JMLR: Workshop and Conference Proceedings 39:80–95, 2014 Asian Conference on Machine Learning

Dual online inference for latent Dirichlet allocation

Khoat Than KHOATTQ@SOICT.HUST.EDU.VN
Tung Doan PHONGTUNG_BMW@YAHOO.COM
Hanoi University of Science and Technology, 1, Dai Co Viet road, Hanoi, Vietnam.

Editor: Hang Li and Dinh Phung

Abstract

Latent Dirichlet allocation (LDA) provides an efficient tool to analyze very large text collections. In this paper, we discuss three novel contributions: (1) a proof for the tractability of the MAP estimation of topic mixtures under certain conditions that might fit well with practices, even though the problem is known to be intractable in the worse case; (2) a provably fast algorithm (OFW) for inferring topic mixtures; (3) a dual online algorithm (DOLDA) for learning LDA at a large scale. We show that OFW converges to some local optima, but under certain conditions it can converge to global optima. The discussion of OFW is general and hence can be readily employed to accelerate the MAP estimation in a wide class of probabilistic models. From extensive experiments we find that DOLDA can achieve significantly better predictive performance and semantic quality, with lower runtime, than stochastic variational inference. Further, DOLDA enables us to easily analyze text streams or millions of documents.

Keywords: Topic models, latent Dirichlet allocation, MAP estimation, stochastic gradient, large-scale learning

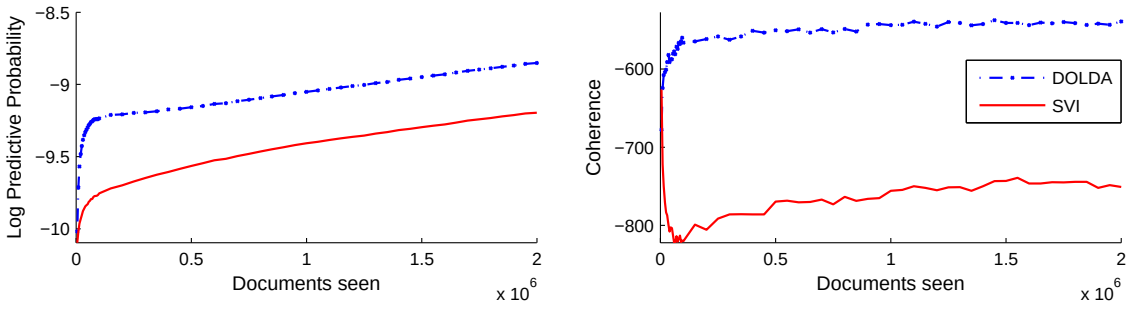
1. Introduction

Latent Dirichlet allocation (LDA) is the class of Bayesian networks that has gained arguably significant interests. It has found successful applications in a wide range of areas including text modeling (Blei, 2012), bioinformatics (Pritchard et al., 2000; Liu et al., 2010), history (Mimno, 2012; Hoffmann, 2013), politics (Grimmer, 2010; Gerrish and Blei, 2012), psychology (Schwartz et al., 2013), to name a few.

One of the core issues in LDA is the estimation of posterior distributions for individual documents. The research community has been studying many approaches for this estimation problem, including variational Bayes (VB) (Blei et al., 2003), collapsed variational Bayes (CVB) (Teh et al., 2007), CVB0 (Asuncion et al., 2009), and collapsed Gibbs sampling (CGS) (Griffiths and Steyvers, 2004; Mimno et al., 2012). Those approaches enable us to easily work with millions of texts (Mimno et al., 2012; Hoffman et al., 2013; Foulds et al., 2013). The quality of LDA in practice is determined by the quality of the inference method being employed. However, none of the mentioned methods has a theoretical guarantee on quality or convergence rate. This is a major drawback of existing inference methods.

Our first contribution in this paper is a proof of the tractability of the following problem:

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



5K	25K	50K	75K	100K	125K	150K	175K
companywide	cased	humana	humana	humana	humana	humana	humana
percentage	governmental	genesis	genesis	genesis	genesis	cellar	cellar
courted	lawbreaker	righted	genomes	cellar	cellar	genesis	researched
millionaire	landed	genomes	dnasequencing	genetically	genetically	genetically	genesis
righted	courted	dnasequencing	genetically	dnasequencing	genomes	scintillating	genetically

Figure 1: (Top left) Predictiveness of the models learned by DOLDA and SVI when analyzing 8 millions of articles in the Pubmed central. (Top right) Semantic quality (measured by Coherence) of the learned models. The higher the better. Note that DOLDA can achieve a high predictiveness level after analyzing a few thounsand articles, but SVI has to analyze milions. The models learned by DOLDA are often significantly more interpretable than those by SVI. (Bottom) Transience of a topic as DOLDA sees more news in New York Times.

for which the multinomial and Dirichlet distributions compose a conjugate pair for the variables $\boldsymbol{d}$ and $\boldsymbol{\theta}$ respectively. In LDA, $\boldsymbol{d}$ represents a document and $\boldsymbol{\theta}$ represents a topic mixture in $\boldsymbol{d}$. This MAP problem was shown to be intractable in the worse case (Sontag and Roy, 2011). Nonetheless, we will show that it is in fact tractable under certain conditions, e.g., $||\boldsymbol{d}||_1$ or the dimensionality of $\boldsymbol{d}$ is sufficiently large. Note that in practice of text modeling, the dimensionality often reaches hundreds of thousands (Asuncion et al., 2011) or even millions (Than and Ho, 2012; Wang et al., 2013). This suggests that the MAP problem in practice might be tractable with high probability. Therefore, our result provides a partial justification for why many existing inference methods empirically succeed even though there is no guarantee on quality.

Our second contribution in this paper is the introduction of an online algorithm, namely *Online Frank-Wolfe* (OFW), for solving the MAP problem (1). We prove that OFW converges to a local maximum of the MAP problem. Under certain conditions, OFW theoretically converges to the global optimum. Furthermore, OFW is able to jump out of local maxima to get close to the global solutions owing to its stochastic nature. Hence, it overcomes many drawbacks of existing inference methods. Those properties help OFW to be more preferable than existing inference methods in many contexts, and provide us real benefits when using OFW in a wide class of probabilistic models, including the LDA-based family.

