}{h^2}.\end{aligned} \tag{7}$$

Note that $S_h = \mathcal{O}(\sqrt{h\log h})$ [6], and hence $\sum_{h=1}^{\infty} c_1\frac{|S_h|}{h^2}$ converges in probability one. Moreover, the term $\sum_{h=1}^{+\infty}\frac{1}{h^2}$ is bounded.

So $\sum_{h=1}^{+\infty}\frac{1}{h}\langle F_h'(\boldsymbol{\theta}_h), \boldsymbol{e}_h-\boldsymbol{\theta}_h\rangle$ is bounded above.

Because $\boldsymbol{e}_t = \arg\max_{\boldsymbol{x}\in\Delta_K}\langle F_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x}\rangle$, so $\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle \geq 0$.

If exists $t_0 > 0, c_3 > 0$ such as $\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle \geq c_3 \forall t > t_0$ then $\sum_{t=1}^{\infty}\frac{1}{t}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle > \sum_{t=1}^{\infty}\frac{c_3}{t}$. And because $\sum_{t=1}^{\infty}\frac{1}{t}$ is not bounded above, so $\sum_{h=1}^{+\infty}\frac{1}{h}\langle F_h'(\boldsymbol{\theta}_h), \boldsymbol{e}_h-\boldsymbol{\theta}_h\rangle \to \infty$, which contradicts with the clause we claimed. Therefore:

$$\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle \to 0 \text{ as } t \to \infty.$$

$$\begin{aligned}&\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle = \\ &= \langle f'(\boldsymbol{\theta}_t)+\frac{S_t}{t}(G_1'(\boldsymbol{\theta}_t)-G_2'(\boldsymbol{\theta}_t)), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle \\ &= \langle f'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle + \langle\frac{S_t}{t}(G_1'(\boldsymbol{\theta}_t)-G_2'(\boldsymbol{\theta}_t)), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle\end{aligned}\ .$$

Since $\frac{S_t}{t} \to 0$, then $\langle f'(\boldsymbol{\theta}_t), \boldsymbol{e}_t-\boldsymbol{\theta}_t\rangle \to 0$. Apply Frank-Wolfe gap criterion, $\boldsymbol{\theta}^*$ is stationary/local maximum of f, which completes the proof.

Besides, in the non-convex optimization field, the idea of how to build the approximate function in G-OPE can be utilized in the case of objective function f which is the sum of two parts $f = g + h$. In each step, choose g or h in Bernoulli distribution with parameter p, and adjust p to adapt with different circumstance. Randomness can help algorithms jump out of local minimum/maximum.

Therefore, to design new stochastic algorithms, we begin with a deterministic version, add a sequence of approximation in the G-OPE style, working with each approximation at each iteration by deterministic update formula. This is an open idea for our future works.

4 Experiments

In this section, we will investigate the performance of G-OPE in real world datasets. G-OPE can play as the core inference step when learning LDA, we will investigate the performance of G-OPE through the performance of Online-OPE [16] when changing its core inference method. So we derived Online-GOPE.

We conducted two experiments. The first one is the effect of parameter p in G-OPE when learning LDA and the second is in comparison Online-GOPE with the current state-of-the-art methods.

4.1 Datasets and Settings

The datasets for our investigation are New York Times and Pubmed[1]. These are very large datasets. The number of documents is large and the size of vocabulary is large also. Details of datasets are presented in Table 1.

To evaluate the performance of learning methods in LDA, we used *Log Predictive Probability (LPP)* and *Normalized Pointwise Mutual Information (NPMI)* measures. These measures is commonly used in topic models. *Predictive Probability* [10] measures the predictiveness and generalization of a model to new data, while NPMI [1, 4] evaluates semantics quality of an individual topic in these models.

Some common parameters is set as follows: the number of topics is $K = 100$, the hyper-parameters in LDA model is $\alpha = \frac{1}{K} = 0.01, \eta = \frac{1}{K} = 0.01$. For each inference method, the number of iterations is $T = 50$. We compare the online learning algorithms together and the mini-batch size is $S = |C_t| = 5000$. For the other state-of-the-art methods, the forgetting rate$\kappa = 0.9$, we fixed $\tau = 1$. These chosen parameters is best for online learning LDA in many previous works.

As algorithms we compares are stochastic, so to avoid randomness, we run each method five times, and report the average results.

The script of experiments is that: for the first experiment, we run Online-GOPE with different values of parameter p then choose the best one. In the second experiment, we compare Online-GOPE obtained with the best parameter p to some methods in learning LDA such as VB, CVB, CGS, OPE.

