Some Methods for Posterior Inference in Topic Models

***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 (VB), collapsed variational Bayes (CVB) and collapsed Gibbs sampling (CGS) do not have any guarantee on neither quality nor rate of convergence. Online Maximum a Posteriori Estimation algorithm (OPE) has more attractive properties than other inference approaches, including theoretical guarantees on convergence rate. In this paper, we introduce four algorithms to improve OPE (so called OPE1, OPE2, OPE3, and OPE4) by combining stochastic bounds. Our new algorithms not only preserve the key advantages of OPE but also can sometimes perform significantly better than OPE. These algorithms have been employed to develop new effective methods for learning topic models from massive/streaming text collections. Empirical results show that our approaches are often more efficient than the state-of-the-art methods.**

***Index Terms*—OPE, Topic models, Posterior Inference, Online MAP Estimation, Large-scale Learning**

I. Introduction

Topic modeling provides a framework to model high-dimensional sparse data. It is also can 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], [6], [7], politics [2], [8], psychology [9].

Originally LDA is applied to model the 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, the estimating posterior distributions for individual document 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], 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 is proved that the inference problem is NP-hard, hence it is intractable [18]. Among mentioned methods, there is only OPE which has a theoretical guarantee on fast convergence. We investigate the operation of OPE and enhance OPE with more quality in different measures.

The main contributions of our paper are:

- 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 bounds of the objective function.
- We introduce new methods for learning LDA from text data. From extensive experiments on two large corpora, we find that some of our methods can reach high performance in important measurements usually used in topic models.
- Our ideas of combining the upper and lower bounds to solve 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.

Organization: The paper is organized into six sections. Section II reviews related work and background. Section III explicitly describes our proposed approaches. Section IV shows the convergence of our new algorithms. Experimental results are discussed in Section V and conclusion is made in Section VI.

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 $\mathbf{A}$. 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 $\overline{\Delta_n}$. 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 $E(x)$ is expectation of random variable x.

II. Posterior Inference

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

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

Each topic mixture $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ represents the contributions of topics to document $\boldsymbol{d}$, while β_{kj} shows the contribution of term *j* to topic *k*. Note that $\boldsymbol{\theta}_d \in \overline{\Delta}_K, \boldsymbol{\beta}_k \in \Delta_V, \forall k$. Both $\boldsymbol{\theta}_d$ and $\boldsymbol{z}_d$ are 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

Algorithm 1 OPE: Online maximum a posteriori estimation

Input: document **d** and model $\{\boldsymbol{\beta}, \alpha\}$
Output: $\boldsymbol{\theta}$ that maximizes $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$

Initialize $\boldsymbol{\theta}_1$ arbitrary in Δ_K
for $t = 1, 2, ...\infty$ **do**
Pick f_t uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$
$F_t := \frac{2}{t} \sum_{h=1}^t f_h$
$\boldsymbol{e}_t := \arg\max_{\boldsymbol{x} \in \Delta_K} < F_t^{'}(\boldsymbol{\theta}_t), \boldsymbol{x} >$
$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}$
end for

$p(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{d} | \boldsymbol{\beta}, \alpha)$. Direct estimation of this distribution is a NP-hard in the worst case [18]. Hence existing inference approaches use different schemes. Some methods such as VB, CVB, CVB0 try to estimate the distribution by maximizing a lower bound of the likelihood $P(\boldsymbol{d}|\boldsymbol{\beta}, \alpha)$, whereas CGS tries to estimate $P(\boldsymbol{z}|\boldsymbol{d}, \boldsymbol{\beta}, \alpha)$. 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 maximum a posterior (MAP) estimation of topic mixture for a given document

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

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

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

It is showed 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 the problem (2) is concave optimization, therefore it can be solved in polynomial time. Unfortunately, in practice LDA, the parameter α is often small, says $\alpha < 1$, which causes (2) to be non-concave optimization.

OPE for doing inference of topic mixtures for documents is developed by Than and Doan in [16]. Details of OPE is 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. A good vertex at each iteration is decided by assessing stochastic approximations to the gradient of objective function $f(\boldsymbol{\theta})$. When the number of iterations goes to infinity, value of $\boldsymbol{\theta}_t$ in OPE will approach to 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).

The rate of convergence of OPE is faster than other methods. Each iteration of OPE requires modest arithmetic operations, then 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. Further, OPE is so general that it can be easily used and applied in a wide range of contexts, including MAP estimation or 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 characters of OPE, some were investigated in [16]. OPE can work well with a complex non-convex objective function such as

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

We can see this as a Difference of Convex (D.C) functions. Denote

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

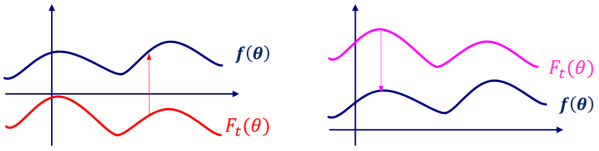
The objective function $f(\boldsymbol{\theta})$ can be rewritten as the difference of two functions as form $f(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta})$. General speaking, the optimization theory has encountered many difficulties in solving the optimization problem with non-convex objective function. Many methods are only good in theory but inapplicable in practice via careful researches. Therefore, instead of directly solving the non-convex optimization with the objective function f, OPE constructs a sequence of stochastic functions F_t that approximates to the objective function of interest by alternatively choosing uniformly from $\{g_1, g_2\}$ in each iteration t, it is guaranteed that F_t converges to f when $t \to \infty$.

