Some Methods for Posterior Inference in Topic Models

Xuan Bui[1,2], Tu Vu[1], Khoat Than[1]

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

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

Correspondence: Xuan Bui, thanhxuan1581@gmail.com

Communication: received 27 February 2018, revised 10 July 2018, accepted 8 August 2018

Online early access: 8 November 2018, Digital Object Identifier: 10.32913/rd-ict.vol2.no15.687

The Area Editor coordinating the review of this article and deciding to accept it was Dr. Trinh Quoc Anh

Abstract: The problem of posterior inference for individual documents is particularly important in topic models. However, it is often intractable in practice. Many existing methods for posterior inference such as variational Bayes, collapsed variational Bayes and collapsed Gibbs sampling do not have any guarantee on either quality or rate of convergence. The online maximum a posteriori estimation (OPE) algorithm has more attractive properties than other inference approaches. In this paper, we introduced four algorithms to improve OPE (namely, OPE1, OPE2, OPE3, and OPE4) by combining two stochastic bounds. Our new algorithms not only preserve the key advantages of OPE but also can sometimes perform significantly better than OPE. These algorithms were employed to develop new effective methods for learning topic models from massive/streaming text collections. Empirical results show that our approaches were often more efficient than the state-of-the-art methods.

Keywords: *Topic models, posterior inference, online maximum a posteriori estimation (OPE), large-scale learning.*

I. Introduction

Topic modeling provides a framework to model high-dimensional sparse data. It can also be seen as an unsupervised learning approach in machine learning. One of the most famous topic models, latent Dirichlet allocation (LDA) [1], has been successfully applied in a wide range of areas including text modeling [2], bioinformatics [3, 4], history [5–7], politics [2, 8], and psychology [9].

Originally, LDA is applied to model a corpus of text documents in which each document is assumed as a random mixture of topics and a topic is a distribution over words. The learning problem is finding the topic distribution of each document and the distribution of words in topics. When learning these parameters, we have to deal with an inference step which is to find the topic distribution of a document with the known distributions of words in topics. Inference problem is, in essence, estimating posterior distributions for individual documents and it is the core problem in LDA. This problem is considered by many researchers in recent years and various learning algorithms such as variational Bayes (VB) [1, 10, 11], collapsed variational Bayes (CVB) [12, 13], CVB0 [14] and collapsed Gibbs sampling (CGS) [7, 15], online maximum a posteriori estimation (OPE) [16], BP-sLDA [17] have been proposed. Inference can be formulated as an optimization problem, ideally, it is a convex optimization. However, the convexity is controlled by a prior parameter which leads to a non-convex problem in practice. Also, it has been proved that the inference problem is NP-hard, hence it is intractable [18]. Among mentioned methods, only OPE has a theoretical guarantee on fast convergence. We investigate the operation of OPE and enhance OPE in terms of different quality measures.

The main contributions of our paper are as follows. First, we investigate the operation of OPE, figure out basic features, and use them to propose new algorithms which are called OPE1, OPE2, OPE3, and OPE4. Those algorithms are derived from combining the upper and lower stochastic bounds of the true objective function. Second, we introduce new methods for learning LDA from text data. From extensive experiments on two large corpora, New York Times and PubMed, we find that some of our methods can achieve high performance in several important measurements usually used in topic models. Third, our ideas of combining the upper and lower stochastic bounds to solve a non-convex inference problem is novel. It has shown effectiveness in topic modeling. Therefore, we believe that this idea can be used in various situations to deal with non-convex optimization.

The paper is organized into six sections. Section II reviews related works and background. Section III ex-

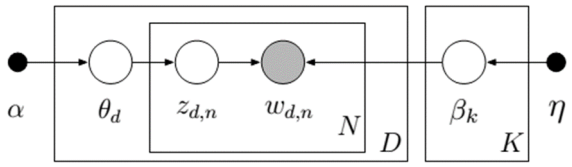


Figure 1. LDA, represented as a graphical model.

plicitly describes our proposed approaches. Experimental results are discussed in Section IV. Section V shows the convergence of our new algorithms and conclusion is in Section VI.

Notations: Throughout the paper, we use the following conventions and notations. Bold faces denote vectors or matrices, x_i the i-th element of vector $\boldsymbol{x}$, and A_{ij} 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 work with text collections of $\boldsymbol{V}$ dimensions (dictionary size). Each document $\boldsymbol{d}$ is represented as a frequency vector, $\boldsymbol{d} = (d_1, .., d_V)^T$, where d_j represents the frequency of the 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 the expectation of the random variable $\boldsymbol{X}$.

II. Posterior Inference

LDA is a generative model for modeling texts and discrete data. It assumes that a corpus is composed from $\boldsymbol{K}$ topics, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, each of which is a sample from a $\boldsymbol{V}$-dimensional Dirichlet distribution, $Dirichlet(\eta)$. Each document $\boldsymbol{d}$ is a mixture of those topics and is assumed to arise from the following generative process:

1) Draw $\boldsymbol{\theta}_d | \alpha \sim Dirichlet(\alpha)$.
2) For the n-th word of $\boldsymbol{d}$,
   - draw topic index $z_{dn} | \boldsymbol{\theta}_d \sim Multinomial(\boldsymbol{\theta}_d)$,
   - draw word $w_{dn} | z_{dn}, \boldsymbol{\beta} \sim Multinomial(\boldsymbol{\beta}_{z_{dn}})$.

Each topic mixture $\boldsymbol{\theta}_d = (\theta_1, \ldots, \theta_K)$ represents the contributions of topics to document $\boldsymbol{d}$, $\theta_k = \Pr(z = k | \boldsymbol{d})$, while $\beta_{kj} = \Pr(w = j | z = k)$ shows the contribution of term j to topic k. Note that $\boldsymbol{\theta}_d \in \Delta_K, \boldsymbol{\beta_k} \in \Delta_V, \forall k$. $\boldsymbol{\theta}_d$ and $\boldsymbol{z}_d$ are respectively hidden and local variables for each document $\boldsymbol{d}$. LDA further assumes that $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are samples of Dirichlet distributions, more specifically, $\boldsymbol{\theta}_d \sim Dirichlet(\alpha)$ and $\boldsymbol{\beta}_k \sim Dirichlet(\eta)$.

The problem of posterior inference for each document $\boldsymbol{d}$, given a model $\{\boldsymbol{\beta}, \alpha\}$, is to estimate the full joint distribution $\Pr(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{d} | \boldsymbol{\beta}, \alpha)$. Direct estimation of this distribution is an NP-hard in the worst case [18]. Existing inference approaches use different schemes. Some methods such as VB, CVB, and CVB0 try to estimate the distribution by maximizing a lower bound of the likelihood $\Pr(\boldsymbol{d} | \boldsymbol{\beta}, \alpha)$, whereas CGS tries to estimate $\Pr(\boldsymbol{z} | \boldsymbol{d}, \boldsymbol{\beta}, \alpha)$. They are being popularly used in topic modeling, but we have not seen any theoretical analysis about how fast they do inference for individual documents.

Other good candidates for posterior inference includes concave-convex procedure (CCCP) [19], stochastic majorization-reduction (SMM) [20], Frank-Wolfe (FW) [21], online Frank-Wolfe (OFW) [22], and threshold linear inverse (TLI) [23]. One might employ CCCP and SMM to do inference in topic models. Those two algorithms are guaranteed to converge to a stationary point of the inference problem. However, the rates of convergence of CCCP and SMM are not clearly analyzed in non-convex circumstances such as inferences in topic models.

