Algorithms of the LDA model

Jaka Špeh, Andrej Muhič, Jan Rupnik
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: {jaka.speh, andrej.muhic, jan.rupnik}@ijs.si

ABSTRACT

We review three algorithms for Latent Dirichlet Allocation (LDA). Two of them are variational inference algorithms: Variational Bayesian inference and Online Variational Bayesian inference and one is Markov Chain Monte Carlo (MCMC) algorithm – Collapsed Gibbs sampling. We compare their time complexity and performance. We find that online variational Bayesian inference is the fastest algorithm and still returns reasonably good results.

1 INTRODUCTION

Nowadays big corpora are used daily. People often search through huge numbers of documents either in libraries or online, using web search engines. Therefore, we need efficient algorithms that enable us efficient information retrieval.

Sometimes appropriate documents are hard to find, especially if you do not have the exact title. One solution is to search using keywords. As many documents do not have keywords, we need to know what certain document is talking about. Therefore, we would like to tag documents with appropriate keywords by clustering them according to their topics.

As the size of corpus increases, manual annotation is not an option. We would like that computers process documents and find their topics automatically. That can be done using machine learning.

Probabilistic graphical models such as Latent Dirichlet Allocation (LDA) allow us to describe a document in terms of probabilistic distributions over topics, and these topics in terms of distributions over words. In order to obtain documents topics and corpus topics (distributions over words), we need to compute posterior distribution. Unfortunately, the posterior is intractable to compute and one must appeal to approximate posterior inference.

Modern approximate posterior inference algorithms fall into two categories: sampling approaches and optimization approaches. Sampling approaches are usually based on MCMC sampling. Conceptual idea of the methods is to generate independent samples from posterior and then reason about documents and corpus topics. Whereas optimization approaches are usually based on variational inference, also called Variational Bayes (VB) for Bayesian models. Variational Bayes methods optimize closeness, in Kullback-Leibler divergence, of simplified parametric distribution to the posterior.

In this paper, we compare one MCMC and two VB algorithms for approximating posterior distribution. In the subsequent sections, we formally introduce LDA model and algorithms. We study performance of algorithms and make comparisons between them. For training and testing set we use articles from Wikipedia. We show that Online Variational Bayesian inference is the fastest algorithm. However the accuracy is lower than in the other two, but the results are still good enough for practical use.

2 LDA MODEL

Latent Dirichlet Allocation [1] is a Bayesian probabilistic graphical model, which is regularly used in topic modeling. It assumes M documents are build in a following fashion. First, a collection of K topics (distributions over words) are drawn from a Dirichlet distribution, $\varphi_k \sim \text{Dirichlet}(\beta)$. Then for m-th document, we:

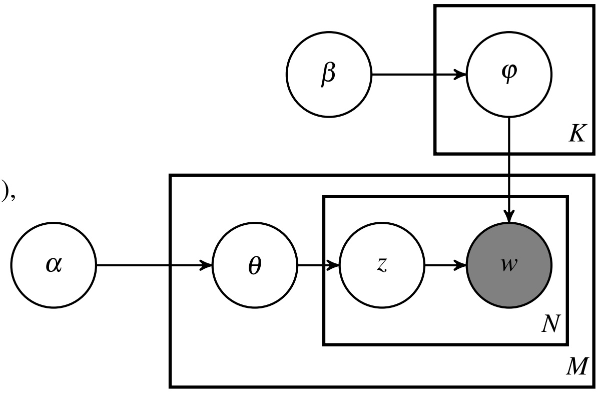

FIGURE 1. Plate notation of LDA.

1. Choose a topic distribution $\theta_m \sim \text{Dirichlet}(\alpha)$.
2. For each word $w_{m,n}$ in m-th document:
   i. choose a topic of the word $z_{m,n} \sim \text{Multinomial}(\theta_m)$,
   ii. choose a word $w_{m,n} \sim \text{Multinomial}(\varphi_{z_{m,n}})$.

LDA can be graphically presented using plate notation (Figure 1).