4.2 The Effect of Bernoulli Parameter p

In this experiment, we investigate how important the value of parameter p is. Because $p \in (0, 1)$, and p is good if it is not close to 0 and 1. So we choose p respectively in $\{0.1, 0.15, ..., 0.9\}$, then run Online-GOPE in two datasets. We report the performance of Online-GOPE in Fig.2 and Fig.3. We can easly observe that p affects very much in the performance in terms of both measures. In Fig.2, Online-GOPE reaches the best performance on New York Times for LPP measure at $p = 0.35$ and for NPMI measure at $p = 0.75$. In Fig.3, Online-GOPE reaches the best performance on Pubmed for LPP measure at $p = 0.4$, for NPMI measure at $p = 0.45$.

This results support our idea about the contributions of likelihood part and prior part of topic proportion inference for a document. The different dataset has the suitable value of p. If we want to get the best performance on the generalization or on semantics quality of topics, we have different p to choose. Therefore G-OPE is very flexible in the real world dataset.

The good values of p depend on how much likelihood part and prior part possess in total. The likelihood depends on the length of the documents. In our datasets, the average length of a document in New York Times is 329 while the average length of a document in Pubmed is 65. That explains why we have different best values of p for each dataset.

4.3 Comparison of G-OPE with Novel Algorithms

In this experiment, we compare Online-GOPE with the best value of p in previous experiment to the original Online-OPE and other methods: Online-VB, Online-CVB, Online-CGS. All of these algorithms try to learn the topics over the words $\boldsymbol{\beta}$ or variational parameters $\boldsymbol{\lambda}$. The difference among these algorithms is the inner inference procedures.

The results is shown in Fig.4 and 5. With suitable parameter p, we obtained G-OPE which was better than OPE, VB, CVB, and CGS on LPP measure. For NPMI measure, all algorithms perform the same, but G-OPE is one of the tops.

This results show that Online-GOPE performs better than not only original OPE, but also the current novel methods. G-OPE works well because of the right choose of controlled parameter p.

[1]The datasets were taken from http://archive.ics.uci.edu/ml/

Table 1. Two data sets for our experiments

Data sets	No.Documents	No.Terms	No.Train	No.Test
New York Times	300000	141444	290000	10000
Pubmed	330000	100000	320000	10000