We consider the following maximum a posteriori (MAP) estimation of topic mixture for a given document $\boldsymbol{d}$:

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

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

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

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

$$\begin{aligned} \log \Pr(\boldsymbol{d} | \boldsymbol{\theta}, \boldsymbol{\beta}) &= \log \prod_j \Pr(w = j | \boldsymbol{d})^{d_j} \\ &= \sum_j d_j \log \Pr(w = j | \boldsymbol{d}) \\ &= \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}. \end{aligned}$$

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

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

It is shown that this problem is NP-hard in the worst case when $\alpha < 1$ by the authors in [18]. In the case of $\alpha \geq 1$, one can easily show that problem (2) is a concave optimization, and therefore can be solved in polynomial time. Unfortunately, in practice, the parameter α is often small, e.g., $\alpha < 1$, which causes (2) to be a non-concave

Algorithm 1: OPE: Online MAP estimation

Input: document $\mathbf{d}$ 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 $\boldsymbol{\theta}_1$ arbitrarily in Δ_K
for $t = 1, 2, \ldots, \infty$ **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

optimization. In this paper, we consider problem (2) in case the hyper-parameter $\alpha < 1$.

The OPE algorithm for doing inference of topic mixtures for documents was developed by Than and Doan in [16]. Details of OPE are presented in Algorithm 1. The operation of OPE is simple. It solves (2) by iteratively finding a vertex of Δ_K as a direction to the optimal solution. A good vertex at each iteration is decided by assessing the stochastic approximations of the gradient of objective function $f(\boldsymbol{\theta})$. When the number of iterations t goes to infinity, value of $\boldsymbol{\theta}_t$ in OPE will approach a local maximal/stationary point. We also find out that OPE, unlike CCCP and SMM, is guaranteed to converge very fast to a local maximal/stationary point of problem (2).

Each iteration of OPE requires modest arithmetic operations, thus OPE is significantly more efficient than CCCP and SMM. Having a clear guarantee helps OPE to overcome many limitations of VB, CVB, CVB0, and CGS. Furthermore, OPE is so general that it can be easily used and applied in a wide range of contexts, including MAP estimation and non-convex optimization. Therefore, OPE overcomes drawbacks of FW, OFW, and TLI.

III. Characteristics of OPE and New Variants

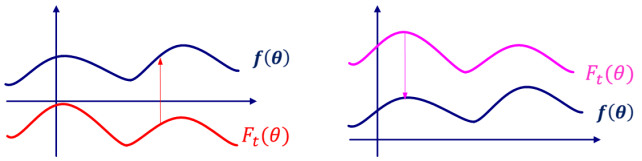
In this section, we figure out more important characteristics of OPE, some were investigated in [16]. OPE can work well with a complex non-convex objective function as follows:

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

Denote

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

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



(a) $F_1(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta})$ (b) $F_1(\boldsymbol{\theta}) = g_2(\boldsymbol{\theta})$

Figure 2. Two cases of initializing stochastic approximating bounds of F_t.

The true objective function $f(\boldsymbol{\theta})$ can be rewritten as

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

We also find that $g_1(\boldsymbol{\theta})$ is concave while $g_2(\boldsymbol{\theta})$ is non-concave when $\alpha < 1$, then $f(\boldsymbol{\theta})$ is non-concave in case $\alpha < 1$.

In general, the optimization theory has encountered many difficulties in solving non-convex optimization problems. Many methods are good in theory but inapplicable in practice. Therefore, instead of directly solving the non-convex optimization with the true objective function $f(\boldsymbol{\theta})$, OPE constructs a sequence of the stochastic functions $F_t(\boldsymbol{\theta})$ that approximates the objective function of interest by uniformly choosing from $\{g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})\}$ in each iteration t. It is guaranteed that F_t converges to f when $t \to \infty$.

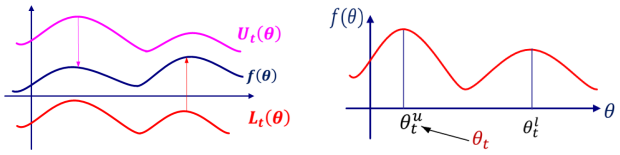
OPE is a stochastic optimization algorithm, can be implemented in a straightforward manner, is computationally efficient and suitable for problems that are large in terms of data and/or parameters. Than and Doan in [16] experimentally and theoretically showed the effectiveness of OPE when applying to the posterior inference of LDA.

By analyzing OPE for more interesting features, we noticed that

$$g_1(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} < 0,$$

$$g_2(\boldsymbol{\theta}) = (\alpha - 1) \sum_{k=1}^{K} \log \theta_k > 0,$$

and $f_t(\boldsymbol{\theta})$ was picked from $\{g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})\}$. Hence, in the first iteration, if we choose $f_1 = g_1$ then $F_1 < f$, which leads the sequence of stochastic functions $F_t(\boldsymbol{\theta})$ approaching $f(\boldsymbol{\theta})$ from below, or it is a lower bound for $f(\boldsymbol{\theta})$. In contrast, if we choose $f_1 = g_2$ in the first iteration, then $F_1 > f$, and the sequence of stochastic functions $F_t(\boldsymbol{\theta})$ approaches $f(\boldsymbol{\theta})$ from above, or it is a upper bound for $f(\boldsymbol{\theta})$ (Figure 2). New perspectives lead us to improvements of OPE. Although OPE is a good candidate for solving posterior inference in topic models, we want to enhance OPE in several different ways. It makes sense that having two stochastic approximating sequences from above and below is better than having one. Therefore, we construct two sequences that both converge to f, one begins



(a) Using upper bound and lower bound of the objective function

(b) Choose the higher point at an iteration

Figure 3. Basic ideas for improving OPE.

Algorithm 2: OPE1: Uniform choice from two stochastic bounds

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 $\boldsymbol{\theta}_1$ arbitrarily in Δ_K

$f_1^l := \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$;

$f_1^u := (\alpha - 1) \sum_{k=1}^K \log \theta_k$

for $t = 2, 3, \ldots \infty$ **do**

Pick f_t^u uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

Pick f_t^l uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

$L_t := \frac{2}{t} \sum_{h=1}^t f_h^l$

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

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

$\boldsymbol{\theta}_{t+1} :=$ pick uniformly from $\{\boldsymbol{\theta}_{t+1}^u, \boldsymbol{\theta}_{t+1}^l\}$

end for

with g_1, called the sequence $\{L_t\}$, and the other begins with g_2, called the sequence $\{U_t\}$ (Figure 3). Using both two stochastic sequences at each iteration gives us more information about the objective function $f(\boldsymbol{\theta})$, so that we can get more chances to reach the maximum of $f(\boldsymbol{\theta})$. In this section, we show four different ideas to improve OPE corresponding to four new algorithms, called OPE1, OPE2, OPE3 and OPE4. Their differences come from the way we combine two approximating sequences $\{U_t\}$ and $\{L_t\}$.

In designing OPE1, we construct two stochastic sequences $\{U_t(\boldsymbol{\theta})\}$ and $\{L_t(\boldsymbol{\theta})\}$ which are similar to $\{F_t(\boldsymbol{\theta})\}$ in OPE. Then, we obtain two sequences $\{\boldsymbol{\theta}_t^u\}$ and $\{\boldsymbol{\theta}_t^l\}$. We pick $\boldsymbol{\theta}_t$ uniformly from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$. OPE1 aims at increasing the randomness of the stochastic algorithm. Getting the idea from a random forest, which constructs a lot of random trees to obtain the average result of all trees, we use randomness to create plenty of choices in our algorithm. We hope that, with full randomness, OPE1 can jump over local stationary points to reach the highest local stationary point.