OPE is one of stochastic optimization algorithms. OPE is straightforward to implement, computationally efficient and suitable for problems that are large in terms of data and/or parameters. Than and Doan in [16] have experimentally and theoretically showed the effectiveness of OPE when applying to the posterior inference of LDA.

By analyzing OPE with more interesting features, we noticed that:

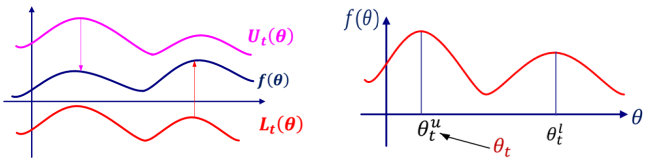
$$g_1(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} < 0\,, \; g_2(\boldsymbol{\theta}) = (\alpha - 1) \sum_{k=1}^K \log \theta_k > 0$$

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 $F_t(\boldsymbol{\theta})$ goes to $f(\boldsymbol{\theta})$ from below, or it is a lower bound for $f(\boldsymbol{\theta})$. In contrast, in the first iteration, if we choose $f_1 = g_2$ then $F_1 > f$, and the sequence of stochastic functions $F_t(\boldsymbol{\theta})$ goes to $f(\boldsymbol{\theta})$ from above, or it is a upper bound for $f(\boldsymbol{\theta})$ (Fig.1).



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

Fig. 1. Two cases of initializing stochastic approximating bounds F_t



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

(b) Choose the higher point at a iteration

Fig. 2. Basic ideas in order to improve OPE

New perspectives leads us to improvements of OPE. Although, OPE is a good candidate for solving a posterior inference in topic models, but we want to enhance OPE in different ways. It makes sense that two stochastic approximating sequences from above and below is better than one. So we construct two sequences that both converging to f, one begins with g_1 called the sequence $\{L_t\}$, another begins with g_2 called the sequence $\{U_t\}$ (Fig. 2).

Using both two stochastic sequences at each iteration gives us more information about objective function f, so that we can get more chances to reach a maximal of f. In this section, we show four different ideas to improve OPE corresponding with four new algorithms, so called OPE1, OPE2, OPE3 and OPE4. These are different from the ways we combine two approximating sequences.

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

Continuing with the idea of rising the randomness, we pick $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$ with probabilities which depends on the value $\{f(\boldsymbol{\theta}_t^u), f(\boldsymbol{\theta}_t^l)\}$. The point at which makes the value of f higher has higher probability of choosing. The probability of selection of $\boldsymbol{\theta}_t$ in OPE2 is smoother than probabilities $\{0.5,\ 0.5\}$ 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 $f(\boldsymbol{\theta}_t^u), f(\boldsymbol{\theta}_t^l)$ and take the point that makes the value of f highest as possible at each iteration (Fig.1). OPE3 works differently from OPE. OPE just constructs 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. Although, the structure of sequence $\{\boldsymbol{\theta}_t\}$ really changes, but

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 $\boldsymbol{\theta}_1$ arbitrary 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 $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

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

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

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

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

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

end for

Algorithm 3 OPE2: Smoothly random 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 $\boldsymbol{\theta}_1$ arbitrary 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 $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

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

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

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

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

$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_{t+1}^u$ 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

OPE's good properties are remained in OPE3.

Another inference algorithm called OPE4 was proposed. We approximate the objective function by a linear combination of the upper bound U_t and lower bound L_t with a suitable

Algorithm 4 OPE3: Choice with the higher value from 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 $\boldsymbol{\theta}_1$ arbitrary 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 $\{\sum_j d_j \log \sum_{k=1}^K \theta_j \beta_{kj}; \quad (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$

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

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

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

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

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

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

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

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

end for

parameter ν, $F_t := \nu U_t + (1 - \nu) L_t$. The usage of both bounds is stochastic in nature and help us reduce the possibility of getting stuck at a local stationary point and this is an efficient approach for escaping saddle points in non-convex optimization. So, our new variant seem to be more appropriate and robust than the OPE. Existing methods become less relevant in high dimensional non-convex optimization. OPE4 is proposed algorithm whose theoretical justification is motivated by ensuring rapid escape from saddle points.

Similar to OPE, OPE4 constructs the sequence $\{\boldsymbol{\theta}_t\}$ converging to $\boldsymbol{\theta}^*$. Although, OPE4 aims at increasing randomness, but OPE4 works differently with OPE. While OPE constructs only one sequence of function F_t, OPE4 constructs again three sequences U_t, L_t, and F_t, in which F_t depending on U_t and L_t. So that, the structure of 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 are remained the same as OPE as analyzed in the next section. So, our new algorithms preserve the key features of OPE.