The topic modeling literature has seen a fast growing interest in designing large-scale learning algorithms (Smola and Narayanamurthy, 2010; Asuncion et al., 2011; Mimno et al., 2012; Than and Ho, 2012; Broderick et al., 2013; Foulds et al., 2013; Patterson and Teh, 2013;

Hoffman et al., 2013). Existing algorithms allow us to easily analyze millions of documents. Those developments are of great significance, even though the posterior estimation is often intractable. Figure 1 illustrates the ability of two large-scale algorithms when working with 8 millions of articles from the Pubmed central. Note that the performance of a method heavily depends on its core inference subroutine. Therefore, existing large-scale learning methods seem to likely remain some of the drawbacks of VB, CVB, CVB0, and CGS.

Our third contribution in this paper is a dual online algorithm (DOLDA) for learning LDA at a large scale. This algorithm owns the online nature when learning the global variables (topics), and employs OFW as the core for inferring local variables ($\boldsymbol{\theta}$) for individual texts. DOLDA overcomes many drawbacks of existing large-scale learning methods owing to the preferable properties of OFW. Figure 1 illustrates the superior behaviors of DOLDA over stochatic variational inference (SVI) by Hoffman et al. (2013). From extensive experiments we find that DOLDA often reaches very fast to a high predictiveness level, and is able to consistently increase the semantic quality and predictiveness of the learned models as observing more data. Therefore, DOLDA provides us an efficient tool for analyzing text streams or collections of big size.

Organization: in the next section, we study the MAP problem (1) and show some interesting properties that fit well with the practice of topic models. We discuss the OFW algorithm for solving (1) in Section 3. We also analyze the convergence property of OFW. Section 4 presents our dual online algorithm for learning LDA from text streams or big text collections. Practical behaviors of DOLDA will be investigated in Section 5. The final section presents some conclusions and discussions.

Notation: Throughout the paper, we use the following conventions and notations. Bold faces denote vectors or matrices. x_i denotes the i^{th} element of vector $\boldsymbol{x}$, and A_{ij} denotes the element at row i and column j of matrix $\boldsymbol{A}$. For a given vector $\boldsymbol{x} = (x_1, ..., x_V)^T$, we denote $\frac{1}{\boldsymbol{x}^a} = (\frac{1}{x_1^a}, ..., \frac{1}{x_V^a})^T$ and denote $diag(\boldsymbol{x})$ as the diagonal matrix whose diagonal entries are $x_1, ..., x_V$, respectively. The unit simplex in the n-dimensional Euclidean space is denoted as $\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \geq 0, \sum_{k=1}^n x_k = 1\}$, and its interior is denoted as $\overline{\Delta}_n$. We will work with text collections with V dimensions (dictionary size). Each document $\boldsymbol{d}$ will be represented as a frequency vector, $\boldsymbol{d} = (d_1, ..., d_V)^T$ where d_j represents the frequency of term j in $\boldsymbol{d}$. Denote n_d as the length of $\boldsymbol{d}$, i.e., $n_d = \sum_j d_j$. The j^{th} unit vector in $\mathbb{R}^n$ is denoted as $\boldsymbol{e}_j$.

2. LDA and approximate inference

LDA is a generative model for modeling texts and discrete data. It assumes that a corpus is composed from K topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$, each of which is a sample from the V-dimensional Dirichlet distribution $Dirichlet(\eta)$. A document $\boldsymbol{d}$ arises 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 $\boldsymbol{\theta}_d = (\theta_{d1}, ..., \theta_{dK})$ represents the contributions of topics to document $\boldsymbol{d}$, while β_{kj} shows the contribution of term j to topic k. Note that $\boldsymbol{\theta}_d \in \Delta_K, \boldsymbol{\beta}_k \in \Delta_V, \forall k$. Both $\boldsymbol{\theta}_d$ and $\boldsymbol{z}_d$ are unobserved variables and are local for each document.

According to Teh et al. (2007), the task of Bayesian inference (learning) given a corpus $\mathcal{C} = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_M\}$ is to estimate the posterior distribution $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\mathcal{C}, \alpha, \eta)$ over the latent topic indicies $\boldsymbol{z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_d\}$, topic mixtures $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M\}$, and topics $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$. The problem of posterior inference for each document $\boldsymbol{d}$, given a model $\{\boldsymbol{\beta}, \alpha\}$, is to estimate the full joint distribution $p(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{d}|\boldsymbol{\beta}, \alpha)$. Direct estimation of this distribution is intractable. Hence existing approaches uses different schemes. VB, CVB, and CVB0 try to estimate the distribution by maximizing a lower bound of the likelihood $p(\boldsymbol{d}|\boldsymbol{\beta}, \alpha)$, whereas CGS (Mimno et al., 2012) tries to estimate $p(\boldsymbol{z}_d|\boldsymbol{d}, \boldsymbol{\beta}, \alpha)$.

2.1. MAP inference of topic mixtures

We now consider the MAP estimation of topic mixture for a given document $\boldsymbol{d}$:

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

It is easy to show that this problem is equivalent to the following one:

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

Sontag and Roy (2011) showed that this problem is NP-hard in the worst case when $\alpha < 1$. In the case of $\alpha \geq 1$, one can easily show that the problem (3) is concave, and therefore it can be solved in polynomial time. Unfortunately, in practice of LDA, the parameter α is often small, says $\alpha < 1$, causing (3) to be nonconcave. That is the reason for why (3) is intractable in the worst case.

2.2. Tractability of the MAP inference problem when $\alpha < 1$

In this section we discuss the tractability of the problem (3) under certain conditions. More specifically, we will show that it is concave with high probability when the dimensionality V or the length of $\boldsymbol{d}$ is large. It implies that in many practical situations, the problem (3) can be solved efficiently.