Algorithm 3: OPE2: Smooth random choice from two stochastic bounds

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 $\boldsymbol{\theta}_1$ arbitrarily in Δ_K

$f_1^l := \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$;

$f_1^u := (\alpha - 1) \sum_{k=1}^K \log \theta_k$

for $t = 2, 3, \ldots \infty$ **do**

Pick f_t^u uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

Pick f_t^l uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

$L_t := \frac{2}{t} \sum_{h=1}^t f_h^l$

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

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

$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_{t+1}^u$ with probability $\frac{\exp f(\boldsymbol{\theta}_{t+1}^u)}{\exp f(\boldsymbol{\theta}_{t+1}^u) + \exp f(\boldsymbol{\theta}_{t+1}^l)}$

and

$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_{t+1}^l$ with probability $\frac{\exp f(\boldsymbol{\theta}_{t+1}^l)}{\exp f(\boldsymbol{\theta}_{t+1}^u) + \exp f(\boldsymbol{\theta}_{t+1}^l)}$

end for

Continuing with the idea of raising the randomness, we pick $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$ with probabilities depending on the value of $\{f(\boldsymbol{\theta}_t^u), f(\boldsymbol{\theta}_t^l)\}$. The higher the value of f is, the higher the probability that the point will be chosen. The probability of selection of $\boldsymbol{\theta}_t$ in OPE2 is smoother than the uniform probability in OPE1. We obtain OPE2 which is detailed in Algorithm 3.

The third idea to improve OPE is based on the greedy approach. We always compare two values of $f(\boldsymbol{\theta}_t^u)$ and $f(\boldsymbol{\theta}_t^l)$ and take the point corresponding to the highest value of f in each iteration (Figure 2). OPE3 works differently from the original OPE. OPE constructs only one sequence $\{\boldsymbol{\theta}_t\}$ while OPE3 creates three sequences $\{\boldsymbol{\theta}_t^u\}$, $\{\boldsymbol{\theta}_t^l\}$, and $\{\boldsymbol{\theta}_t\}$ depending on each other. Even though the structure of the sequence $\{\boldsymbol{\theta}_t\}$ really changes, OPE's good properties remain in OPE3.

Another inference algorithm called OPE4 was proposed. We approximate the true objective function $f(\boldsymbol{\theta})$ by a linear combination of the upper bound U_t and the lower bound L_t with a suitable parameter ν, $F_t := \nu U_t + (1 - \nu) L_t$. The usage of both bounds is stochastic in nature and helps us reduce the possibility of getting stuck at a local stationary point. This is an efficient approach for escaping saddle points in non-convex optimization. Our new variant seems to be more appropriate and robust than the OPE. Existing

Algorithm 4: OPE3: Higher-value choice from stochastic bounds

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 $\boldsymbol{\theta}_1$ arbitrarily in Δ_K

$f_1^l := \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$;

$f_1^u := (\alpha - 1) \sum_{k=1}^K \log \theta_k$

for $t = 2, 3, ...\infty$ **do**

Pick f_t^u uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; \quad (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

Pick f_t^l uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_k j; \ (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

$L_t := \frac{2}{t} \sum_{h=1}^t f_h^l$

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

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

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

end for

Algorithm 5: OPE4: Linear combination of stochastic bounds

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 $\boldsymbol{\theta}_1$ arbitrarily in Δ_K

$f_1^l := \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$;

$f_1^u := (\alpha - 1) \sum_{k=1}^K \log \theta_k$

for $t = 2, 3, ...\infty$ **do**

Pick f_t^u uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; \ (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

Pick f_t^l uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; \quad (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

$L_t := \frac{2}{t} \sum_{h=1}^t f_h^l$

$F_t := \nu . U_t + (1 - \nu) L_t$

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

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

end for

methods become less relevant in high dimensional non-convex optimization. The theoretical justification of OPE4 is motivated by ensuring rapid escape from saddle points.

Similar to OPE, OPE4 constructs the sequence $\{\boldsymbol{\theta}_t\}$ converging to $\boldsymbol{\theta}^*$. OPE4 also aims at increasing the randomness, but it works differently compared to OPE. While OPE constructs only one sequence of function F_t, OPE4 constructs three sequences U_t, L_t, and F_t, in which F_t depends on U_t and L_t. Therefore, the structure of the main sequence F_t is actually changed.

One can recognize that our new algorithms double the computation of OPE at each iteration. However, the rates of convergence of OPE3 and OPE4 remain the same as of OPE as analyzed in the next section. That means, our new algorithms still preserve the key features of OPE.

IV. Experiments

In this section, we investigate the practical performance of our new variants. Since OPE, OPE1, OPE2, OPE3, and OPE4 can play the role as the core subroutine of large-scale learning methods for LDA, we will investigate the performance of these inference algorithms through ML-OPE and Online-OPE [24] by replacing their inference core. We also see how helpful our new algorithms for posterior inference are. Replacing OPE by our new variants in ML-OPE and Online-OPE, we obtain eight new algorithms for learning LDA, called ML-OPE1, Online-OPE1, ML-OPE2, Online-OPE2, ML-OPE3, Online-OPE3, ML-OPE4, and Online-OPE4. Our results provide comparisons between OPE and these four new variants of OPE.

TABLE I
Datasets for Experiment

Data set	No.docs	No.terms	No.doc train	No.doc test
New York Times	300,000	141,444	290,000	10,000
PubMed	330,000	100,000	320,000	10,000

1. Datasets

We used the two large corpora as shown in Table I. The PubMed dataset consists of 330,000 articles from the PubMed Central and the New York Times (NYT) dataset consists of 300,000 news pieces[1]. Each of the learning methods are run five times on each dataset and average results are reported.

2. Parameter Settings

To compare our new methods with OPE, all free parameters receive the same values as in [16]. Below are parameter settings:

- **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.