IV. Analysis of Convergence

From extensive experiments, OPE3 and OPE4 are more efficient than OPE on the two datasets when applying in two learning methods with LDA. Therefore, we focused on the convergence of OPE3 and OPE4.

Theorem 1 (Convergence of OPE3)

Consider the objective function $f(\boldsymbol{\theta})$ in problem (2), given fixed $\boldsymbol{d}$, $\boldsymbol{\beta}$, α. For OPE3, with probability one, the followings hold:

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

2. $\boldsymbol{\theta}_t$ converges to a local maximal/stationary point of $f(\boldsymbol{\theta})$ at a rate of $\mathcal{O}(1/t)$.

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 $\boldsymbol{\theta}_1$ arbitrary 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

Proof:

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

Consider the sequence U_t. Let a_t and b_t be the number of times that we have already picked g_1 and g_2 respectively after t iterations to construct U_t. Note that $a_t + b_t = t$.

Denote $S_t = a_t - b_t$. We have

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

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

$$U_t^{'} - f^{'} = \frac{S_t}{t}(g_1^{'} - g_2^{'}) \tag{3}$$

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

$$E(f_t^u) = \frac{1}{2} g_1 + \frac{1}{2} g_2 = \frac{1}{2} f$$

$$E(U_t) = E(\frac{2}{t} \sum_{h=1}^t f_h^u) = \frac{2}{t} \sum_{h=1}^t E(f_h^u) = \frac{2}{t} \sum_{h=1}^t \frac{1}{2} f = \frac{2}{t} . \frac{t}{2} f = f \tag{4}$$

So U_t is an unbiased estimation of f. Based on proof in [16], combining this with (3), (4) we conclude that the sequence $U_t \to f$, the derivative sequence $U_t^{'} \to f^{'}$ as $t \to +\infty$.

Consider

$$\langle U_t^{'}(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t} \rangle =$$

$$= \langle U_t^{'}(\boldsymbol{\theta}_t) - f^{'}(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t} \rangle + \langle f^{'}(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t^u - \boldsymbol{\theta}_t}{t} \rangle$$

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

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

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

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

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

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

Since $\boldsymbol{e}_t^u$ and $\boldsymbol{\theta}_t$ belong to $\overline{\Delta}_K$, $|\langle g_1'(\boldsymbol{\theta}_t)-g_2'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle|$ and $\|\boldsymbol{e}_t^u-\boldsymbol{\theta}_t\|^2$ are bounded above for any t.

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

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

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

$$\sum_{t=1}^{\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle \leq \sum_{t=1}^{\infty} c_1\frac{|S_t|}{t^2}+f(\boldsymbol{\theta}_\infty)-f(\boldsymbol{\theta}_1)+\sum_{t=1}^{\infty}\frac{c_1L}{t^2}$$

Because f is bounded then $f(\boldsymbol{\theta}_\infty)$ is bounded, and $\sum_{t=1}^{\infty}\frac{c_1L}{t^2}$ also is bounded. Applying proof in [16] we have $\sum_{t=1}^{\infty} c_1\frac{|S_t|}{t^2}$ converges in probability. So $\sum_{t=1}^{\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle$ is bounded.

We have $\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle \geq 0$. If exists $t_0 > 0, c_3 > 0$ such as $\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle \geq c_3 \forall t > t_0$ then

$$\sum_{t=1}^{\infty}\frac{1}{t}\langle U_t'(\boldsymbol{\theta}_t), \boldsymbol{e}_t^u-\boldsymbol{\theta}_t\rangle$$

is divergent.

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

If $\|\boldsymbol{e}_t^u-\boldsymbol{\theta}_t\| \to 0$ as $t \to \infty$ then $\boldsymbol{\theta}_t$ is approaching a vertex of simplex $\overline{\Delta}_K$. In contrast, $\|\boldsymbol{e}_t^u-\boldsymbol{\theta}_t\| \to c$ with $c > 0$ or $\boldsymbol{e}_t^u-\boldsymbol{\theta}_t$ is divergent, then $U_t'(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.

Because of $U_t' \to f'$ as $t \to \infty$ and U_t', f' are continuous, then $f'(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.

It is easy to show that the components of $f_t'(\boldsymbol{\theta}_t)$ are

$$f'(\theta_{kt}) = \sum_j d_j\frac{\beta_{kj}}{\sum_{k=1}^K \theta_{kt}\beta_{kj}}+(\alpha-1)\frac{1}{\theta_{kt}}$$

Consider a real sequence $\{x_n\} \in (0,1), n = 1,2,...,$ we have

$$\sum_j d_j\frac{\beta_{kj}}{\sum_{k=1}^K x_n\beta_{kj}}+(\alpha-1)\frac{1}{x_n}$$

converges, so the sequence $\{x_n\}$ converges.