Let $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$ be the objective function of (3). We have its first and second derivatives at a $\boldsymbol{\theta} \in \overline{\Delta}_K$ as follows:

$$\frac{\partial f}{\partial \theta_t} = \sum_j d_j \frac{\beta_{tj}}{\sum_{k=1}^K \theta_k \beta_{kj}} + \frac{\alpha - 1}{\theta_t}, \quad t = \overline{1, K} \tag{4}$$

$$f'' = -\boldsymbol{A}^T \boldsymbol{A} - (\alpha - 1) diag\left(\frac{1}{\boldsymbol{\theta}^2}\right), \tag{5}$$

where $\boldsymbol{A} = (a_{ji})_{V \times K}$ for $a_{ji} = \frac{\beta_{ij}\sqrt{d_j}}{\sum_{k=1}^K \theta_k \beta_{kj}}$. Note that, denoting $\boldsymbol{B} = \boldsymbol{A} diag(\boldsymbol{\theta})$,

$$
\begin{aligned}
f'' &= diag\left(\frac{1}{\boldsymbol{\theta}}\right).[(1-\alpha)\boldsymbol{I}_K - diag(\boldsymbol{\theta}).(\boldsymbol{A}^T\boldsymbol{A}).diag(\boldsymbol{\theta})].diag\left(\frac{1}{\boldsymbol{\theta}}\right) \\
&= diag\left(\frac{1}{\boldsymbol{\theta}}\right).[(1-\alpha)\boldsymbol{I}_K - \boldsymbol{B}^T\boldsymbol{B}].diag\left(\frac{1}{\boldsymbol{\theta}}\right).
\end{aligned} \tag{6}
$$

A classical result in Algebra (Abadir and Magnus, 2005, exercise 8.28) says that for any symmetric matrix $\boldsymbol{A}$ and nonsingular $\boldsymbol{Y}$, the product $\boldsymbol{Y}\boldsymbol{A}\boldsymbol{Y}^T$ is positive semidefinite if and only if $\boldsymbol{A}$ is positive semidefinite. Consequently, the matrix $(1-\alpha)\boldsymbol{I}_K - \boldsymbol{B}^T\boldsymbol{B}$ decides negative semidefiniteness of f''. This implies the negative semidefiniteness of $(1-\alpha)\boldsymbol{I}_K - \boldsymbol{B}^T\boldsymbol{B}$ decides the concavity of f. As a result, we have

Lemma 1 *Denote* $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k\beta_{kj} + (\alpha-1)\sum_{k=1}^K \log\theta_k$, *for nonnegative vectors* $\boldsymbol{d}$ *and* $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$. *Then* f *is concave over* $\overline{\Delta}_K$ *if the matrix* $(1-\alpha)\boldsymbol{I}_K - \boldsymbol{B}^T\boldsymbol{B}$ *is negative semidefinite.*

Let $\lambda_{min}(\boldsymbol{B}^T\boldsymbol{B})$ be the least eigenvalue of $\boldsymbol{B}^T\boldsymbol{B}$. One can easily see that if $\lambda_{min}(\boldsymbol{B}^T\boldsymbol{B}) \geq 1-\alpha$ then matrix $(1-\alpha)\boldsymbol{I}_K - \boldsymbol{B}^T\boldsymbol{B}$ is negative semidefinite.

Lemma 2 *Function* $f(\boldsymbol{\theta})$ *in Lemma 1 is concave over* $\overline{\Delta}_K$ *if* $\lambda_{min}(\boldsymbol{B}^T\boldsymbol{B}) \geq 1-\alpha$.

The remaining task to see concavity of f is to investigate the conditions for which $\lambda_{min}(\boldsymbol{B}^T\boldsymbol{B}) \geq 1-\alpha$. Note that $\boldsymbol{\theta}$ is a random variable and belongs to the simplex Δ_K. Therefore $\boldsymbol{B}$ is a random matrix. This observation suggests that one can use results from the theory of random matrix to investigate $\boldsymbol{B}$. However, it seems to be very difficult, because of the unknown distribution of the elements in $\boldsymbol{B}$. Indeed, the elements of $\boldsymbol{B}$ are

$$
B_{jt} = \frac{\sqrt{d_j}\beta_{tj}\theta_t}{\sum_{k=1}^K \theta_k\beta_{kj}}, \quad t = \overline{1,K},\ j = \overline{1,V} \tag{7}
$$

Even though $\boldsymbol{\theta}$ follows the Dirichlet distribution with parameter α, deciding the distribution of B_{jt} is really difficult. As a result, further assumptions on $\boldsymbol{B}$ are needed to see $\lambda_{min}(\boldsymbol{B}^T\boldsymbol{B})$.

There is a nice result by Rudelson and Vershynin (2009) for estimating the least singular value of a random matrix with sub-Gaussian entries. Recall that a random variable ξ is called *sub-Gaussian* if its tail is dominated by that of the standard normal random variable, i.e., if there exists $D > 0$ such that

$$
\Pr(|\xi| > t) \leq 2\exp\left(\frac{-t^2}{D^2}\right) \quad \text{for all } t > 0. \tag{8}
$$

The minimal D is called the sub-Gaussian moment of ξ. Note that any bounded random variables are sub-Gaussian.

Theorem 3 *(Rudelson and Vershynin, 2009) Let* $\boldsymbol{S}$ *be a* $V \times K$ *random matrix with* $V \geq K$, *whose elements are independent copies of a sub-Gaussian random variable with unit variance, zero mean, and sub-Gaussian moment* D. *Then for every* $\varepsilon \geq 0$, *we have*

$$
\Pr\left(s_{min}(\boldsymbol{S}) \leq \varepsilon\left(\sqrt{V} - \sqrt{K-1}\right)\right) \leq (C\varepsilon)^{V-K+1} + e^{-cV}, \tag{9}
$$

where constants $C, c > 0$ depend polynomially on the sub-Gaussian moment D; $s_{min}(\boldsymbol{S})$ is the least singular value of $\boldsymbol{S}$.

This theorem essentially says that the least singular value of a tall random matrix is not so small. In particular, $s_{min}(\boldsymbol{S}) > \varepsilon\left(\sqrt{V} - \sqrt{K-1}\right)$ with high probability. $s_{min}(\boldsymbol{S})$ increases as V increases with high probability.

Returning to the matrix $\boldsymbol{B}$ in (7), we easily observe that $0 \leq B_{jt} \leq \sqrt{n_d}$ for any (j,t). It suggests that each B_{jt} is a copy of a sub-Gaussian variable. Further, $B_{jt}/\sqrt{n_d}$ is a copy of a sub-Gaussian variable with variance ≤ 1. As a result, Theorem 3 implies the following.