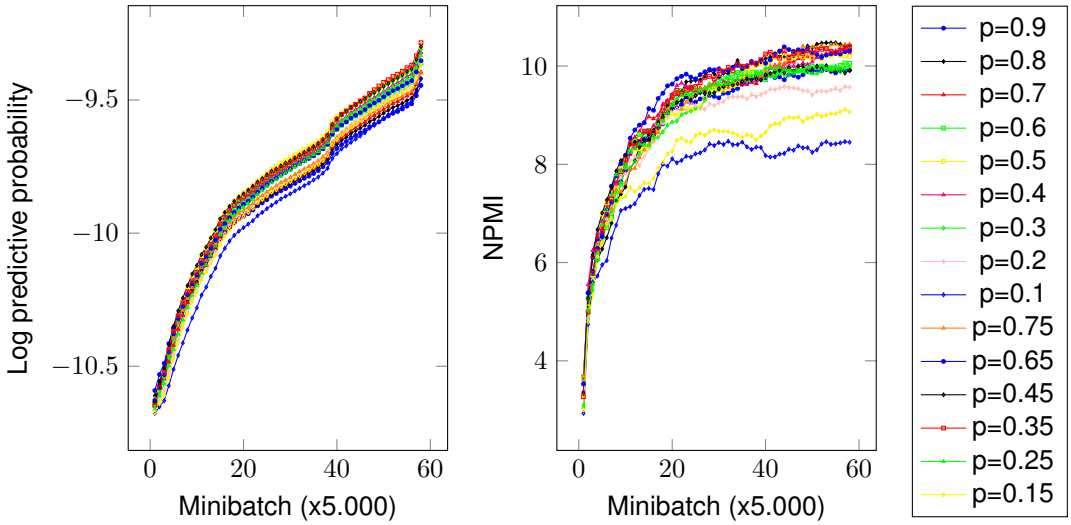

Fig. 2. Online-GOPE with different values of p on New York Times

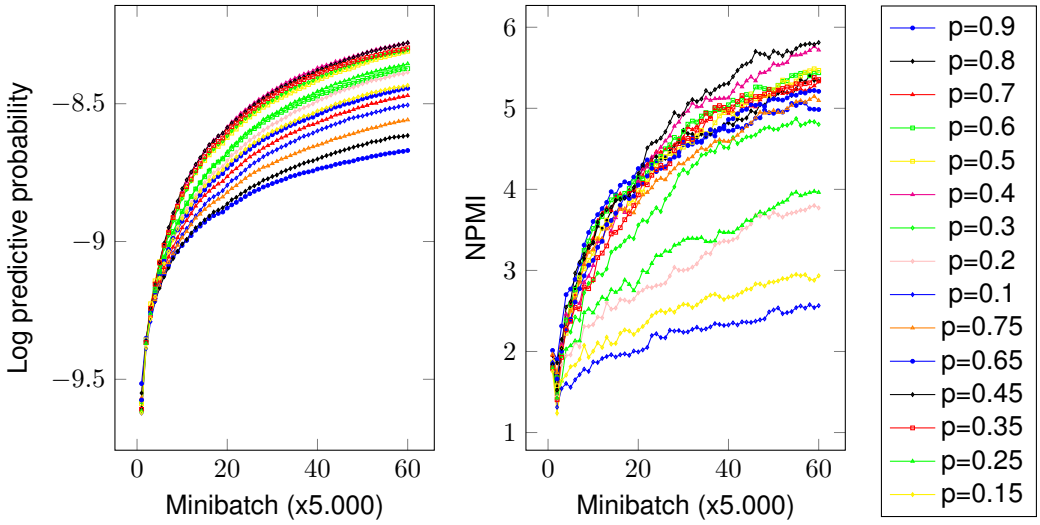

Fig. 3. Online-GOPE with different values of p on Pubmed

5 Conclusion

We have discussed how posterior inference for individual texts in topic models can be done efficiently with our method. In theory, G-OPE remains the guarantee on quality and convergence

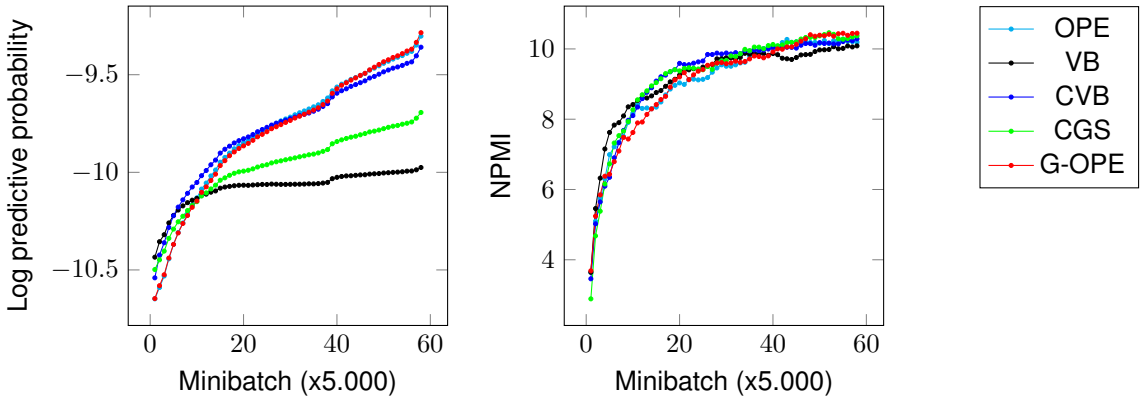

Fig. 4. Online-GOPE compares with Online-OPE, Online-VB, Online-CVB and Online-CGS on New York Times dataset. Higher is better

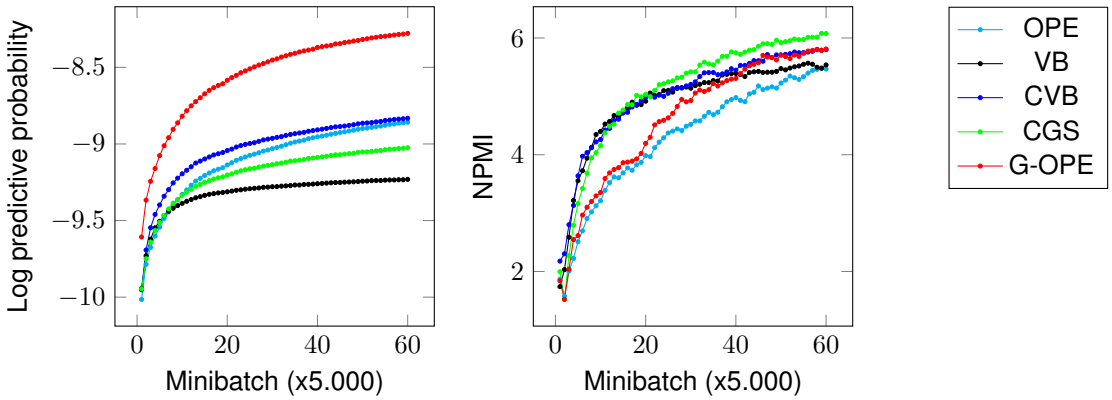

Fig. 5. Online-GOPE compares with Online-OPE, Online-VB, Online-CVB and Online-CGS on Pubmed dataset. Higher is better

rate of original OPE algorithm, which is the most important character among existing state-of-the-art inference methods. In practice, the parameter p of Bernoulli distribution in our method is a flexible way to deal with different datasets.