Because $f_t'(\boldsymbol{\theta}_t)$ converges in K-dimensional space, each component of $f'(\boldsymbol{\theta}_t)$ converges as $t \to \infty$ or $f'(\theta_{kt})$ converges to a finite constant. So θ_{kt} converges when $t \to \infty$, which leads to vector $\boldsymbol{\theta}_t$ converges to $\boldsymbol{\theta}^*$ as $t \to \infty$.

On the other side, $f'(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$ or $\lim_{t\to\infty} f'(\boldsymbol{\theta}_t) = 0$, and $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ as $t \to \infty$.

Then,

$$\lim_{t\to\infty} f'(\boldsymbol{\theta}_t) = f'(\boldsymbol{\theta}^*) = 0$$

In other words, $\boldsymbol{\theta}_t$ converges in probability to a stationary point $\boldsymbol{\theta}^*$ of f.

Due to the update $\boldsymbol{\theta}_{t+1}^u := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t^u-\boldsymbol{\theta}_t}{t}$ and $\boldsymbol{\theta}_{t+1}^l := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t^l-\boldsymbol{\theta}_t}{t}$ it is easy to show that $\boldsymbol{\theta}_t^u$,$\boldsymbol{\theta}_t^l$ and $\boldsymbol{\theta}_t$ converge to $\boldsymbol{\theta}^*$ with the rate $O(1/t)$, which complete the proof.■

Theorem 2 (Convergence of OPE4)

Consider the objective function $f(\boldsymbol{\theta})$ in problem (2), given fixed $\boldsymbol{d}$, $\boldsymbol{\beta}$, α. For OPE4, with probability one, the followings hold:

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

2. $\boldsymbol{\theta}_t$ converges to a local maximal/stationary point of $f(\boldsymbol{\theta})$ at a rate of $\mathcal{O}(1/t)$.

Proof:

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

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

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

$$E(f_t^u) = E(f_t^l) = \frac{1}{2}g_1+\frac{1}{2}g_2 = \frac{1}{2}f$$

$$E(U_t) = E(\frac{2}{t}\sum_{h=1}^{t} f_h^u) = \frac{2}{t}\sum_{h=1}^{t} E(f_h^u) = \frac{2}{t}\sum_{h=1}^{t}\frac{1}{2}f = \frac{2}{t}\cdot\frac{t}{2}f = f$$

$$E(L_t) = E(\frac{2}{t}\sum_{h=1}^{t} f_h^l) = \frac{2}{t}\sum_{h=1}^{t} E(f_h^l) = \frac{2}{t}\sum_{h=1}^{t}\frac{1}{2}f = \frac{2}{t}\cdot\frac{t}{2}f = f$$

$$E(F_t) = \nu E(U_t)+(1-\nu)E(L_t) = \nu f+(1-\nu)f = f$$

Denote

$$S_t^u = a_t - b_t,\ S_t^l = c_t - d_t$$

and

$$S_t = max(|S_t^u|, |S_t^l|)$$

We have

$$U_t = \frac{2}{t}(a_tg_1+b_tg_2) \qquad a_t+b_t = t$$

$$L_t = \frac{2}{t}(c_tg_1+d_tg_2) \qquad c_t+d_t = t$$

$$U_t - f = \frac{S_t^u}{t}(g_1-g_2) \qquad L_t - f = \frac{S_t^l}{t}(g_1-g_2)$$

$$U_t' - f' = \frac{S_t^u}{t}(g_1'-g_2') \qquad L_t' - f' = \frac{S_t^l}{t}(g_1'-g_2')$$

We obtain

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

So F_t is an unbiased estimation of f. We conclude that the sequence $U_t \to f$, the derivative sequence $U_t' \to f'$ as $t \to +\infty$. Similarly we have $L_t \to f$, the derivative sequence $L_t' \to f'$ as $t \to +\infty$.

Due to the update $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \frac{\boldsymbol{e}_t-\boldsymbol{\theta}_t}{t}$ we conclude that $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ as $t \to +\infty$ at the rate $O(1/t)$.

Consider

$$\langle F'_t(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle =$$

$$\langle F'_t(\boldsymbol{\theta}_t) - f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle$$

$$= \langle (\nu\frac{S^u_t}{t} + (1-\nu)\frac{S^l_t}{t})(g'_1(\boldsymbol{\theta}_t) - g'_2(\boldsymbol{\theta}_t)), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle + \langle f'(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle$$

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

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

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

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

$$\langle F'_t(\boldsymbol{\theta}_t), \frac{\boldsymbol{e}_t - \boldsymbol{\theta}_t}{t}\rangle \leq c_1\frac{S_t}{t^2} + f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t) + \frac{c_1 L}{t^2} \quad (6)$$

Summing both sides of (6) for all t we have

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

It is easy to show that $\sum_{t=1}^{\infty}\frac{c_1 L}{t^2}$ also is bounded. Applying proof in [16], we have $\sum_{t=1}^{\infty} c_1\frac{S_t}{t^2}$ converges in probability. So $\sum_{t=1}^{\infty}\frac{1}{t}\langle F'_t(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle$ is bounded.