Theorem 4 *For given nonnegative $\boldsymbol{d}, \boldsymbol{\beta}, \alpha$, and $V \geq K$, consider matrix $\boldsymbol{B}$ whose elements are defined as in (7). Assume all elements of $\boldsymbol{B}/\sqrt{n_d}$ are copies of a sub-Gaussian variable with variance 1, and mean 0, and sub-Gaussian moment D. Then for every $\varepsilon > 0$, we have*

$$\Pr\left(s_{min}\left(\frac{1}{\sqrt{n_d}}\boldsymbol{B}\right) \leq \varepsilon\left(\sqrt{V} - \sqrt{K-1}\right)\right) \leq (C\varepsilon)^{V-K+1} + e^{-cV}, \tag{10}$$

$$\Pr\left(\lambda_{min}\left(\frac{1}{n_d}\boldsymbol{B}^T\boldsymbol{B}\right) \leq \varepsilon^2\left(\sqrt{V} - \sqrt{K-1}\right)^2\right) \leq (C\varepsilon)^{V-K+1} + e^{-cV}, \tag{11}$$

where constants $C, c > 0$ depend polynomially only on D.

Inequality (11) suggests that with probability at least $1 - (C\varepsilon)^{V-K+1} - e^{-cV}$, we have

$$\lambda_{min}\left(\boldsymbol{B}^T\boldsymbol{B}\right) > \varepsilon^2 n_d\left(\sqrt{V} - \sqrt{K-1}\right)^2 \tag{12}$$

Combining this observation with Lemma 2, we obtain the following result.

Corollary 5 *Using the notations and assumptions in Theorem 4, consider the function $f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1)\sum_{k=1}^K \log\theta_k$. If $\varepsilon^2 n_d\left(\sqrt{V} - \sqrt{K-1}\right)^2 \geq 1 - \alpha$, for any $\varepsilon \in \left(0, \frac{1}{C}\right)$, then $f(\boldsymbol{\theta})$ is concave over $\overline{\Delta}_K$ with probability at least $1 - (C\varepsilon)^{V-K+1} - e^{-cV}$.*

This corollary immediately implies our main result about the tractability of the MAP inference problem (3).

Theorem 6 (Concavity of MAP inference) *Using the notations and assumptions in Theorem 4, for any $\varepsilon \in \left(0, \frac{1}{C}\right)$, if*

$$\varepsilon^2 n_d\left(\sqrt{V} - \sqrt{K-1}\right)^2 \geq 1 - \alpha, \tag{13}$$

then the MAP inference problem (3) is concave over $\overline{\Delta}_K$ with probability at least $1 - (C\varepsilon)^{V-K+1} - e^{-cV}$.

For the first time the tractability of the MAP inference problem in LDA has been proved, in the case of $\alpha < 1$. This is in contrast with the intractability result in the worst case by Sontag and Roy (2011). From (13) one can easily find that when the document length n_d or the dimensionality V are very large, the MAP problem is concave with high probability.

Corollary 7 (Concavity for long documents) *Using the assumptions in Theorem 4,*

+ *if* $n_d(\sqrt{V}-\sqrt{K-1})^4 \geq C^4(1-\alpha)^2$ *then the problem (3) is concave with probability at least* $1-(n_d)^{-\frac{1}{4}(V-K+1)}-e^{-cV}$.

+ *As* $n_d \to +\infty$, *the problem (3) is concave with probability at least* $1-e^{-cV}$.

Proof The first statement can be derived from Theorem 6 by choosing $\varepsilon = \frac{1}{C}n_d^{-\frac{1}{4}}$. The second statement thus follows. ■

Corollary 8 (Concavity for high dimensionality) *Using the notations and assumptions in Theorem 4, let* K *and* n_d *be fixed. Then the MAP problem (3) is concave over* $\overline{\Delta}_K$ *with probability 1 as* $V \to +\infty$.

2.3. Connection to practices

Both Corollaries 7 and 8 show the two extremes for which the MAP problem in LDA is tractable. Corollary 7 suggests that as the document length grows, the MAP problem is more likely tractable, and can be solved in polynomial time. This result is similar in merit with the one by Tang et al. (2014), who showed that the LDA model is more likely to be recovered as the document length grows.

The result in Corollary 8 is much more interesting, because it seems to fit well with the practice of topic models. When modeling text collections, the dimensionality V (dictionary size) easily reaches hundreds of thousands (Asuncion et al., 2011; Smola and Narayanamurthy, 2010) or even millions (Than and Ho, 2012; Wang et al., 2013). For such collections, posterior inference for individual texts is concave with very high probability. Inequality (13) tells even more about the tractability of problem (3) in practice. We observe that K in practice is often significantly less than V. Further, the document length in some types of texts might be large (Mimno et al., 2012). Combination of those factors makes (13) to happen more likely, and hence (3) is concave with high probability. All of these observations suggest that the results in Theorem 6, Corollaries 7 and 8 provide a new perspective that explains the practical success of many existing inference methods.

3. Online Frank-Wolfe for MAP inference

We have discussed the tractability of the MAP inference problem in the last section. Next we present a novel algorithm for doing inference of topic mixtures for documents. Our algorithm, namely *Online Frank-Wolfe* (OFW), is theoretically guaranteed to converge to a local maximum. In some cases, it converges to the global solutions. To the best of our knowledge, OFW is the first algorithm for posterior inference for individual texts that has a guarantee on inference quality and convergence rate. Hence, OFW overcomes many drawbacks of VB, CVB, CVB0, and CGS.

Details of OFW is presented in Algorithm 1.[1] It is worth noting that OFW is a careful adaptation of the general algorithm by Hazan and Kale (2012). A crucial point is that

1. In practice, we can set $a = 1/\sqrt{\ell}+\epsilon$ for a very small constant ϵ, says $\epsilon = 10^{-10}$, to assure that $\boldsymbol{\theta}_\ell$ always stays in the interior of Δ_K.

Algorithm 1 Online Frank-Wolfe for MAP inference

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 $\overline{\Delta}_K = \{\boldsymbol{x} \in \mathbb{R}^K : \sum_{k=1}^K x_k = 1, \boldsymbol{x} > 0\}$.
for $\ell = 1, ..., \infty$ **do**
  Pick f_ℓ uniformly from $\{\sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj};\ (\alpha - 1) \sum_{k=1}^K \log \theta_k\}$
  $F_\ell := \frac{2}{\ell} \sum_{h=1}^{\ell} f_h$
  $i' := \arg\max_i \nabla F_\ell(\boldsymbol{\theta}_\ell)_i$; (maximal partial gradient)
  $a := 1/\sqrt{\ell}$;
  $\boldsymbol{\theta}_{\ell+1} := a\boldsymbol{e}_{i'} + (1 - a)\boldsymbol{\theta}_\ell$.
end for

OFW here has significantly better bounds on both quality and convergence rate than that of Hazan and Kale (2012).