Probability of the LDA model is

$$p(\mathbf{w}, \mathbf{z}, \theta, \varphi \mid \alpha, \beta) = \prod_{k=1}^{K} p(\varphi_k|\beta) \prod_{m=1}^{M} \left(p(\theta_m|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) p(w_{m,n}|\varphi_{z_{m,n}}) \right).$$

We can analyse a corpus of documents by computing the posterior distribution of the hidden variables $(\mathbf{z}, \theta, \varphi)$ given a document $(\mathbf{w})$. This posterior reveals latent structure in the corpus that can be used for prediction or data exploration. Unfortunately, this distribution cannot be computed directly [1], and is usually approximated using Markov Chain Monte Carlo (MCMC) methods or variational inference.

3 ALGORITHMS

In the following subsections, we will derive one MCMC algorithm and two variational Bayes algorithms for the approximation of the posterior inference.

3.1 Collapsed Gibbs sampling

In the Collapsed Gibbs sampling we first integrate θ and φ out.

$$p(\mathbf{z}, \mathbf{w} \mid \alpha, \beta) = \int_{\theta} \int_{\varphi} p(\mathbf{z}, \mathbf{w}, \theta, \varphi \mid \alpha, \beta)\, d\theta\, d\varphi.$$

The goal of Collapsed Gibbs sampling here is to approximate the distribution $p(\mathbf{z} \mid \mathbf{w}, \alpha, \beta)$. Conditional probability $p(\mathbf{w} \mid \alpha, \beta)$ does not depend on $\mathbf{z}$, therefore Gibbs sampling equations can be derived from $p(\mathbf{z}, \mathbf{w} \mid \alpha, \beta)$ directly. Specifically, we are interested in the following conditional probability

$$p(z_{m,n} \mid \mathbf{z}_{\neg(m,n)}, \mathbf{w}, \alpha, \beta),$$

where $\mathbf{z}_{\neg(m,n)}$ denotes all z-s but $z_{m,n}$. Note that for Collapsed Gibbs sampling we need only to sample a value for $z_{m,n}$ according to the above probability. Thus we only need the probability mass function up to scalar multiplication. So, the distribution can be simplified [4, page 22] as:

$$(1) \qquad p(z_{m,n} = k \mid \mathbf{z}_{\neg(m,n)}, \mathbf{w}, \alpha, \beta) \propto \frac{n^{(v)}_{k,\neg(m,n)} + \beta}{\sum_{v=1}^{V} (n^{(v)}_{k,\neg(m,n)} + \beta)} \, (n^{(k)}_{m,\neg(m,n)} + \alpha),$$

where $n^{(v)}_{k}$ refers to the number of times that term v has been observed with topic k, $n^{(k)}_{m}$ refers to the number of times that topic k has been observed with a word of document m, and $n^{(\cdot)}_{\cdot,\neg(m,k)}$ indicate that the n-th token in m-th document is excluded from the corresponding $n^{(v)}_{k}$ or $n^{(k)}_{m}$.

Corpus and document topics can be obtained by [4, page 23]:

$$\varphi_{k,v} = \frac{n^{(t)}_{k} + \beta}{\sum_{v=1}^{V} (n^{(t)}_{k} + \beta)}, \quad \theta_{m,k} = \frac{n^{(k)}_{m} + \alpha}{\sum_{k=1}^{K} (n^{(k)}_{m} + \alpha)}.$$

In Collapsed Gibbs sampling algorithm, we need to remember values of three variables: $z_{m,n}$, $n^{(k)}_{m}$, and $n^{(v)}_{k}$, and some sums of these variables for efficiency. The algorithm first initializes $\mathbf{z}$ and computes $n^{(k)}_{m}$, $n^{(v)}_{k}$ according to the initialized values. Then in one iteration of the algorithm we go over all words of all documents, sample values of $z_{m,n}$ according to Equation (1), and recompute $n^{(k)}_{m}$, $n^{(v)}_{k}$. Then one has to decide when (from which iteration/s) to take a sample or samples and which criteria to choose to check if Markov chain has converged.