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

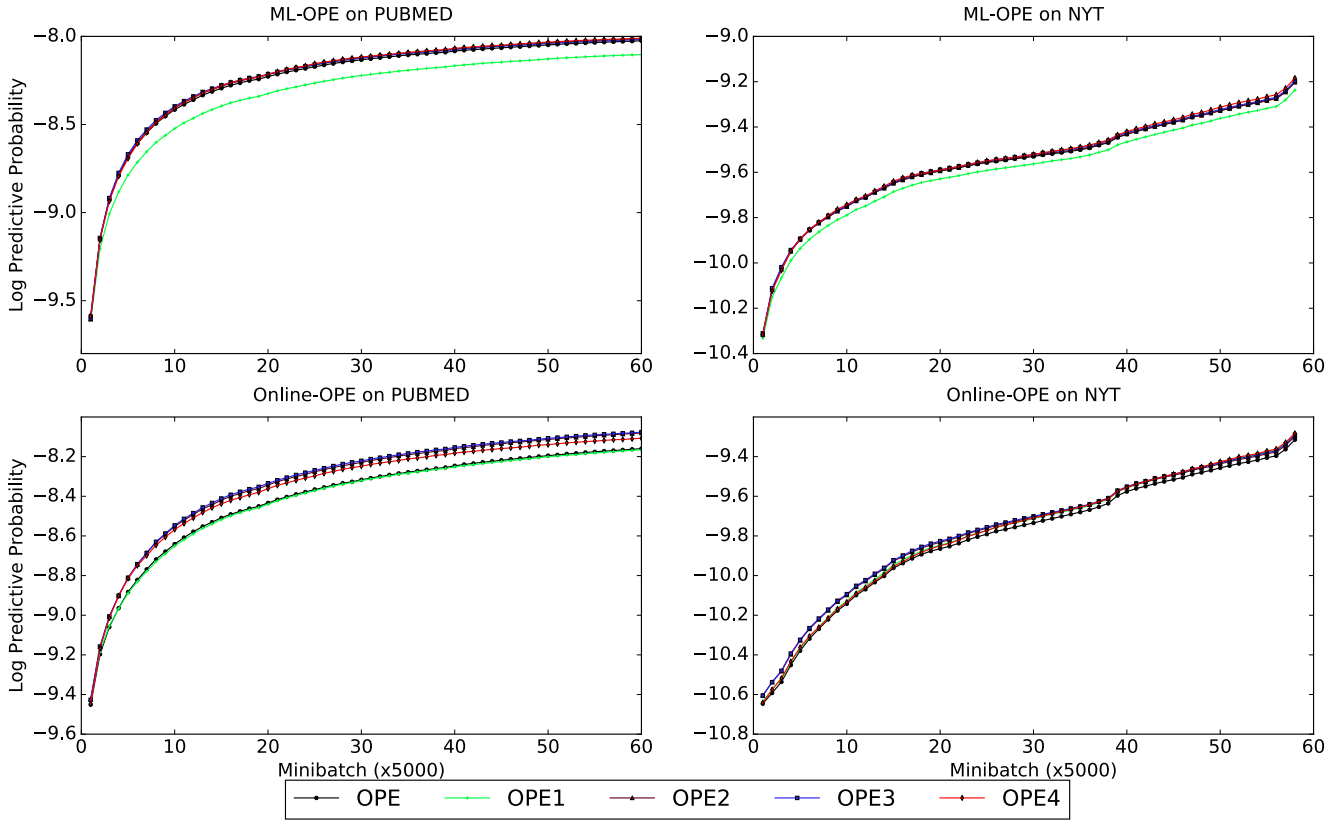


Figure 4. Results of new algorithms compared with the OPE. It can be seen that some new algorithms still have better performance than that of OPE.

- **Inference parameters:** The number of iterations was chosen as $T = 20$.
- **Learning parameters:** Mini-batch size $S = |C_t| = 5000$. $\kappa = 0.9$ and $\tau = 1$ adapted best for existing inference methods. The best value for parameter ν in OPE4 was selected from $\{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$ for each experiment.

3. Evaluation Measures

We used two measures: Predictive Probability [7] and NPMI [25]. Predictive probability measures the predictability and generalization of a model to new data, while NPMI evaluates semantic quality of an individual topic. Details of the measures are presented in Appendix A and B.

4. Evaluation Results

Figure 4 and Figure 5 present evaluation results. We split the results into two figures corresponding to the measures.

Variants of OPE aim to seek the parameter $\boldsymbol{\theta}$ that maximizes a function $f(\boldsymbol{\theta})$ on a simplex using stochastic bounds. Then its results are used to update parameters of a model. ML-OPE updates the direct model parameter $\boldsymbol{\beta}$ and Online-OPE updates the variational parameter $\boldsymbol{\lambda}$. The quality of the parameter $\boldsymbol{\theta}$ found by OPE affects directly the quality of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$.

In practice, OPE is fast and stable. Stability is shown by the number of iterations T. The predictability level that OPE obtain after 20 iterations ($T = 20$) is the same as after 100 iterations ($T = 100$). That means OPE converges very fast. The authors also [16] did experiments by running OPE for 10 times and observed that obtained results were not different. We show that fewer iterations are needed to yield a useful approximation if the rate of convergence is higher. Improving a fast and stable algorithm is not easy, we can neither increase the number of iterations nor run it many times. We need to change the structure of sequences that OPE uses to maximize the objective function.

Figure 4 shows that OPE1 and OPE2 are working worse than the remaining algorithms. The way OPE1 and OPE2 work does not increase the randomness of the approximation. At each iteration, both OPE1 and OPE2 randomly choose one of the two values in $\{\boldsymbol{\theta}^u, \boldsymbol{\theta}^l\}$. Thus, for many consecutive iterations, we may have selected the values of $\boldsymbol{\theta}$ which actually make the objective function f decrease. OPE3 overcomes this problem. OPE3 selects the point $\boldsymbol{\theta}$

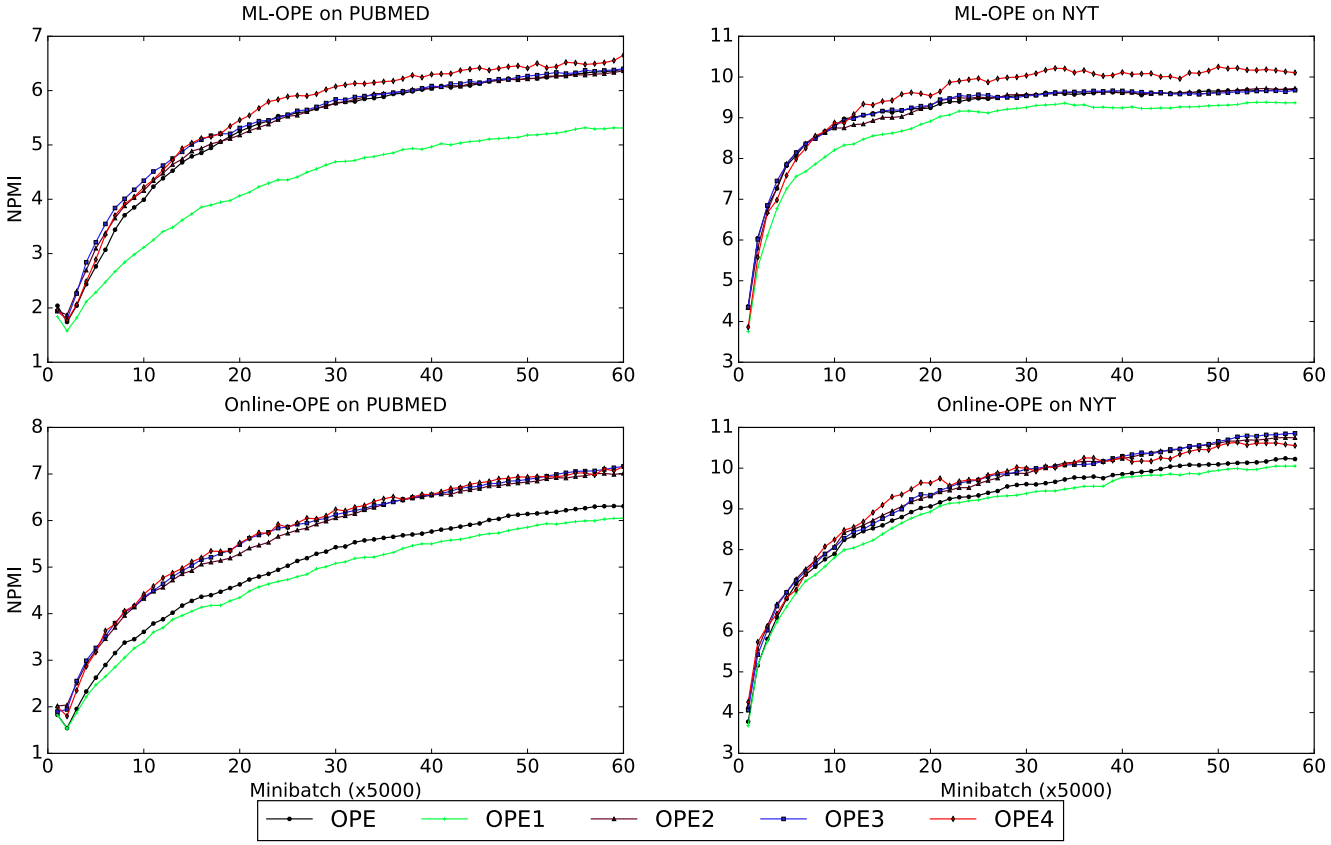


Figure 5. Results of new algorithms when compared to OPE on the NPMI measure. It can be seen that the new algorithms are as good as or even better than OPE.

such that it always increases the value of the objective function f. Therefore, the quality of the learned parameter $\boldsymbol{\theta}$ is better, then the quality of the parameter $\boldsymbol{\beta}$ is better. Notice that the log predictive probability obtained by OPE3 is higher than corresponding results of OPE1 or OPE2. Similar to OPE3, OPE4 with a fit parameter ν obtains good result. Although there are differences in results between methods, the differences are very small. Therefore, in this case the log predictive probability does not reflect well the effectiveness of the improvements. Because the log predictive probability depends on the quality of the parameter $\boldsymbol{\beta}$ of ML-OPE and Online-OPE and it demonstrates that the quality of the parameter $\boldsymbol{\theta}$ is not improved after the inference process.

