A Simple Stochastic Gradient Variational Bayes for Latent Dirichlet Allocation

Tomonari Masada[1(✉)] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki, Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper proposes a new inference for the latent Dirichlet allocation (LDA) [4]. Our proposal is an instance of the stochastic gradient variational Bayes (SGVB) [9,13]. SGVB is a general framework for devising posterior inferences for Bayesian probabilistic models. Our aim is to show the effectiveness of SGVB by presenting an example of SGVB-type inference for LDA, the best-known Bayesian model in text mining. The inference proposed in this paper is easy to implement from scratch. A special feature of the proposed inference is that the logistic normal distribution is used to approximate the true posterior. This is counterintuitive, because we obtain the Dirichlet distribution by taking the functional derivative when we lower bound the log evidence of LDA after applying a mean field approximation. However, our experiment showed that the proposed inference gave a better predictive performance in terms of test set perplexity than the inference using the Dirichlet distribution for posterior approximation. While the logistic normal is more complicated than the Dirichlet, SGVB makes the manipulation of the expectations with respect to the posterior relatively easy. The proposed inference was better even than the collapsed Gibbs sampling [6] for not all but many settings consulted in our experiment. It must be worthwhile future work to devise a new inference based on SGVB also for other Bayesian models.

Keywords: Text mining · Topic models · variational Bayesian inference

1 Introduction

When we use Bayesian probabilistic models for data mining applications, we need to infer the posterior distribution. While the Markov Chain Monte Carlo (MCMC) is undoubtedly important [5,14], this paper focuses on the variational Bayesian inference (VB). Therefore, we first present an outline of VB.

Let $\boldsymbol{x}$ be a set of the random variables whose values are observed. A probabilistic model for analyzing the observed data $\boldsymbol{x}$ can be specified unambiguously by its full joint distribution $p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Theta})$, where $\boldsymbol{z}$ denotes the discrete latent variables and $\boldsymbol{\Theta}$ the continuous ones. In VB, we maximize the log of the evidence $p(\boldsymbol{x})$, which is obtained from the full joint distribution by marginalizing $\boldsymbol{z}$ and

O. Gervasi et al. (Eds.): ICCSA 2016, Part IV, LNCS 9789, pp. 232–245, 2016.
DOI: 10.1007/978-3-319-42089-9_17

$\boldsymbol{\Theta}$ out, i.e., $p(\boldsymbol{x}) = \int \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}$. However, the maximization of $\log p(\boldsymbol{x})$ is generally intractable. Therefore, we instead maximize its lower bound:

$$\log p(\boldsymbol{x}) = \log \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\Theta}) \frac{p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Theta})}{q(\boldsymbol{z}, \boldsymbol{\Theta})} d\boldsymbol{\Theta} \geq \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\Theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Theta})}{q(\boldsymbol{z}, \boldsymbol{\Theta})} d\boldsymbol{\Theta}. \quad (1)$$

We have introduced an approximate posterior $q(\boldsymbol{z}, \boldsymbol{\Theta})$ in Eq. (1) to apply Jensen's inequality. If we put the true posterior $p(\boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{x})$ in place of the approximate posterior, Jensen's inequality holds with equality. However, the true posterior is typically intractable. Therefore, in VB, the inference of the approximate posterior is the main task.

In this paper, we consider the latent Dirichlet allocation (LDA) [4], the best-known Bayesian model in text mining, as our target. LDA and its extensions have been applied to solve a wide variety of text mining problems [7,10,12,15,17]. It is known that the performance of LDA measured in terms of test set perplexity heavily depends on how the inference is conducted [2]. Therefore, to provide a new proposal relating to the posterior inference for LDA is highly relevant to text mining research.

The main contribution of this paper is to propose a new VB-type inference for LDA. The proposed inference was better than the VB presented in [4] in terms of test set perplexity in all situations consulted by our experiment. For brevity, we call the VB presented in [4] standard.[1]

In the standard VB, the true posterior distribution is approximated by the Dirichlet distribution. This is because the Dirichlet is obtained analytically by taking the functional derivative after applying a mean field approximation. It has been shown experimentally that the standard VB works as well as other inference methods [2]. In our proposed inference, we apply the same mean field approximation. However, we do not use the Dirichlet for approximating the true posterior. Nevertheless, the proposed inference could achieve a better perplexity than the standard VB in our experiment. Interestingly, our method was better even than the collapsed Gibbs sampling (CGS) [6] in not all but many situations.

The proposed inference for LDA is based on the stochastic gradient variational Bayes (SGVB), which has been proposed by the two papers [9,13] almost simultaneously. SGVB can be regarded as a general framework for obtaining VB-type inferences for a wide range of Bayesian probabilistic models. Precisely, SGVB provides a general framework for obtaining a Monte Carlo estimate of the log-evidence lower bound, i.e., the lower bound of $\log p(\boldsymbol{x})$ in Eq. (1). In this paper, we utilize SGVB to devise an inference easy to implement for LDA. We use the logistic normal distribution [1] for approximating the true posterior. While the logistic normal is more complicated than the Dirichlet, we can obtain a simple VB-type inference owing to SGVB.

[1] Precisely speaking, the VB presented in [4] performs a point estimation for the per-topic word multinomial distributions. In the VB we call standard here, a Bayesian inference is performed also for the per-topic word multinomial distributions, not only for the per-document topic multinomial distributions.