Besides, the spiritual idea in building approximation functions from G-OPE can be easily extended to a wide class of maximum a posteriori estimation or non-convex problems. By exploiting G-OPE carefully, we have derived an efficient method Online-GOPE for learning LDA from data streams or large corpora. As a result, it is the good candidate to help us to work with text streams and big data.

6 Predictive Probability

Predictive Probability shows the predictiveness and generalization of a model M on new data.

We followed the procedure in [12] to compute this measurement. For each document in a testing dataset, we divided randomly into two disjoint parts w_{obs} and w_{ho} with a ratio of 80:20. We next did inference for w_{obs} to get an estimate of $\mathbb{E}(\theta^{obs})$. Then we approximated the predictive probability as:

$$\Pr(w_{ho}|w_{obs},\mathcal{M}) \simeq \prod_{(w\in w_{ho})} \sum_{k=1}^{K} \mathbb{E}(\theta_k^{obs})\mathbb{E}(\beta_{kw}),$$

$$LogPredictiveProbability = \log \frac{\Pr(w_{ho}|w_{obs}, \mathcal{M})}{|w_{ho}|},$$

where $\mathcal{M}$ is the model to be measured. We estimated $\mathbb{E}(\beta_k) \propto \lambda_k$ for the learning methods which maintain a variational distribution (λ) over topics. Log Predictive Probability was averaged from 5 random splits, each was on 1000 documents.

7 NPMI

NPMI measurement helps us to see the coherence or semantic quality of individual topics. According to [11], NPMI agrees well with human evaluation on interpretability of topic models. For each topic t, we take the set $\{w_1, w_2, \ldots, w_n\}$ of top n terms with highest probabilities. We then computed:

$$NPMI(t) = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_j, w_i)}{P(w_j)P(w_i)}}{-\log P(w_j, w_i)},$$

where $P(w_i, w_j)$ is the probability that terms w_i and w_j appear together in a document. We estimated those probabilities from the training data. In our experiments, we chose top $n = 10$ terms for each topic. Overall, NPMI of a model with K topics is averaged as:

$$NPMI = \frac{1}{K} \sum_{t=1}^{K} NPMI(t).$$

Acknowledgements

This research is funded by Thai Nguyen University of Information and Communication Technology (ICTU) under grant number T2018-07-01.

References

1. **Aletras, N. & Stevenson, M. (2013).** Evaluating topic coherence using distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics*, pp. 13–22.
2. **Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009).** On smoothing and inference for topic models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 27–34.
3. **Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003).** Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022.
4. **Bouma, G. (2009).** Normalized (pointwise) mutual information in collocation extraction. *German Society for Computational Linguistics & Language Technology*, pp. 31–40.
5. **Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977).** Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B*, pp. 1–38.
6. **Feller, W. (1943).** The general form of the so-called law of the iterated logarithm. *Transactions of the American Mathematical Society*, Vol. 54, No. 3, pp. 373–402.
7. **Frank, M. & Wolfe, P. (1956).** An algorithm for quadratic programming. *Naval Research Logistics*, Vol. 3, No. 1-2, pp. 95–110.
8. **Griffiths, T. L. & Steyvers, M. (2004).** Finding scientific topics. *Proceedings of the National academy of Sciences*, volume 101, National Acad Sciences, pp. 5228–5235.
9. **Hazan, E. & Kale, S. (2012).** Projection-free online learning. *Proceedings of the 29th International Conference on International Conference on Machine Learning*, Omnipress, pp. 1843–1850.
10. **Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. W. (2013).** Stochastic variational inference. *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 1303–1347.
11. **Lau, J. H., Newman, D., & Baldwin, T. (2014).** Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539.

12. **Mimno, D., Hoffman, M. D., & Blei, D. M. (2012).** Sparse stochastic inference for latent dirichlet allocation. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1515–1522.
13. **Reddi, S. J., Sra, S., Póczos, B., & J.Smola, A. (2016).** Stochastic frank-wolfe methods for non-convex optimization. *Proceedings of 54th Annual Allerton Conference on Communication, Control, and Computing*, IEEE, pp. 1244–1251.
14. **Sontag, D. & Roy, D. (2011).** Complexity of inference in latent dirichlet allocation. *Advances in neural information processing systems*, pp. 1008–1016.
15. **Teh, Y. W., Kurihara, K., & Welling, M. (2007).** Collapsed variational inference for hdp. *Proceedings of Advances in Neural Information Processing Systems*, pp. 1481–1488.
16. **Than, K. & Doan, T. (2015).** Guaranteed inference in topic models. *arXiv preprint arXiv:1512.03308*.

Article received on 17/12/2017; accepted on 15/02/2018.
Corresponding author is Xuan Bui.