NPMI reveals evidently the quality of the parameter $\boldsymbol{\theta}$ learned through five algorithms. Figure 5 shows that NPMI is significantly improved by these new OPE variants.

We find out that OPE1 obtains the poorest result. OPE2 and OPE3 are better than OPE. And OPE4 shows the best results. The idea of OPE2 comes from the combination of OPE1 and OPE3 (OPE2 is a hybrid algorithm combining OPE1 and OPE3). OPE2 chooses the parameter $\boldsymbol{\theta}$ with a probability depending on the value of the function $f(\boldsymbol{\theta})$ at two bounds (the higher the value of the function $f(\boldsymbol{\theta})$ at a point is, the higher the probability at that point is). Thus, OPE3 is the same as OPE2 when the probability at the upper bound is 1 and the probability at the lower bound is 0. NPMI is computed directly from the learned parameter $\boldsymbol{\theta}$. It is easy to notice that the quality of the parameter $\boldsymbol{\theta}$ is significantly improved with the construction of a new approximation of the function f from OPE2 and OPE3. OPE4 is shown to be more effective when the best parameter ν is chosen. The parameter $\boldsymbol{\theta}$ is appropriately chosen in our experiments, then OPE4 is more complex than other algorithms. By adding the appropriate parameter ν, we have increased the quality of the model because, in machine learning theory, the more complex the model is, the higher the accuracy it achieves.

It is easy to see that OPE3 makes ML-OPE3 and Online-OPE3 become more efficient. OPE3 demonstrates our idea of using two random sequences of functions to approximate the objective function $f(\boldsymbol{\theta})$. The idea of increasing the randomness and the greedy of the algorithm is exploited here. Firstly, two random sequences of function are used to raise our participants and information relevant to the

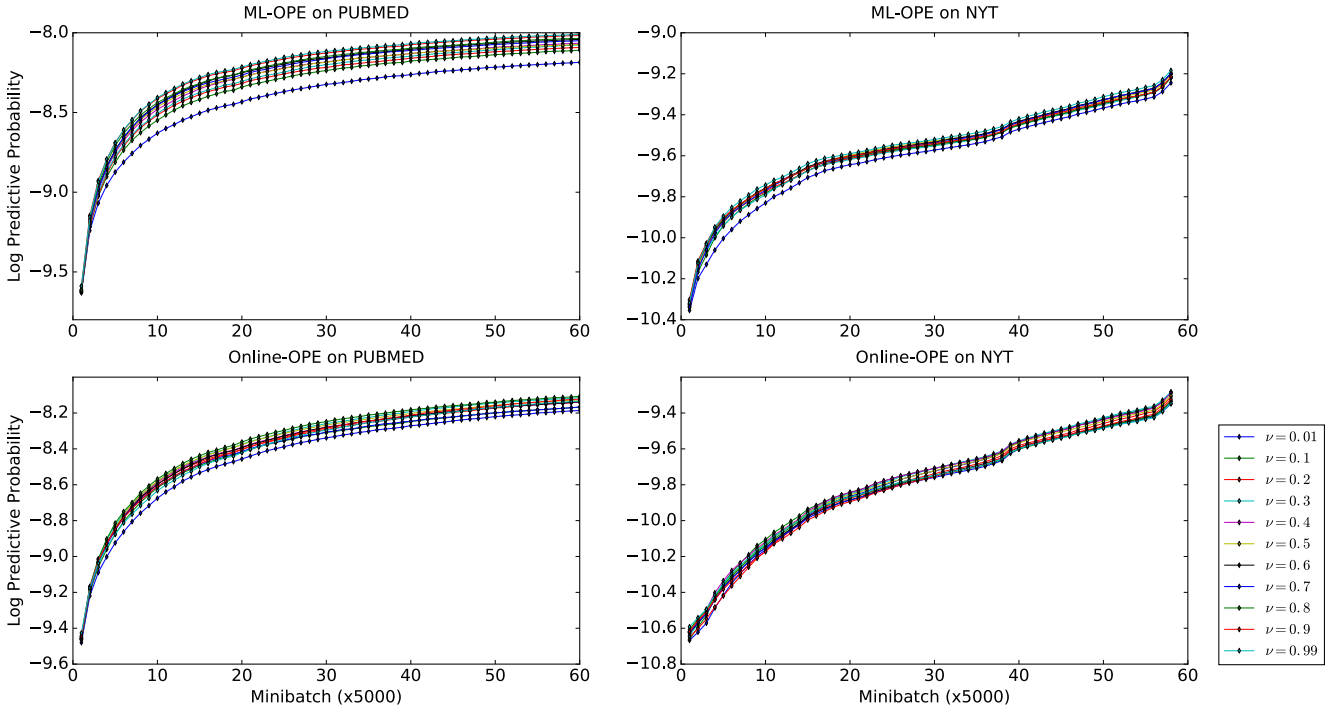


Figure 6. OPE4 with different values of ν using the Predictive Probability measure.

objective function. Hence, in the next iteration, we have more choices in $\boldsymbol{\theta}_t$. Secondly, choosing $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$ makes the value of $f(\boldsymbol{\theta})$ higher after each iteration, that comes from the idea of greedy algorithms. It maybe the 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 parameters.

In the experiment with OPE4, we introduce the parameter ν to construct $F_t(\boldsymbol{\theta}_t) = \nu U_t(\boldsymbol{\theta}_t) + (1-\nu)L_t(\boldsymbol{\theta}_t)$. Therefore, we increase the number of parameters in the model and we have to choose the parameter ν empirically. The parameter ν that we used for each dataset is usually 0.01 or 0.99, that means the stochastic bounds always follow one direction below or above. OPE4 uses a linear combination of the upper bound U_t and the lower bound L_t. The bounds U_t and L_t converge to the objective function f, so the linear combination F_t improves the convergence speed and the quality of the approximation.

OPE4 is the simplest way to combine the bounds. We can utilize OPE4 to invent more complicated combinations which may result in better approximations. Besides, OPE4 can be expanded by using not only two but also many stochastic bounds to approximate an objective function, which is an open approach to investigate. We notice that, with both measures, OPE3 and OPE4 are better than OPE1 and OPE2, especially when using the NPMI measure.

By changing variables and bound functions, we obtain two new algorithms (OPE3 and OPE4) that are more effective than OPE. We show that our approach outperforms the state-of-the-art approaches of posterior inference in LDA.

TABLE II
THE BEST VALUE OF ν CHOSEN WITH THE TWO DATASETS VIA THE TWO MEASURES.

Method	Measure	New York Times	Pubmed
ML-OPE4	LPP	$\nu = 0.6$	$\nu = 0.99$
ML-OPE4	NPMI	$\nu = 0.4$	$\nu = 0.99$
Online-OPE4	LPP	$\nu = 0.3$	$\nu = 0.8$
Online-OPE4	NPMI	$\nu = 0.5$	$\nu = 0.9$

5. Effect of Parameter ν in OPE4

In Section II, we find out that OPE has many more good characteristics than existing algorithms. The above experiments showed that OPE3 and OPE4 outperform OPE. Especially, we find that OPE4 is the most efficient for almost all datasets. However, the effectiveness of OPE4 depends on how the parameter ν is chosen. To see the effect of the parameter ν, we run the algorithm with different values of ν from the set $\{0.01, 0.1, 0.2...0.9, 0.99\}$, because $0 < \nu < 1$, while the other parameters are fixed (See Figure 6 and Figure 7).