3.2 Variational Bayesian inference

This algorithm was proposed in the original LDA paper [1].

In Variational Bayesian inference (VB) the true posterior is approximated by a simpler distribution $q(\mathbf{z}, \theta, \phi)$, which is indexed by a set of free parameters [6]. We choose a fully factorized distribution q of the form

$$q(z_{m,n} = k) = \psi_{m,n,k},$$
$$q(\theta_m) = \text{Dirichlet}(\theta_m \mid \gamma_m),$$
$$q(\varphi_k) = \text{Dirichlet}(\varphi_k \mid \lambda_k).$$

The posterior is parameterized by ψ, γ and λ. We refer to λ as corpus topics and γ as documents topics.

The parameters are optimized to maximize the Evidence Lower Bound (ELBO):

$$(2) \quad \log p(\mathbf{w} \mid \alpha, \beta) \geq \mathscr{L}(\mathbf{w}, \psi, \gamma, \lambda) = \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{z}, \theta, \varphi \mid \alpha, \beta)] - \mathbb{E}_q[\log q(\mathbf{z}, \theta, \varphi)].$$

Maximizing the ELBO is equivalent to minimizing the Kullback-Leibler divergence between $q(\mathbf{z}, \theta, \varphi)$ and the posterior $p(\mathbf{z}, \theta, \varphi \mid \mathbf{w}, \alpha, \beta)$.

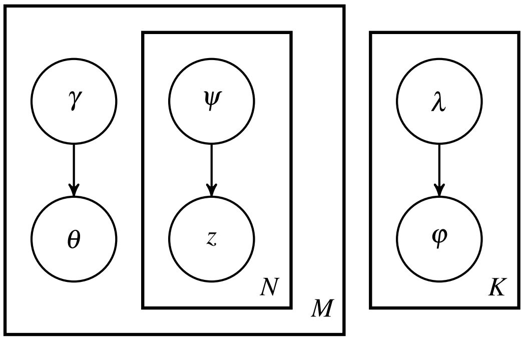

FIGURE 2. Plate notation of parameterized distribution q.

ELBO $\mathscr{L}$ can be optimized using coordinate ascent over the variational parameters (detailed derivation in [1, 2]):

(3) $\psi_{m,v,k} \propto \exp\left\{\mathbb{E}_q[\log\theta_{m,k}] + \mathbb{E}_q[\log\varphi_{k,v}]\right\}$,

(4) $\gamma_{m,k} = \alpha + \sum_{v=1}^{V} n_{m,v}\psi_{m,v,k}$,

(5) $\lambda_{k,v} = \beta + \sum_{m=1}^{M} n_{m,v}\psi_{m,v,k}$,

where $n_{m,v}$ is the number of terms v in document m. The expectations are

$$\mathbb{E}_q[\log\theta_{m,k}] = \Psi(\gamma_{m,k}) - \Psi\left(\sum_{\widetilde{k}=1}^{K}\gamma_{m,\widetilde{k}}\right),$$
$$\mathbb{E}_q[\log\varphi_{k,v}] = \Psi(\lambda_{k,v}) - \Psi\left(\sum_{\widetilde{v}=1}^{V}\lambda_{k,\widetilde{v}}\right),$$

where Ψ denotes the digamma function (the first derivative of the logarithm of the gamma function).

The updates of the variational parameters are guaranteed to converge to a stationary point of the ELBO. We can make some parallels with Expectation-Maximization (EM) algorithm [3]. Iterative updates of γ and ψ until convergence, holding λ fixed, can be seen as "E"-step, and updates of λ, given γ and ψ, can be seen as "M"-step.

The Variational Bayesian inference algorithm first initializes λ randomly. Then for each documents does the "E"-step: initializes γ randomly and then until γ converges does the coordinate ascend using Equations (3) and (4). After γ converges, the algorithm performs the "M"-step: sets λ using Equation (4). Each combination "E" and "M"-step improves ELBO. Variational Bayesian inference finishes after relative improvement of $\mathscr{L}$ is less than a preprescribed limit or after we reach maximum number of iterations. We define an iteration as "E" + "M"-step.