Theorem 9 (Convergence for concavity) *Consider the objective function $f(\boldsymbol{\theta})$ in problem (3), given fixed $\boldsymbol{d}, \boldsymbol{\beta}, \alpha$. Assuming f is concave over $\overline{\Delta}_K$, Algorithm 1 returns an iterate $\boldsymbol{\theta}_\ell$ with a regret bound of $O\left(1/\sqrt{\ell}\right)$ after ℓ iterations, i.e., $\max_{\boldsymbol{\theta}} F_\ell(\boldsymbol{\theta}) - F_\ell(\boldsymbol{\theta}_\ell) \leq C/\sqrt{\ell}$ for some constant $C > 0$. Furthermore, $F_\ell(\boldsymbol{\theta}) \to f(\boldsymbol{\theta})$ as $\ell \to +\infty$, implying that Algorithm 1 converges to the global solution at a rate of $O\left(1/\sqrt{\ell}\right)$.*

Proof The regret bound has been proven in Theorem 3.1 by Hazan and Kale (2012) (for the setting $b = 0, d = 0.5$).

Denote $g_1 = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$ and $g_2 = (\alpha - 1) \sum_{k=1}^K \log \theta_k$. Let a_ℓ and b_ℓ be the number of times that we have already picked g_1 and g_2 respectively after ℓ iterations. Due to the uniform sampling of f_ℓ, we have that $a_\ell + b_\ell = \ell$ and $a_\ell/\ell \to 0.5, b_\ell/\ell \to 0.5$ as $\ell \to +\infty$. Therefore, $F_\ell = \frac{2}{\ell} \sum_{t=1}^{\ell} f_t = \frac{2}{\ell}(a_\ell g_1 + b_\ell g_2) \to f$ as $\ell \to +\infty$. ■

It is worth remarking that OFW follows the common greedy approach, using gradient to guide the direction for searching. Therefore when $f(\boldsymbol{\theta})$ is not concave, OFW is able to converge to a local maximum because of the last statement in Theorem 9. As a result, we have the following.

Corollary 10 (Convergence for non-concavity) *Consider the objective function $f(\boldsymbol{\theta})$ for fixed $\boldsymbol{d}, \boldsymbol{\beta}, \alpha$. If f is non-concave, Algorithm 1 converges to a local maximum of (3) at a rate of $O\left(1/\sqrt{\ell}\right)$.*

Comparing with other inference approaches (including VB, CVB, CVB0 and CGS), our algorithm has many preferable properties. The most attractive property of OFW is the theoretical guarantees on quality and convergence rate, as shown in Theorem 9 and Corollary 10. Existing inference methods often do not have any guarantee. In the case of non-concavity of the MAP problem (3), OFW is able to jump out of local maxima to get close to the global solutions owing to its stochastic nature. This is another interesting

Algorithm 2 DOLDA: online learning for latent Dirichlet allocation

Input: $K, \alpha > 0, \tau \geq 0, \kappa \in (0.5, 1]$
Output: $\boldsymbol{\beta}$
Initialize $\boldsymbol{\beta}^0$ randomly
for $t = 1, ..., \infty$ **do**
  Pick a set $\mathcal{C}_t$ of documents
  For each $\boldsymbol{d} \in \mathcal{C}_t$ do inference for $\boldsymbol{d}$ by OFW to get $\boldsymbol{\theta}_d$, given $\boldsymbol{\beta}^{t-1}$
  Compute intermediate topics $\hat{\boldsymbol{\beta}}^t$ as: $\hat{\beta}^t_{kj} \propto \sum_{\boldsymbol{d} \in \mathcal{C}_t} d_j \theta_{dk}$
  Set step-size: $\rho_t = (t + \tau)^{-\kappa}$
  Update topics: $\boldsymbol{\beta}^t := (1 - \rho_t)\, \boldsymbol{\beta}^{t-1} + \rho_t \hat{\boldsymbol{\beta}}^t$
end for

property. One can easily see that the memory requirement for implementing OFW is very modest, says $O(K)$, which is significantly less than that of VB, CVB, CVB0, and CGS.

OFW is very general for solving the MAP problem (2). Hence, it can be adapted easily to a large class of probabilistic models, including the LDA-based family, for which the multinomial distribution is used to model a discrete variable $\boldsymbol{d}$ with a Dirichlet prior.

4. Dual online algorithm for learning LDA

In this section we describe a novel algorithm, namely *Dual Online Algorithm* (DOLDA), for learning LDA from large corpora. DOLDA employs OFW to do MAP inference for individual documents, and the online scheme (Hoffman et al., 2013) to infer the global variables (topics). Hence, the online nature appears in both local and global inference phases. Note that the inference of local variables by OFW has theoretical guarantees on quality and convergence rate. Such a property might help DOLDA be more attractive than other large-scale learning methods.

4.1. Derivation of batch learning

We first discuss how to design a batch learning algorithm using OFW. The MAP learning problem, given a corpus $\mathcal{C} = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_M\}$ and $\alpha > 0$, is to estimate the topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$ that maximize

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \sum_{\boldsymbol{d} \in \mathcal{C}} \log \Pr\left(\boldsymbol{\theta}_d, \boldsymbol{d} | \boldsymbol{\beta}, \alpha\right) \\ &= \sum_{\boldsymbol{d} \in \mathcal{C}} \left(\sum_j d_j \log \sum_{k=1}^{K} \theta_{dk} \beta_{kj} + (\alpha - 1) \sum_{k=1}^{K} \log \theta_{dk} \right) + constant \end{aligned} \tag{14}$$

We maximize $\mathcal{L}(\boldsymbol{\beta})$ by alternating the following two steps until convergence: Step 1 maximizes $\Pr\left(\boldsymbol{\theta}_d, \boldsymbol{d} | \boldsymbol{\beta}, \alpha\right)$ by OFW to infer $\boldsymbol{\theta}_d$ for each $\boldsymbol{d} \in \mathcal{C}$; Step 2 maximizes $\mathcal{L}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ for fixed $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M$. By using the same arguments as Than and Ho (2012) we can arrive at the following formula to update $\boldsymbol{\beta}$ in Step 2:

$$\beta_{kj} \propto \sum_{\boldsymbol{d} \in \mathcal{C}} d_j \theta_{dk} \tag{15}$$