We have $\langle F'_t(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \geq 0$. 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 $\sum_{t=1}^{\infty}\frac{1}{t}\langle F'_t(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle$ is divergent. Therefore, $\langle F'_t(\boldsymbol{\theta}_t), \boldsymbol{e}_t - \boldsymbol{\theta}_t\rangle \to 0$ as $t \to \infty$.

If $\|\boldsymbol{e}_t - \boldsymbol{\theta}_t\| \to 0$ as $t \to \infty$ then $\boldsymbol{\theta}_t$ is approaching to the vertex of simplex $\overline{\Delta}_K$. In contrast, $\|\boldsymbol{e}_t - \boldsymbol{\theta}_t\| \to c$ with $c > 0$ or $\boldsymbol{e}_t - \boldsymbol{\theta}_t$ is divergent, then $F'_t(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.

Because of $F'_t \to f'$ as $t \to \infty$ and F'_t, f' are continuous and $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ then

$$\lim_{t\to\infty} F'(\boldsymbol{\theta}_t) = \lim_{t\to\infty} f'(\boldsymbol{\theta}_t) = f'(\boldsymbol{\theta}^*) = 0$$

Therefore, $\boldsymbol{\theta}_t$ converges in probability to a stationary point $\boldsymbol{\theta}^*$ of f, which completes the proof. ■

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

V. Experiment

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

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

TABLE I
DATASETS FOR EXPERIMENT

A. Datasets

We used the two large corpora in Table I. Dataset PubMed consists of 330,000 articles from the PubMed central and dataset New York Times (NYT) consists of 300,000 news [1]. To avoid randomness, the learning methods for each dataset is run five times and reported its average results.

B. Parameter settings

To compare our methods with OPE, all free parameter are the same as in [16].

- **Model parameters:** The number of topics $K = 100$, the hyper-parameters $\alpha = \frac{1}{K}$ and $\eta = \frac{1}{K}$. These parameters is commonly used in topic models.
- **Inference parameters:** The number of iterations was chosen as $T = 20$.
- **Learning parameters:** Mini-batch size $S = |C_t| = 5000$. $\kappa = 0.9$,$\tau = 1$ adapted best for existing inference methods. The best parameter ν in OPE4 was selected from $\{0.01, 0.1, 0.2, .., 0.9, 0.99\}$ for each dataset and learning method.

C. Evaluation measures

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

D. Evaluation Results

Fig.3 and Fig.4 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 direct model parameter $\boldsymbol{\beta}$, Online-OPE updates variational parameter $\boldsymbol{\lambda}$. The quality of parameter $\boldsymbol{\theta}$ found by OPE affects directly to the quality of parameters $\{\boldsymbol{\beta}, \boldsymbol{\lambda}\}$.

In practice, OPE performs fast and stable. Stability is shown by the number of iterations T. Just $T = 20$ iterations for OPE results in the same predictive level as $T = 100$. Hence, OPE converges very fast. OPE is also a stable algorithm. The authors [16] did experiments by running 10 times OPE and observed that the 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 iteration nor run it many times. We need to change the structure of sequences that OPE use to maximize goal