We show some results obtained by running OPE4 with different values of ν between 0 and 1. From Figure 6 and Figure 7, the best values for ν are close to either 1 or 0.5. Details are presented in Table II.

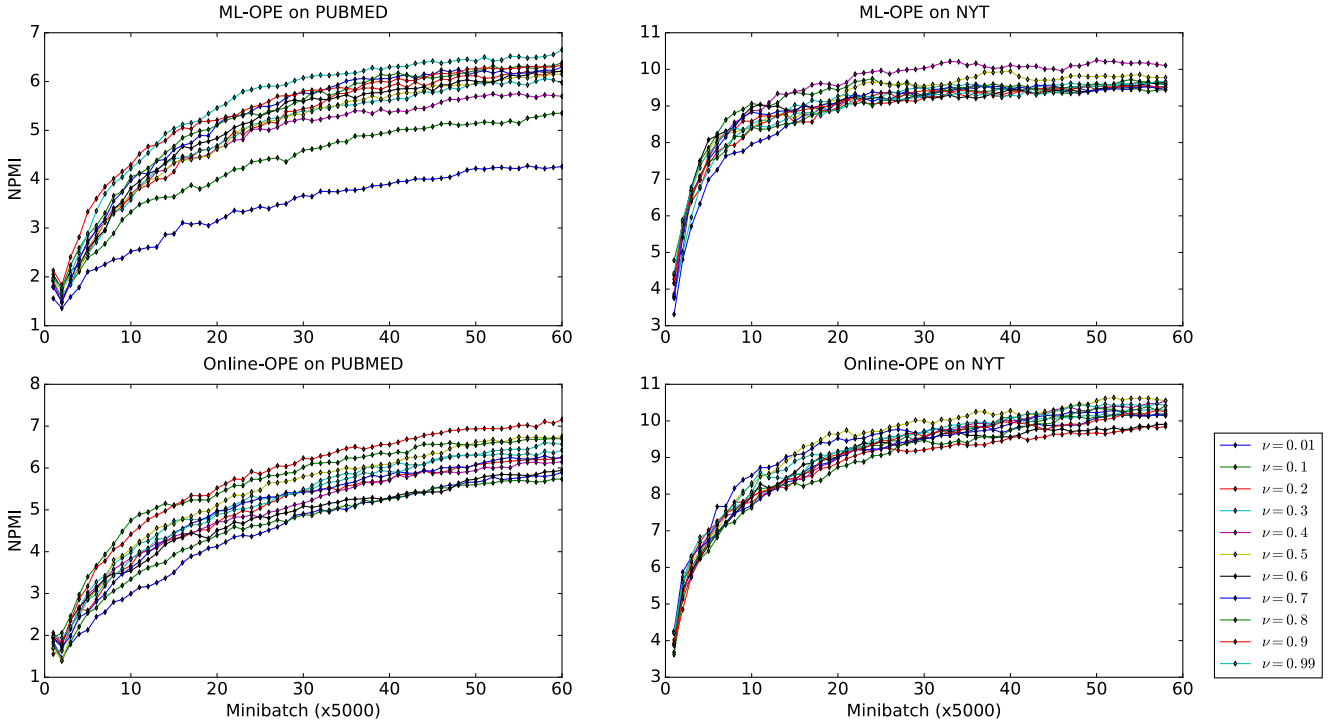


Figure 7. OPE4 with different values of ν using the NPMI measure.

OPE4 works efficiently when using the upper bound, the lower bound, or the average of the two bounds. We suppose that, when the parameter ν is close to either 0 or 1, OPE4 works like OPE. The best value of the parameter ν is calculated from experimental data. By finding the best value of the parameter ν, OPE4 performs better than OPE, but the trade-off is the extra running time needed to find the best value of ν. This step is neccessary, because inappropriate choices of ν might significantly affect the performance of OPE4.

V. Analysis of Convergence

From extensive experiments, we find that OPE3 and OPE4 are more efficient than OPE on the two datasets when applied in two learning methods for LDA. Therefore, we focused on the convergence of OPE3 and OPE4 algorithms.

Theorem 1 (Convergence of OPE3): Consider the objective function $f(\boldsymbol{\theta})$ in problem (2), given fixed $\boldsymbol{d}$, $\boldsymbol{\beta}$, and α. For OPE3, with probability of 1, the following holds:

1) For any $\boldsymbol{\theta} \in \Delta_K$, $U_t(\boldsymbol{\theta})$ and $L_t(\boldsymbol{\theta})$ converge to $f(\boldsymbol{\theta})$ as $t \to +\infty$,
2) $\boldsymbol{\theta}_t$ converges to a local maximal/stationary point of $f(\boldsymbol{\theta})$.

Proof: The objective function $f(\boldsymbol{\theta})$ is a non-convex. The criterion used for the convergence analysis is important in non-convex optimization. For unconstrained optimization problems, the gradient norm $\|\nabla f(\boldsymbol{\theta})\|$ is typically used to measure the convergence, because $\|\nabla f(\boldsymbol{\theta})\| \to 0$ captures the convergence to a stationary point. However, this criterion can not be used for constrained optimization problems. Instead, we use the "Frank-Wolfe gap" criterion in [26].

Denote

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

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

Firstly, we consider the sequence $\{U_t\}$. Let a_t and b_t respectively be the number of times that we have already picked g_1 and g_2 after t iterations to construct $\{U_t\}$.

Note that $a_t + b_t = t$. Denote $S_t = a_t - b_t$. We have

$$U_t = \frac{2}{t}(a_t g_1 + b_t g_2), \tag{3}$$

$$U_t - f = \frac{S_t}{t}(g_1 - g_2), \tag{4}$$

$$U'_t - f' = \frac{S_t}{t}(g'_1 - g'_2). \tag{5}$$

Since f_t^u is chosen uniformly from $\{g_1,\ g_2\}$ then

$$\begin{aligned}\boldsymbol{E}(f_t^u) &= \frac{1}{2}g_1 + \frac{1}{2}g_2 = \frac{1}{2}f,\\ \boldsymbol{E}(U_t) &= \boldsymbol{E}(\frac{2}{t}\sum_{h=1}^{t} f_h^u) = \frac{2}{t}\sum_{h=1}^{t}\boldsymbol{E}(f_h^u) = \frac{2}{t}\sum_{h=1}^{t}\frac{1}{2}f\\ &= \frac{2}{t}\cdot\frac{t}{2}f = f. \end{aligned} \tag{6}$$

So $U_t(\boldsymbol{\theta})$ is an unbiased estimation of $f(\boldsymbol{\theta})$.

For each iteration t of OPE3 we have to pick uniformly randomly an f_t^u from $\{g_1,\ g_2\}$. We make a correspondence between f_t^u and a uniformly random variable X_t on $\{1, -1\}$. This correspondence is an one-to-one mapping. So $S_t = a_t - b_t$ can be represented as $S_t = X_1 + \cdots + X_t$.

Applying the iterated logarithm in [27] we have $S_t = O(\sqrt{t\log t})$, suggesting $\frac{S_t}{t} \to 0$ as $t \to +\infty$. Combining this with (4), we conclude that the sequence $U_t \to f$ with the probability of 1. Also, due to (5), the derivative sequence $U_t' \to f'$ as $t \to +\infty$. The convergence holds for any $\boldsymbol{\theta} \in \overline{\Delta}_K$.