4.2. Adaptation to online learning

Formula (15) forms the batch learning of topics for LDA. We use the simple method by Hoffman et al. (2013) to design an online algorithm from (15). More specifically, the algorithm repeats the following steps:

- Sample a set $\mathcal{C}_t$ of documents. Optimize the local variables for each document $\boldsymbol{d} \in \mathcal{C}_t$, given the global variable $\boldsymbol{\beta}^{t-1}$ in the last step.
- Form an intermediate global variable $\hat{\boldsymbol{\beta}}^t$ for $\mathcal{C}_t$.
- Update the global variable to be a weighted average of the intermediate $\hat{\boldsymbol{\beta}}^t$ and $\boldsymbol{\beta}^{t-1}$.

Hoffman et al. (2013) show that with such a scheme, the online algorithm works as stochastic natural accent on the global variable. When working on a fixed corpus, the algorithm will converge to a stationary point of the objective function.

Algorithm 2 is an adaptation from the batch learning (15) to online learning, using the technique by Hoffman et al. (2013). Note that the step-size $\rho_t = (t+\tau)^{-\kappa}$ must satisfies two conditions: $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$. These conditions are to assure that the learning algorithm will converge. $\kappa \in (0.5, 1]$ is the forgeting rate, the higher the lesser the algorithm weighs the role of new data.

5. Empirical evaluation

In this section, we evaluate the practical performance of DOLDA and OFW. We first want to see the predictiveness and semantic quality of the models which are learned by DOLDA. We then want to see how fast DOLDA learns a qualified model as more data come. Finally, we want to see how fast OFW does inference in practice, despite of the theoretical guarantee for fast convergence in Section 3. To this end, we take *stochastic variational inference* (SVI) (Hoffman et al., 2013) into consideration as the state-of-the-art method for learning LDA at a large scale.[2]

We use two large data sets for evaluation: *Pubmed* consisting of 8.2 millions of medical articles from the pubmed central; *New York Times* consisting of 300K news.[3] The vocabulary size (V) of each corpus is more than 110,000. For each corpus we set aside randomly 1000 documents for testing, and used the remaining for learning.

Parameter settings: There are some parameters (τ, κ) in SVI and DOLDA that have to be carefully considered. To avoid any possible bias in comparisons, we follow the study by Hoffman et al. (2013) to select the best values for those parameters. According to Hoffman et al. (2013), the performance of SVI does not depend heavily on τ, and is better for greater values of κ. Therefore we chose $\tau = 1$ and $\kappa = 0.9$ in our experiments. Larger size of minibatches often helps SVI to learn better, hence we chose minibatches of size 5000. We allowed at most 50 iterations for VB in SVI and OFW in DOLDA to do inference for individual documents. More interations did not help VB consistently increase the quality in our observation. The hyperparamenters for the Dirichlet priors in LDA were set as $\alpha = 1/K, \eta = 1/K$, which were suggested in previous studies.

2. SVI was taken from http://www.cs.princeton.edu/~blei/downloads/onlineldavb.tar.
3. The data were retrieved from http://archive.ics.uci.edu/ml/datasets/

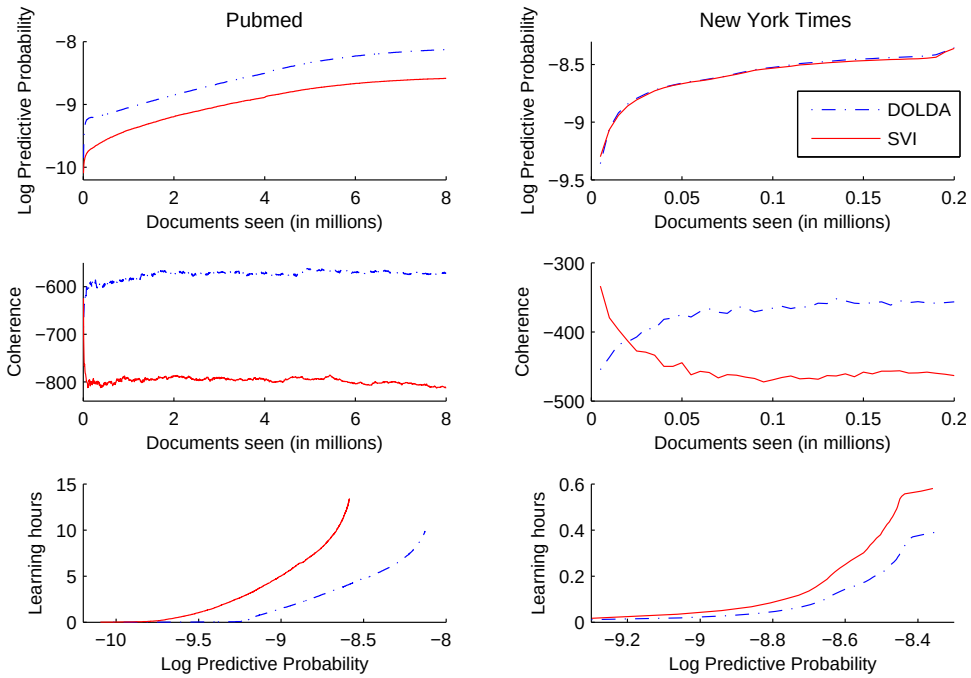


Figure 2: Performance of DOLDA and SVI on two large corpora when learning 100-topic LDA. The higher the better for *Predictive Probability* and *Coherence*, whereas lower is better for *Learning hours*. The last row shows how long the two methods reach to the same generalization level.

Performance measures: We used *Coherence* and *Predictive Probability* to measure the quality of a model which has been learned from the training data. Coherence (Mimno et al., 2011) measures the semantic quality of individual topics, while predictive probability (Hoffman et al., 2013) measures the predictiveness and generalization to new data.