After algorithm finishes γ represents documents topics and λ represents corpus topics.

3.3 Online Variational Bayesian inference

Previously described algorithm has constant memory requirements. It requires full pass through the entire corpus each iteration. Therefore, it is not naturally suited when new data is constantly arriving. We would like an algorithm that gets the data, calculates the data topics and updates the existing corpus topics.

Let us modify previous algorithm and make desired one. First, we factorize ELBO (Equation (2)) into:

$$\begin{aligned}&\mathscr{L}(\mathbf{w},\psi,\gamma,\lambda) = \\ &\quad\sum_{m=1}^{M}\left\{\mathbb{E}_q[\log p(w_m \mid \theta_m, z_m, \varphi)] + \mathbb{E}_q[\log p(z_m \mid \theta_m)]\right.\\ &\quad - \mathbb{E}_q[\log q(z_m)] + \mathbb{E}_q[\log p(\theta_m \mid \alpha)] - \mathbb{E}_q[\log q(\theta_m)]\\ &\quad \left.+ \left(\mathbb{E}_q[\log p(\varphi \mid \beta)] - \mathbb{E}_q[\log q(\varphi)]\right)/M\right\}.\end{aligned}$$

Note that we bring the per corpus topics terms into the summation over documents, and divide them by the number of documents M. This allows us to look at the maximization of the ELBO according to the parameters ψ and γ for each document individually. Therefore, we first maximize ELBO according to the ψ and γ as in previous algorithm with λ fixed. Then we choose such λ for which the ELBO is as high as possible. Let $\gamma(w_m,\lambda)$ and $\psi(w_m,\lambda)$ be the values of γ_m and ψ_m produced by the "E"-step. Our goal is to find λ that maximizes

$$\mathscr{L}(\mathbf{w},\lambda) = \sum_{m=1}^{M}\ell_m\left(w_m,\gamma(w_m,\lambda),\psi(w_m,\lambda),\lambda\right),$$

where $\ell_m(w_m,\gamma(w_m,\lambda),\psi(w_m,\lambda),\lambda)$ is the m-th document's contribution to ELBO.

Then we compute $\widetilde{\lambda}$, the setting of λ that would be optimal with given ψ if our entire corpus consisted of the single document w_m repeated M times:

$$\widetilde{\lambda}_{k,v} = \beta + M n_{m,v}\psi_{m,v,k}.$$

Here M is the number of available documents, the size of the corpus. Then we update λ using convex combination of its previous value and $\widetilde{\lambda}$: $\lambda = (1-\rho_m)\lambda + \rho_m\widetilde{\lambda}$, where the weight is $\rho_m = (\tau_0 + m)^{-\kappa}$. Unknowns have special meaning: $\tau_0 \geq 0$ slows down the early iteration and κ controls the rate at which old values $\widetilde{\lambda}$ are forgotten.

To sum up. The algorithm firstly initializes λ randomly. Then, on a given document, performs "E"-step as in Variational Bayesian inference. Next it updates λ as discussed above. Finally it moves on the new document and repeats everything. The algorithm terminates after all documents are processed.

This algoritem is called Online Variational Bayesian inference (Online VB) and was proposed by Hofffman, Blei and Bach in [5].

4 EXPERIMENTS

We ran several experimets to evaluate algorithms of the LDA model. Our purpose was to compare the time complexity and performance of previously described algorithms. For training and testing corpora we used Wikipedia.