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

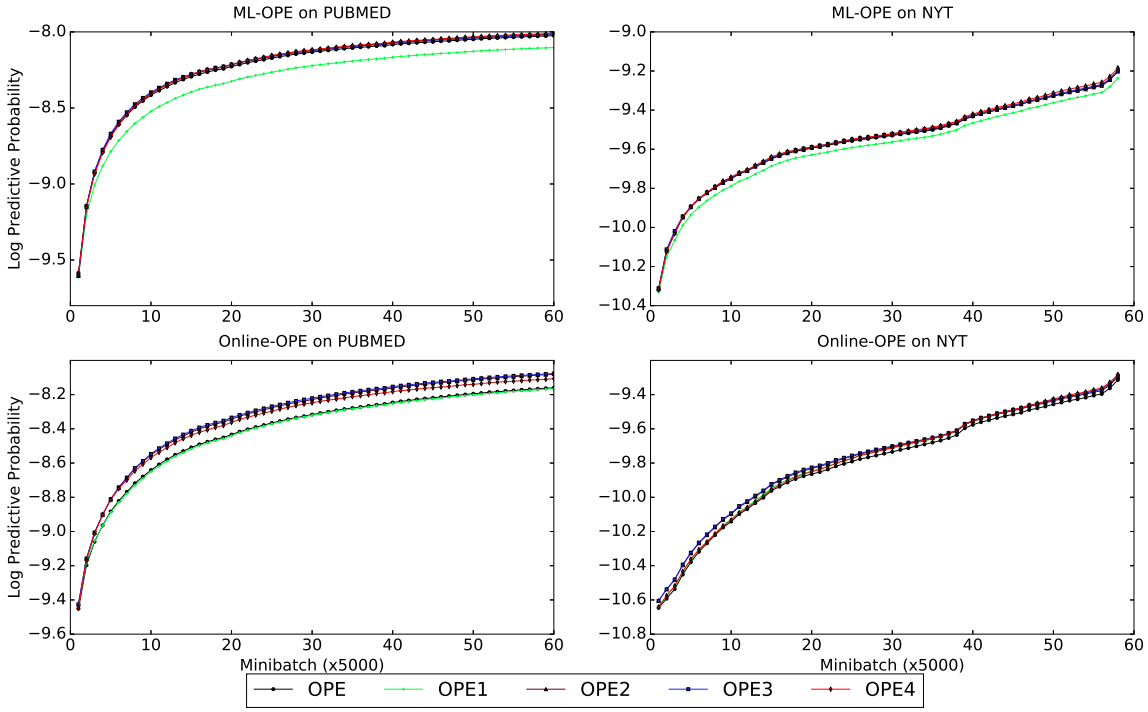


Fig. 3. Results of new algorithms compare with the OPE. Through which to see some new algorithms still ensures good as or even better than OPE.

function. Internal behavior of OPE is interesting and influence clues to enhance algorithm.

Fig. 3 shows that OPE1 and OPE2 are working worse than the remaining algorithms. The way of working of OPE1 and OPE2 is not to promote the idea of increasing the randomness of approximation algorithm. At each iteration, both OPE1 and OPE2 randomly choose one of the two values $\{\boldsymbol{\theta}^u, \boldsymbol{\theta}^l\}$. Thus, maybe for many consecutive iterations, we have selected the values $\boldsymbol{\theta}$ making the objective function f decrease. OPE3 improve this difficulty. OPE3 select point $\boldsymbol{\theta}$ that always makes the value of objective function f higher. Therefore, the quality of parameter $\boldsymbol{\theta}$ learned is better, then the quality of parameter $\boldsymbol{\beta}$ is better. Notice that log predictive probability of OPE3 is higher than ones of OPE1 or OPE2. Similar to OPE3, OPE4 with fit parameter ν has result well. Although, there are differences in results between methods, the differences are very small. Therefore, they do not reflect well the effectiveness of the improvements. Log predictive probability depends on the quality of parameter $\boldsymbol{\beta}$ of ML-OPE and Online-OPE and demonstrates that the quality of parameter $\boldsymbol{\theta}$ is not improved after the inference process.

NPMI measure reveals evidently the quality of the parameter $\boldsymbol{\theta}$ learned through five algorithms. Fig.4 shows that NPMI measure is significantly improved by new OPE variants.

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

It is easy to see that OPE3 makes ML-OPE3 and Online-OPE3 work more efficient. OPE3 demonstrates our idea of using two random sequences of functions to approximate an objective function $f(\boldsymbol{\theta})$. The idea of increasing the randomness and greedy of the algorithm is exploited here. Firstly, two random sequences of function are used to raise our participants and information relevant to objective function. Hence, in the next iteration, we have more choices in $\boldsymbol{\theta}_t$. Secondly, choosing $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$ that makes the value of $f(\boldsymbol{\theta})$ higher in each iteration comes from idea of greedy algorithms. There are many ways of choices here, but we design a best way to create $\boldsymbol{\theta}_t$ from $\{\boldsymbol{\theta}_t^u, \boldsymbol{\theta}_t^l\}$. This approach is simple and there is no need for extra parameter.

In OPE4 experiment, we introduce parameter ν to construct $F(\boldsymbol{\theta}_t) = \nu U(\boldsymbol{\theta}_t) + (1-\nu) L(\boldsymbol{\theta}_t)$. So, we increase the number of parameters in the model and we have to choose the parameter ν empirically. Parameter ν that we find in each dataset is usually about 0.01 or 0.09, which means stochastic bounds