Coherence computation: To calculate the coherence of a topic k, we first choose the set $V^k = \{v_1^k, ..., v_t^k\}$ of the top t terms that have highest probabilities in that topic, and then compute $C(k, V^k) = \sum_{m=2}^{t} \sum_{l=1}^{m-1} \log \frac{D(v_m^k, v_l^k)+1}{D(v_l^k)}$ where $D(v)$ is the document frequency of term v, $D(u, v)$ is the number of documents that contain both terms u and v. In our investigation, we chose top $t = 20$ terms, and coherence is averaged across all topics: $Coherence = \frac{1}{K} \sum_{k=1}^{K} C(k, V^k)$.

Predictive Probability shows the predictiveness and generalization of a model $\mathcal{M}$ on new data. We followed the procedure in (Hoffman et al., 2013) to compute this quantity. 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 then did inference for w_{obs} and then estimate the predictive distribution $\Pr(w_{ho}|w_{obs}, \mathcal{M})$ where $\mathcal{M}$ is the model to be measured. Predictive Probability was averaged from 5 random splits, each was on 1000 documents.

5.1. Performance of DOLDA

We first want to see how well DOLDA learns in comparison with SVI. Figure 2 presents the results on two corpora. One can easily observe that as seeing more documents, both DOLDA and SVI reached to better predictiveness levels with a fast rate. For Pubmed, DOLDA performed significantly better than SVI even just after seeing a few thounsands of documents. DOLDA often reached at the same generalization level (measured by log predictive probability) as SVI within a much less runtime. SVI often needed much more time and data to reach the same prediction level as DOLDA. This demonstrates the goodness of our algorithm.

The fast rate of generalization of the two methods seems to inherit from the goodness of their inference algorithms. VB has been known to work well in many previous studies and empirically has a good convergence rate, even though no explicit guarantee has been derived. This might be the main reason for the good performance of SVI in terms of predictiveness. Remember from Section 3 that OFW has a provably fast rate of convergence and a theoretically good bound on inference quality. Also, OFW is the core subroutine of DOLDA. As a result, DOLDA seems to owe the good performance to the goodness of OFW.

There is a strange behavior of SVI in terms of Coherence. The second row in Figure 2 shows that the semantic quality of the models learned by SVI gradually decreases as seeing more data. According to Mimno et al. (2011), Coherence agrees highly with human assessment about the semantic quality and interpretability of individual topics. This suggests that the interpretabilty of topics seems to decrease as SVI sees more documents. This behavior is unexpected in practice, and should be studied further. In contrast, DOLDA can learn more interpretable topics as seeing more documents. There is a big gap between DOLDA and SVI in terms of coherence, just after learning from a few thounsands of documents. All of those observations suggest the superior behaviors of DOLDA.

5.2. Sensitivity of DOLDA

We next investigate the effects of the parameters on the performance of DOLDA. The parameters include: the forgetting rate κ, τ, the number L of interations for OFW, and the minibatch size. Inappropriate choices of those parameters might affect significantly the performance of DOLDA. To see the effect of a parameter, we changed its values in a finite set, but fixed the other parameters. Results of our experiments are depicted in Figure 3.

We observe that κ and L did not significantly affect the performance of DOLDA. These behaviors of DOLDA are interesting and beneficial in practice. Indeed, we do not have to consider much about the effect of the forgetting rate κ and thus no expensive model selection is necessary. Figure 3(b) reveals a much more interesting behavior of OFW. One easily observes that more iterations in OFW did not necessarily help the performance of DOLDA. Just $L = 10$ iterations for OFW resulted in a comparable predictiveness level as $L = 100$. It suggests that OFW converges very fast in practice, and that $L = 10$ might be enough for practical employments of OFW. This behavior is really beneficial in practice, especially for massive data or streaming data.

τ and minibatch size did affect DOLDA significantly. Similar with the observation by Hoffman et al. (2013) for SVI, we observe that DOLDA performed consistently better as the minibatch size increased. It can reach to a very high predictiveness level with a fast

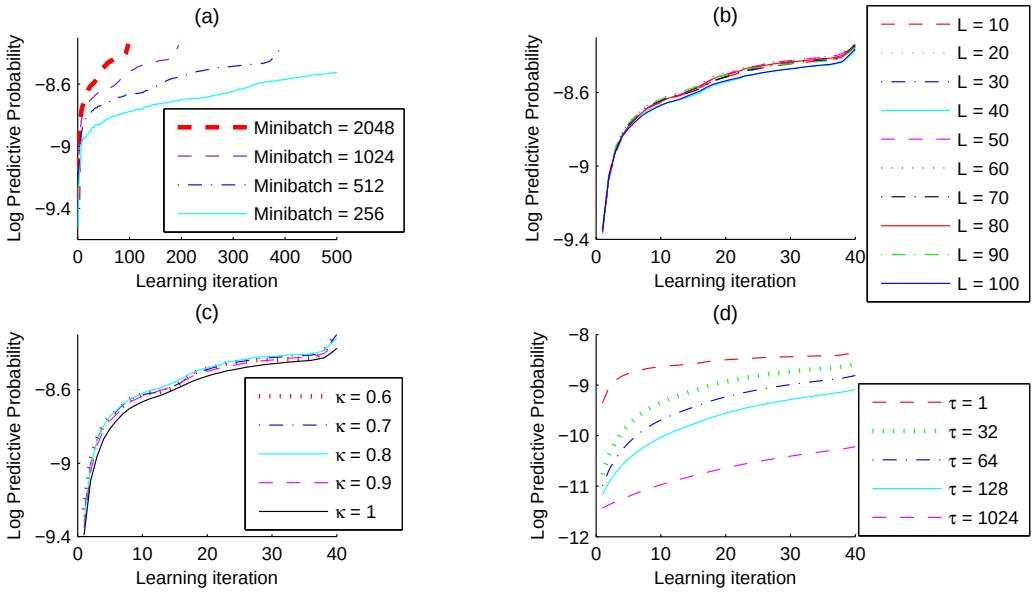


Figure 3: Sensitivity of DOLDA when changing parameters. (a) Change the minibatch size when fixed $\{\kappa = 0.9, \tau = 1, L = 50\}$. (b) Change the number L of iterations for OFW when fixed $\{\kappa = 0.9, \tau = 1\}$. (c) Change the forgetting rate κ when fixed $\{\tau = 1, L = 50\}$. (d) Change τ when fixed $\{\kappa = 0.9, L = 50\}$. The minibatch size in the cases of (b), (c), (d) is 5000. All of these experiments were done on New York Times, with $K = 100$ topics.