Consider

$$\begin{aligned}&\langle U_t'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}\rangle =\\ &= \langle U_t'(\boldsymbol{\theta}_t) - f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}\rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}\rangle\\ &= \frac{S_t}{t^2}\langle g_1^{'}(\boldsymbol{\theta}_t) - g_2^{'}(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}\rangle.\end{aligned}$$

Note that g_1 and g_2 are Lipschitz continuous on $\overline{\Delta}_K$. Hence there exists a constant L such that

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

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

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

$$f(\boldsymbol{\theta}_{t+1}^u) \le f(\boldsymbol{\theta}_{t+1}).$$

Since $\boldsymbol{e}_t^u$ and $\boldsymbol{\theta}_t$ belong to Δ_K, the quantity $|\langle g_1'(\boldsymbol{\theta}_t) - g_2'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle|$ and $\|\boldsymbol{e}_t^u - \boldsymbol{\theta}_t\|^2$ are upper-bounded for any t. Therefore, there exists a constant $c_1 > 0$ such that

$$\langle U_t'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t}\rangle \le c_1\frac{|S_t|}{t^2} + f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t) + \frac{c_1 L}{t^2}. \tag{7}$$

Summing both sides of (7) for all t, we have

$$\begin{aligned}&\sum_{t=1}^{+\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle\\ &\quad\le \sum_{t=1}^{+\infty} c_1\frac{|S_t|}{t^2} + f(\boldsymbol{\theta}_{+\infty}) - f(\boldsymbol{\theta}_1) + \sum_{t=1}^{+\infty}\frac{c_1 L}{t^2}.\end{aligned} \tag{8}$$

Because $f(\boldsymbol{\theta})$ is bounded then $f(\boldsymbol{\theta}_{+\infty})$ is bounded.

Note that $S_t = O(\sqrt{t\log t})$ [27], hence $\sum_{t=1}^{+\infty} c_1\frac{|S_t|}{t^2}$ converges with the probability of 1 and $\sum_{t=1}^{+\infty}\frac{L}{t^2}$ is also bounded. Therefore, the right-hand side of (8) is finite.

In addition, $\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u\rangle > \langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t\rangle$ for any $t > 0$ because of $e_t^u = \arg\max_{x\in\overline{\Delta}_K}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{x}\rangle$. Therefore, we obtain the following:

$$0 \le \sum_{t=1}^{+\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle < \infty. \tag{9}$$

In other words, the series $\sum_{t=1}^{+\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle$ converges to a finite constant. Note that $\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle \ge 0$ for any t. If there exists a constant $c_2 > 0$ satisfying $\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle \ge c_2$ for an infinite number of t's, then the series $\sum_{t=1}^{+\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle$ could not converge to a finite constant, which is in contrary to (9). Therefore,

$$\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle \to 0\ as\ t \to +\infty. \tag{10}$$

Because of $U_t' \to f'$ as $t \to \infty$ and U_t', f' are continuous, combining with (10) we have

$$\langle f'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u - \boldsymbol{\theta}_t\rangle \to 0\ as\ t \to +\infty. \tag{11}$$

Using the "Frank-Wolfe gap" criterion in [26], from (11) we have $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ as $t \to +\infty$. In other words, $\boldsymbol{\theta}_t$ converges in term of the probability to a stationary point $\boldsymbol{\theta}^*$ of $f(\boldsymbol{\theta})$. □

Theorem 2 (Convergence of OPE4): Consider the objective function $f(\boldsymbol{\theta})$ in problem (2), given fixed $\mathbf{d}$, $\boldsymbol{\beta}$, α. For OPE4, with probability of 1, the following holds:

1) For any $\boldsymbol{\theta} \in \Delta_K$, $F_t(\boldsymbol{\theta})$ converges to $f(\boldsymbol{\theta})$ as $t \to +\infty$,
2) $\boldsymbol{\theta}_t$ converges to a local maximal/stationary point of $f(\boldsymbol{\theta})$.

Proof: Denote

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

Let a_t and b_t respectively be the number of times that we have already picked g_1 and g_2 after t iterations to construct U_t. Similarly, let c_t and d_t respectively be the number of times that we have already picked g_1 and g_2 after t iterations to construct L_t.

Since f_t^u is chosen uniformly from $\{g_1, g_2\}$ then

$$\boldsymbol{E}(f_t^u) = E(f_t^l) = \frac{1}{2}g_1 + \frac{1}{2}g_2 = \frac{1}{2}f,$$
$$\boldsymbol{E}(U_t) = \boldsymbol{E}(\frac{2}{t}\sum_{h=1}^{t} f_h^u) = \frac{2}{t}\sum_{h=1}^{t}\boldsymbol{E}(f_h^u) = \frac{2}{t}\sum_{h=1}^{t}\frac{1}{2}f = f,$$
$$\boldsymbol{E}(L_t) = \boldsymbol{E}(\frac{2}{t}\sum_{h=1}^{t} f_h^l) = \frac{2}{t}\sum_{h=1}^{t}\boldsymbol{E}(f_h^l) = \frac{2}{t}\sum_{h=1}^{t}\frac{1}{2}f = f,$$
$$\boldsymbol{E}(F_t) = \nu\boldsymbol{E}(U_t) + (1-\nu)\boldsymbol{E}(L_t) = \nu f + (1-\nu)f = f.$$

Denote

$$S_t^u = a_t - b_t,$$
$$S_t^l = c_t - d_t,$$
$$S_t = \max\{|S_t^u|, |S_t^l|\}.$$

We have

$$U_t = \frac{2}{t}(a_t g_1 + b_t g_2) \qquad a_t + b_t = t$$
$$L_t = \frac{2}{t}(c_t g_1 + d_t g_2) \qquad c_t + d_t = t$$
$$U_t - f = \frac{S_t^u}{t}(g_1 - g_2) \qquad L_t - f = \frac{S_t^l}{t}(g_1 - g_2)$$
$$U_t' - f' = \frac{S_t^u}{t}(g_1' - g_2') \qquad L_t' - f' = \frac{S_t^l}{t}(g_1' - g_2').$$

We obtain

$$F_t = \nu U_t + (1-\nu)L_t$$
$$F_t - f = \nu(U_t - f) + (1-\nu)(L_t - f)$$
$$= (\nu\frac{S_t^u}{t} + (1-\nu)\frac{S_t^l}{t})(g_1 - g_2)$$
$$F_t' - f' = (\nu\frac{S_t^u}{t} + (1-\nu)\frac{S_t^l}{t})(g_1' - g_2').$$

So F_t is an unbiased estimation of f.

Applying the iterated logarithm in [27] we have $S_t^u = O(\sqrt{t\log t})$ and $S_t^l = O(\sqrt{t\log t})$, suggesting $\frac{S_t^u}{t} \to 0$ and $\frac{S_t^l}{t} \to 0$ as $t \to +\infty$. Hence, we conclude that the sequence $U_t \to f$ and the derivative sequence $U_t' \to f'$ as $t \to +\infty$. Similarly, we have $L_t \to f$ and the derivative sequence $L_t' \to f'$ as $t \to +\infty$.

Consider

$$\langle F_t'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle =$$
$$= \langle F_t'(\boldsymbol{\theta}_t) - f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle =$$
$$= \langle(\nu\frac{S_t^u}{t} + (1-\nu)\frac{S_t^l}{t})(g_1'(\boldsymbol{\theta}_t) - g_2'(\boldsymbol{\theta}_t)), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle +$$
$$+ \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle.$$

Note that g_1 and g_2 are Lipschitz continuous on $\overline{\Delta}_K$. Hence there exists a constant L such that

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

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

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 $\|\boldsymbol{e}_t^u - \boldsymbol{\theta}_t\|^2$ are bounded. Therefore, there exists a constant $c_1 > 0$ such that

$$\langle F_t'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle \le c_1\frac{S_t}{t^2} + f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t) + \frac{c_1 L}{t^2}. \quad (12)$$

Summing both sides of (12) for all t we have

$$\sum_{t=1}^{+\infty}\frac{1}{t}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle$$
$$\le \sum_{t=1}^{+\infty} c_1\frac{S_t}{t^2} + f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_1) + \sum_{t=1}^{+\infty}\frac{c_1 L}{t^2}. \quad (13)$$

Because $f(\boldsymbol{\theta})$ is bounded then $f(\boldsymbol{\theta}^*)$ is bounded.