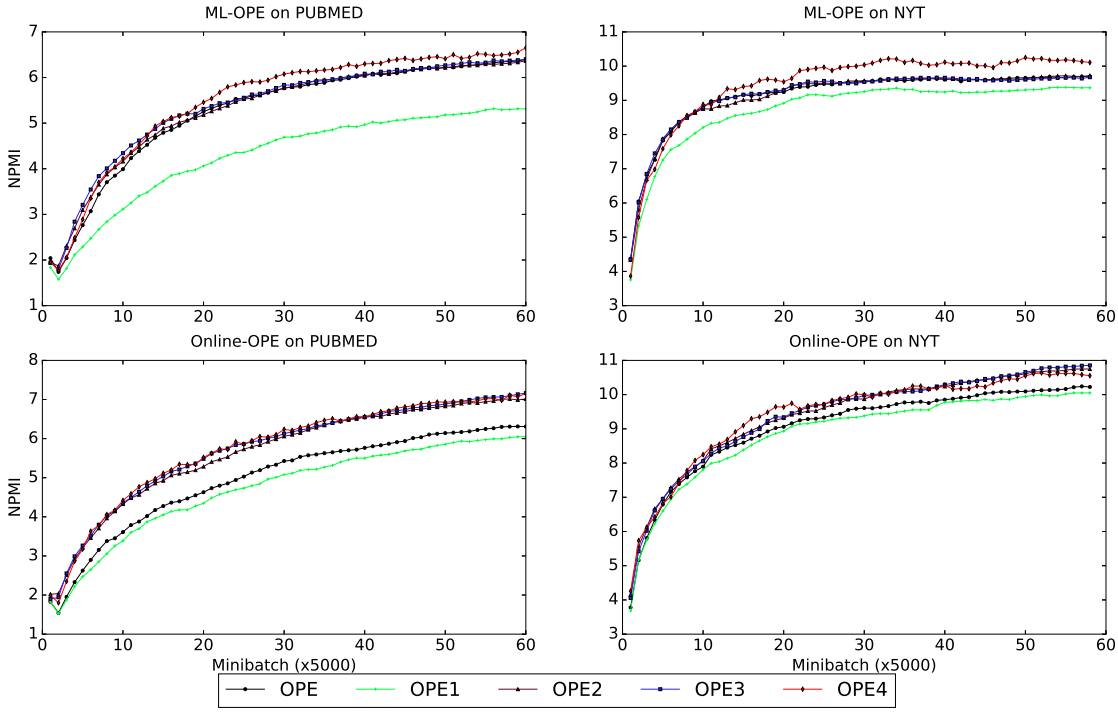


Fig. 4. Results of new algorithms when compare with OPE on NPMI measure. Through which to see the new algorithms still ensures good as or even better than OPE.

always follow one direction below or above. OPE4 uses a linear combination of upper bound U_t and lower bound L_t. The bounds U_t and L_t converge to the objective function f, so the linear combination F_t improves convergence speed and quality of approximation.

OPE4 is the simplest way to combine bounds. We can utilize OPE4 to invent more complicated combination which may result in better approximation. 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 NPMI measure.

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

E. Effect of parameter ν in OPE4

In section II, we find out that OPE has many good characteristics than other algorithms before. Above experiments showed that OPE3 and OPE4 outperform OPE. Especially, we find that OPE4 is the most efficient in almost data sets. However, the effectiveness of OPE4 depends on how the parameter ν is chosen. To see the effect of parameter ν, we changed its values in $\{0.01, 0.1, 0.2...0.9, 0.99\}$ because $0 < \nu < 1$, and of course we fixed the other parameters.

We show some different results of OPE4 when we chose parameter ν between 0 and 1. In the experiments, the value of parameter ν is approximately 1 or 0.5. Although OPE4 works efficiently when using the upper or lower bounds, or the average of two bounds. We suppose that, while the parameter ν is approximately 0 or 1, OPE4 returns the OPE. The value of parameter ν is calculated from experimental data. Due to the choice of parameter ν, OPE4 works better than OPE, but we have to trade-off the running time to find the best parameter ν. Depending on the dataset, we will have to find the suitable value of parameter ν from various parameters. Inappropriate choices of this might affect significantly the performance of OPE4.

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 more fast and effective in practice, then they can be easily extended to a wide class of probabilistic models. By exploiting a new variants of OPE carefully, we have arrived at eight efficient methods for learning LDA from data streams or large corpora. As a result, they are good candidates to help us deal with text streams and big data.