Efficiency was measured by using perplexity on held-out data, which is defined as

$$\text{perplexity}(\mathbf{w}_{\text{test}}, \lambda) = \exp\left\{ -\frac{\sum_{m=1}^{M} \log p(\mathbf{w}_m \mid \lambda)}{\sum_{m=1}^{M} N_m} \right\},$$

where N_m denotes number of words in m-th document. Since we cannot directly compute $\log p(\mathbf{w}_m \mid \lambda)$, we use ELBO as approximation:

$$\begin{aligned} &\text{perplexity}(\mathbf{w}_{\text{test}}, \lambda) \\ &\leq \exp\{-\textstyle\sum_{m=1}^{M} (\mathbb{E}_q[\log p(\mathbf{w}_m, \mathbf{z}_m, \theta_m \mid \varphi)] \\ &\quad - E_q[\log q(\mathbf{z}_m, \theta_m \mid \varphi)]) / \textstyle\sum_{m=1}^{M} N_m\}. \end{aligned}$$

We tested three algorithms and ran experiments for 10.000, 20.000, ..., 80.000 documents as a training set for corpora topics. Later we evaluated perplexity on 100 held-out documents. Size of vocabulary was around 150.000 words.

In all experiments α and β are fixed at 0.01 and the number of topics K is equal to 100. For Collapsed Gibbs sampling, no experiment converged. The criteria was relative change in $\mathbf{z}$ variable; change did not get under 20% in 1000 iterations.

In Variational Bayesian inference the "E"-step and the "M"-step converge if relative change in γ is under 0.001 and relative improvement of the ELBO is under 0.001, respectively. If there is no convergence, we terminate after 100 iterations for both "E" and "M"-step. However algorithm always converged in less than 20 iterations.

In Online Variational Bayesian inference limit for the "E"-step was the same as in Variational Bayesian inference. Batchsize was 100 documents, τ_0 was 1024 and κ was equal to 0.7 as proposed in [5].

The fastest algorithm is Online VB, other two have similar time complexity with a large note: VB algorithm converged every time while Gibbs sampling algorithm did not converge. Unexpected, Online VB does not perform as well as other two but in practice still gives reasonably good results. Our future goal is to explain the results obtained by experiments.

Therefore we recommend Online VB algorithm for practical use, if time is a factor.

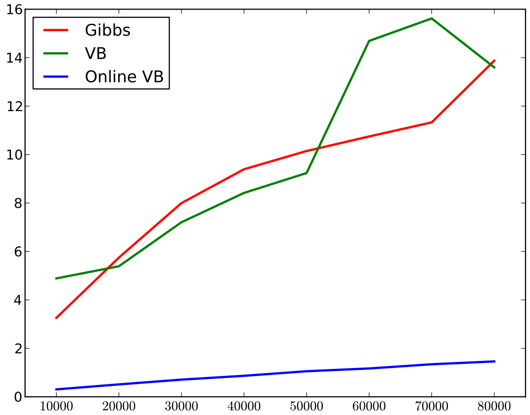

FIGURE 3. Time used by the algorithms (in hours) given the number of the documents.

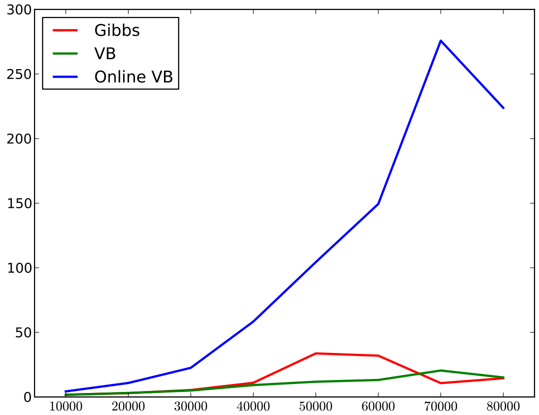

FIGURE 4. Perplexity on held-out documents as a function of number of documents analyzed.

5 ACKNOWLEDGMENT

The authors gratefully acknowledge that the funding for this work was provided by the project XLIKE (ICT-257790-STREP).

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Wim De Smet and Marie-Francine Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, SWSM '09, pages 57–64, New York, NY, USA, 2009. ACM.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[4] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, Darmstadt, Germany, 2005.

[5] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864. Curran Associates, Inc., 2010.

[6] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999.