Note that $S_t = O(\sqrt{t\log t})$ [27], so $\sum_{t=1}^{+\infty} c_1\frac{S_t}{t^2}$ converges with the probability of 1 and $\sum_{t=1}^{+\infty}\frac{L}{t^2}$ is also bounded. Hence, the right-hand side of (13) is finite.

In addition, $\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t\rangle > \langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t\rangle$ for any $t > 0$ because of $e_t = \arg\max_{x\in\overline{\Delta}_K}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{x}\rangle$. Therefore, we obtain the following

$$0 \le \sum_{t=1}^{+\infty}\frac{1}{t}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle < +\infty. \quad (14)$$

In other words, the series $\sum_{t=1}^{+\infty}\frac{1}{t}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle$ converges to a finite constant. Note that $\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \ge 0$ for any t. If there exists a constant $c_3 > 0$ satisfying $\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \ge c_3$ for an infinite number of t's, then the series $\sum_{t=1}^{+\infty}\frac{1}{t}\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle$ could not converge to a finite constant, which is in contrary to (14). Therefore,

$$\langle F_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \to 0 \; as \; t \to +\infty. \quad (15)$$

Because of $F_t' \to f'$ as $t \to \infty$ and F_t', f' are continuous, combining with (15) we have

$$\langle f'(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \to 0 \; as \; t \to +\infty. \quad (16)$$

Using the "Frank-Wolfe gap" criterion in [26], we have $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ as $t \to +\infty$. In other words, $\boldsymbol{\theta}_t$ converges in term of the probability to a stationary point $\boldsymbol{\theta}^*$ of $f(\boldsymbol{\theta})$. □

The above theorems provide theoretical guarantees on the fast convergence for our algorithms.

VI. Conclusion

We have discussed how posterior inference for individual texts in topic models can be done efficiently. We now provide four theoretically justified algorithms (called OPE1, OPE2, OPE3, and OPE4) to deal well with this problem. They all have a theoretical guarantee on fast convergence rate. OPE3 and OPE4 can do inference faster and more effectively in practice, and they can be easily extended to a wide class of probabilistic models. By exploiting four new variants of OPE carefully, we have derived eight efficient methods for learning LDA from data streams or large corpora. As the result, they are good candidates to help us deal with text streams and big data.

Acknowledgement

This research is funded by the Office of Naval Research Global (ONRG), Air Force Office of Scientific Research (AFOSR), and Asian Office of Aerospace Research & Development (AOARD) under Award Numbers N62909-18-1-2072 and 17IOA031.

Appendix A
Predictive Probability

Predictive Probability shows the predictability and generalization of a model $\mathcal{M}$ on new data. We followed the procedure in [7] to compute this measure. For each document in a test dataset, we randomly divided it into two disjoint parts, w_{obs} and w_{ho}, with a ratio of 80:20. Next, we did inference for w_{obs} to get an estimate of $\boldsymbol{E}(\boldsymbol{\theta}^{\text{obs}})$. Then, we approximated the predictive probability as

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

$$\text{Log Predictive Probability} = \log\frac{\Pr(w_{\text{ho}}|w_{\text{obs}},\mathcal{M})}{|w_{\text{ho}}|}$$

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

Appendix B
NPMI

The NPMI measure helps us see the coherence or the semantic quality of individual topics. According to [28], NPMI agrees well with human evaluation on the 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

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

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

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

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] B. Liu, L. Liu, A. Tsykin, G. J. Goodall, J. E. Green, M. Zhu, C. H. Kim, and J. Li, "Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation," *Bioinformatics*, vol. 26, no. 24, pp. 3105–3111, 2010.

[4] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, p. 945, 2000.

[5] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[6] L. Yao, D. Mimno, and A. McCallum, "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*. ACM, 2009, pp. 937–946.

[7] M. Hoffman, D. M. Blei, and D. M. Mimno, "Sparse stochastic inference for latent dirichlet allocation," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. New York, NY, USA: ACM, 2012, pp. 1599–1606.

[8] 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.

[9] H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar, "Toward personality insights from language exploration in social media," in *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[10] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28. PMLR, 2013, pp. 280–288.

[11] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association, to appear*, 2016.

[12] Y. W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for hdp," in *Advances in neural information processing systems*, 2007, pp. 1481–1488.

[13] 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.
[14] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 27–34.
[15] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
[16] K. Than and T. Doan, "Guaranteed inference in topic models," *arXiv preprint arXiv:1512.03308*, 2015.
[17] J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, and L. Deng, "End-to-end learning of lda by mirror-descent back propagation over a deep architecture," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 1765–1773.
[18] D. Sontag and D. Roy, "Complexity of inference in latent dirichlet allocation," in *Neural Information Processing System (NIPS)*, 2011.
[19] A. L. Yuille, A. Rangarajan, and A. Yuille, "The concave-convex procedure (cccp)," *Advances in neural information processing systems*, vol. 2, pp. 1033–1040, 2002.
[20] J. Mairal, "Stochastic majorization-minimization algorithms for large-scale optimization," in *Neural Information Processing System (NIPS)*, 2013.
[21] K. L. Clarkson, "Coresets, sparse greedy approximation, and the frank-wolfe algorithm," *ACM Trans. Algorithms*, vol. 6, no. 4, pp. 63:1–63:30, 2010.
[22] E. Hazan and S. Kale, "Projection-free online learning," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.
[23] S. Arora, R. Ge, F. Koehler, T. Ma, and A. Moitra, "Provable algorithms for inference in topic models," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 2859–2867.
[24] K. Than and T. Doan, "Dual online inference for latent Dirichlet allocation," in *Proceedings of the Sixth Asian Conference on Machine Learning*, D. Phung and H. Li, Eds., vol. 39, 2015, pp. 80–95.
[25] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Association for Computational Linguistics, 2013, pp. 13–22.
[26] S. J. Reddi, S. Sra, B. Póczos, and A. J.Smola, "Stochastic frank-wolfe methods for nonconvex optimization," in *Proceedings of 54th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2016, pp. 1244–1251.
[27] W. Feller, "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, 1943.
[28] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.

Xuan Bui received B.S (2003) from Vietnam National University and M.S (2007) from Thai Nguyen University, Vietnam. She is currently a member of the Data Science Laboratory, within the School of Information and Communication Technology, Hanoi University of Science and Technology. Her research interests include non-convex optimization in machine learning, stochastic optimization, topic model and big data.

Tu Vu received B.S (2016) from Hanoi University of Science and Technology (HUST), Vietnam. He is currently a member of the Data Science Laboratory, within the School of Information and Communication Technology, HUST. His research interests include topic model, stochastic optimization and big data.

Khoat Than is currently the Director of Data Science Laboratory, within the School of Information and Communication Technology, Hanoi University of Science and Technology. He received B.S (2004) from Vietnam National University, M.S (2009) from Hanoi University of Science and Technology, and Ph.D. (2013) from Japan Advanced Institute of Science and Technology. He joins the Program Committees of various leading conferences, including ICML, NIPS, IJCAI, ICLR, PAKDD, ACML. His recent research interests include representation learning, stochastic optimization, topic modeling, dimensionality reduction, large-scale modeling, big data.