Appendix A
Predictive Probability

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

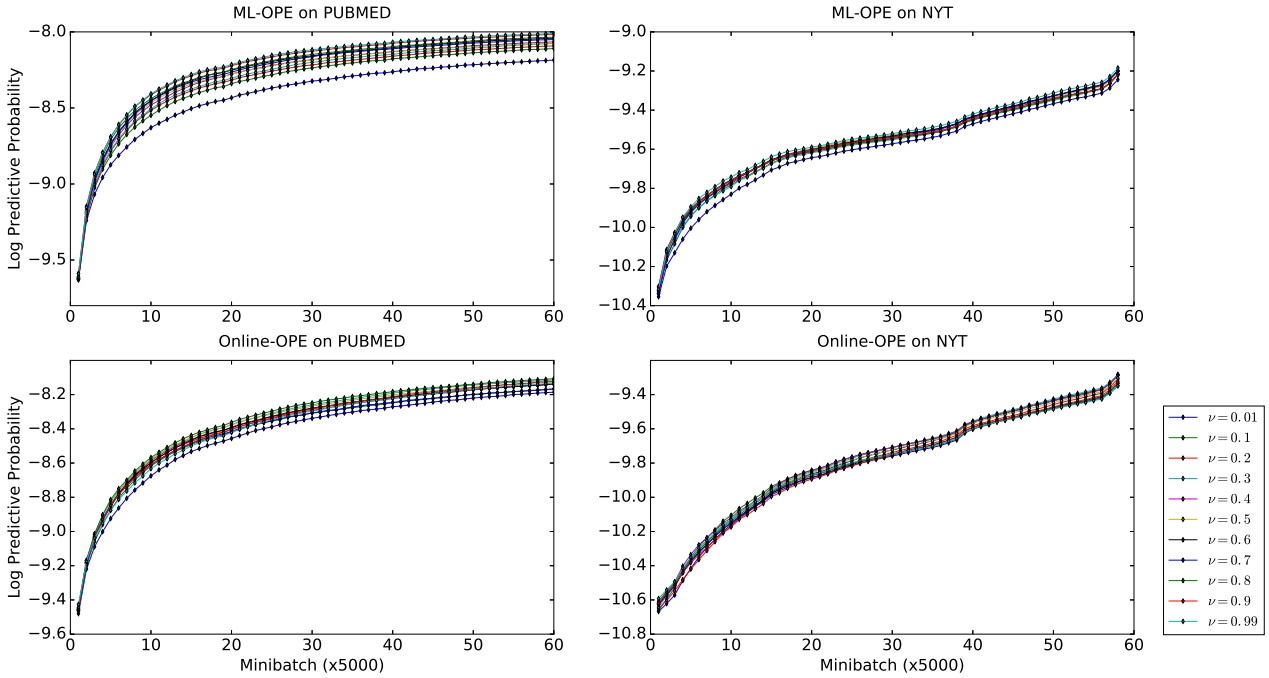


Fig. 5. OPE4 with different selected parameter ν using Predictive Probability measure

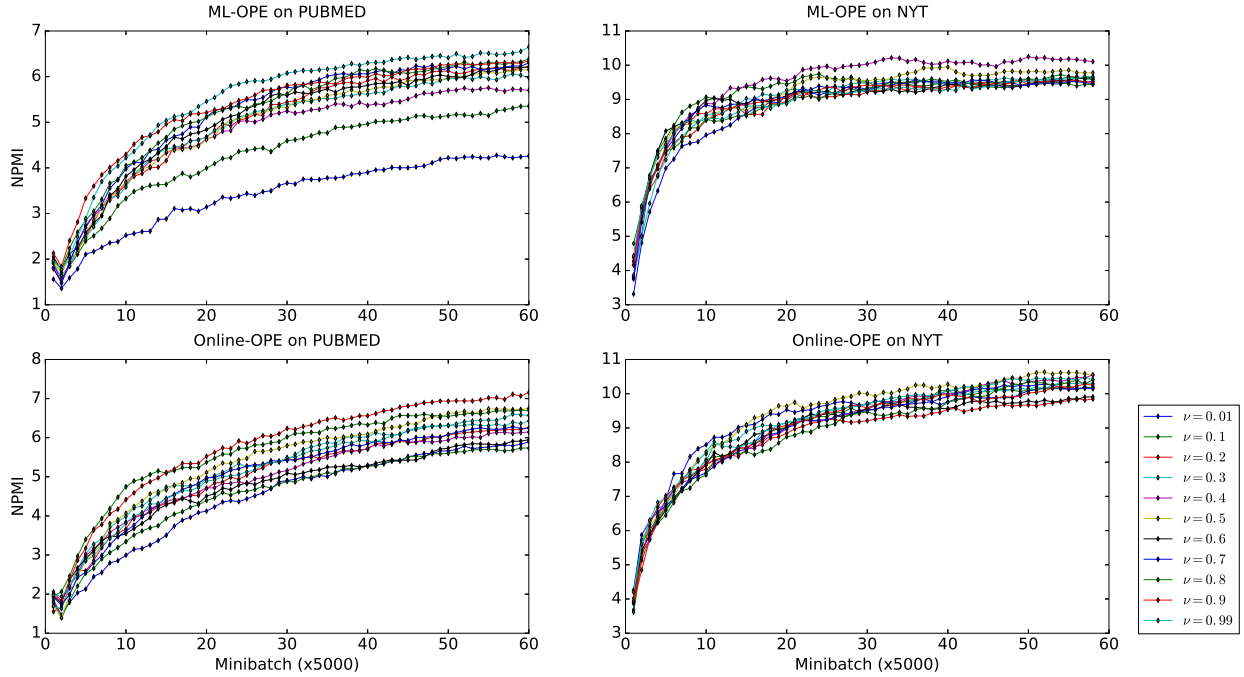


Fig. 6. OPE4 with different selected parameter ν using NPMI measure

in [7] 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.

APPENDIX B
NPMI

NPMI measurement helps us to see the coherence or semantic quality of individual topics. According to [26], 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:

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

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*, pp. 262–272, Association for Computational Linguistics, 2011.

[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*, pp. 937–946, ACM, 2009.

[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), pp. 1599–1606, ACM, 2012.

[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, pp. 280–288, PMLR, 2013.

[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*, pp. 1481–1488, 2007.

[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*, pp. 1353–1360, 2006.

[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*, pp. 27–34, AUAI Press, 2009.

[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*, pp. 1765–1773, 2015.

[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*, pp. 2859–2867, 2016.

[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, pp. 80–95, 2015.

[25] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pp. 13–22, Association for Computational Linguistics, 2013.

[26] 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*, pp. 530–539, 2014.