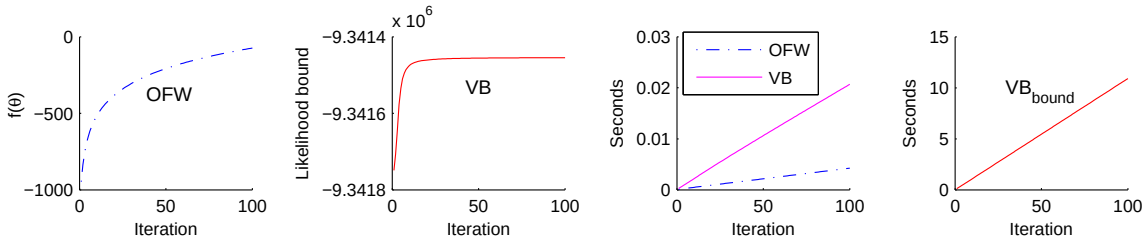


Figure 4: Convergence and inference time of OFW and VB as the number of iterations increase. The first two subplots show how fast OFW and VB maximize their objective functions, while the last two subplots show how long they took. The last subplot shows how long VB did inference when the lower bound of $\Pr(\boldsymbol{d}|\boldsymbol{\beta}, \alpha, \eta)$ was used to check convergence. Note that VB did hundreds of times faster than VB_{bound}, i.e., checking bounds for convergence in VB requires intensive time.

rate. In contrast, DOLDA performed worse as τ increased. The method performed best at $\tau = 1$. It is worth noting that the dependence between the performance of DOLDA and $\{\tau$, minibatch size$\}$ is monotonic. Such a behavior enables us to easily choose a good setting for the parameters of DOLDA in practice.

5.3. Convergence, time, and stability of OFW

Next we investigate the performance of OFW. We want to see its convergence rate, inference time, and stability. To this end, we took the 100-topic LDA as a fixed model which has been learned by SVI previously from New York Times; and then we did inference on individual testing documents by OFW and VB. Both methods were allowed 100 iterations to do inference on a document. Results are depicted in Figure 4.

Observing Figure 4 we find that both method converged very fast. VB reached convergence just after $L = 20$ iterations, while OFW consistently improved approximate solutions with a high rate as allowing more iterations. The early convergence of VB is very interesting and might be beneficial in practice. Note that such an early convergence does not neccessarily reflect the goodness of approximating $\Pr(\boldsymbol{d}|\boldsymbol{\beta}, \alpha, \eta)$, since VB is just able to compute a lower bound. The inferior performance of SVI on Pubmed in terms of predictive probability suggests that there might be a big gap between the likelihood and the lower bound used in VB. Hence, VB might require more than 20 iterations to do inference on each document. In contrast, few iterations for OFW may be sufficient to help DOLDA learn well, as illustrated in Figure 3(b). This shows a better behavior of OFW over VB.

OFW often performed significantly faster than VB, since no convergence check is required and the computation in OFW is very basic. Meanwhile, VB needs to compute a lower bound of the likelihood which is very expensive to estimate, since VB does inference by maximizing that lower bound. That is the reason for why VB performed hundreds of times more slowly than OFW. Hoffman et al. (2013) suggested to avoid the lower bound computation to reduce runtime, which agrees with our experiments as illustrated in the third subplot in Figure 4.

Stability of OFW: our last investigation is whether or not OFW performs stably in practice. We have to consider this behavior as there are two probabilistic steps in OFW: initialization of $\boldsymbol{\theta_1}$ and pick of f_ℓ. To see the stability, we took 100 testing documents to do inference given the 100-topic LDA model previously learned by DOLDA from New York Times. For each document, we did 10 random runs for OFW and then saved the objective values of the last iterates. We found that all of 10 objective values centralized around their mean with a standard deviation of 2.4035 (on average among 100 documents). Such a small deviation strongly suggests that OFW can perform very stably.

6. Conclusion

We have presented three main contributions in this paper. The proof of the tractability of the MAP problem is not only for LDA, but a wide class of probabilistic models. It partially explains why existing inference methods succeed in practice of topic modeling, inspite of no theoretical guarantee on quality and convergence rate. Our method, OFW for solving the MAP problem, has many nice properties that existing ones do not have. OFW can be readily employed to do the posterior estimation in a wide class of models. Finally, our

online algorithm (DOLDA) for learning LDA at a large scale has many preferable behaviors in practice. DOLDA overcomes existing learning algorithms for LDA in various aspects, including a theoretical guarantee on inference quality, a fast rate of convergence to a high predictiveness level, and a streaming nature for dealing with text streams.

Acknowledgments

This work was partly supported by Asian Office of Aerospace R&D under agreement number FA2386-13-1-4046.

References

Karim M. Abadir and Jan R. Magnus. *Matrix Algebra*. Cambridge University Press, 2005.

A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh. On smoothing and inference for topic models. In *Proceedings of the UAI*, pages 27–34, 2009.

Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed estimation of topic models for document analysis. *Statistical Methodology*, 8(1):3–17, 2011.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(3):993–1022, 2003.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael Jordan. Streaming variational bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2013.

James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the KDD*, pages 446–454. ACM, 2013.

Sean Gerrish and David Blei. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems*, 2012.

T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.

Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th Annual International Conference on Machine Learning (ICML)*, 2012.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Leah Hoffmann. Looking back at big data. *Communications of the ACM*, 56(4):21–23, 2013. doi: 10.1145/2436256.2436263.

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*, 26(24):3105, 2010.

David Mimno. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, 5(1):3, 2012.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272, 2011.

David Mimno, Matthew D. Hoffman, and David M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *Proceedings of the 29th Annual International Conference on Machine Learning*, 2012.

Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

H Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Martin EP Seligman, Lyle H Ungar, Eduardo Blanco, Michal Kosinski, and David Stillwell. Toward personality insights from language exploration in social media. In *AAAI Spring Symposium Series*, 2013.

Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

David Sontag and Daniel M. Roy. Complexity of inference in latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198, 2014.

Y.W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, volume 19, page 1353, 2007.

Khoat Than and Tu Bao Ho. Fully sparse topic models. In Peter Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7523 of *Lecture Notes in Computer Science*, pages 490–505. Springer, 2012.

Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Trans. Inf. Syst.*, 31(1):5:1–5:44, 2013. doi: 10.1145/2414782.2414787.