In the next section, we describe SGVB based on [9], which gives an explanation easy to understand for those familiar with LDA. Our description does not cover the full generality of SGVB, partly because we focus only on LDA as our target. We then provide the details of our proposal in Sect. 3. Section 4 presents the results of an evaluation experiment, where we compared our proposal with other methods including the standard VB and CGS. Section 5 concludes the paper with discussion on worthwhile future work.

2 Stochastic Gradient Variational Bayes

The log-evidence lower bound in Eq. (1) can be rewritten as follows:

$$\mathcal{L}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Theta})] - \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\log q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})], \tag{2}$$

where $\boldsymbol{\Lambda}$ denotes the parameters of the approximate posterior $q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})$, and $\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\cdot]$ denotes the expectation with respect to $q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})$. We assume that $q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})$ factorizes as $q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})$. Then we can write $\mathcal{L}(\boldsymbol{\Lambda})$ as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Lambda}) =& \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Theta})] \\ &- \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})}[\log q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})] - \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})}[\log q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})].\end{aligned} \tag{3}$$

Our task is to estimate the expectations on the right hand side of Eq. (3).

In this paper, we estimate the log-evidence lower bound of the latent Dirichlet allocation (LDA) by using the stochastic gradient variational Bayes (SGVB) [9,13]. SGVB is a general framework for obtaining a Monte Carlo estimate of the log-evidence lower bound for a wide variety of Bayesian probabilistic models. Note that SGVB cannot provide an estimate of the expectation with respect to the distribution for the discrete random variables [11]. However, we can perform an estimation as in the standard VB for LDA [4] with resect to $\boldsymbol{z}$.

In SGVB, we can assume that the approximate posterior $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})$ depends on the observed data $\boldsymbol{x}$. We do not consider this option here and thus do not explore the full generality of SGVB. However, by making the approximate posterior not dependent on $\boldsymbol{x}$, we can make the proposed inference simple.

When we apply SGVB, the approximate posterior should meet at least two requirements. SGVB estimates the expectations for the continuous variables $\boldsymbol{\Theta}$ by the Monte Carlo method. Therefore, the approximate posterior $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})$ should be a distribution from which we can draw samples. This is the first requirement. Let $\boldsymbol{\Theta}^{(l)}$, $l = 1, \ldots, L$ be the samples drawn from $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})$. Then $\mathcal{L}(\boldsymbol{\Lambda})$ in Eq. (3) is estimated as

$$\begin{aligned}\hat{\mathcal{L}}(\boldsymbol{\Lambda}) =& \frac{1}{L}\sum_{l=1}^{L}\left\{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})}[\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Theta}^{(l)})] - \log q(\boldsymbol{\Theta}^{(l)}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})\right\} \\ &- \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})}[\log q(\boldsymbol{z}|\boldsymbol{\Lambda}_{\boldsymbol{z}})].\end{aligned} \tag{4}$$

In SGVB, we maximize $\hat{\mathcal{L}}(\boldsymbol{\Lambda})$ in Eq. (4) in place of $\mathcal{L}(\boldsymbol{\Lambda})$ in Eq. (3). To maximize $\hat{\mathcal{L}}(\boldsymbol{\Lambda})$, we need to obtain its derivatives with respect to the relevant variables.

Therefore, $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})$ should be a distribution that makes $\hat{\mathcal{L}}(\boldsymbol{\Lambda})$ differentiable. This is the second requirement. As described below, the inference proposed in this paper for LDA can be regarded as an example of SGVB.

3 Our Proposal

3.1 Lower Bound Estimation

We first describe LDA. Let D, K, and V denote the numbers of documents, latent topics, and vocabulary words, respectively. The parameters of the per-document topic multinomial distributions and the parameters of the per-topic word multinomial distributions are represented as $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ for $d = 1, \ldots, D$ and $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV})$ for $k = 1, \ldots, K$, respectively. Then the full joint distribution of LDA is written as follows:

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) &= \prod_{d=1}^{D} p(\boldsymbol{x}_d|\boldsymbol{z}_d, \boldsymbol{\phi}) p(\boldsymbol{z}_d|\boldsymbol{\theta}_d) \cdot \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \cdot \prod_{k=1}^{K} p(\boldsymbol{\phi}_k|\beta) \\
&= \prod_{d=1}^{D} \prod_{i=1}^{N_d} \phi_{z_{di} x_{di}} \theta_{d z_{di}} \cdot \prod_{d=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \cdot \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{v=1}^{V} \phi_{kv}^{\beta-1}, \qquad (5)
\end{aligned}
$$

where N_d is the length of the dth document. x_{di} is an observed variable whose value is the vocabulary word appearing as the ith token of the dth document. z_{di} is a latent variable whose value is the topic to which the ith word token of the dth document is assigned. The notation $\phi_{z_{di} x_{di}}$ is equivalent to ϕ_{kv} when $x_{di} = v$ and $z_{di} = k$. α and β are the hyperparameters of the symmetric Dirichlet priors for $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$, respectively.

We propose a new inference method for LDA based on SGVB explained in Sect. 2. However, SGVB is applicable only to the continuous latent variables. Therefore, in LDA, SGVB works only for the $\boldsymbol{\theta}_d$s and the $\boldsymbol{\phi}_k$s.

With the mean field approximation $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \approx \prod_{d=1}^{D} \prod_{i=1}^{N_d} q(z_{di}) \cdot \prod_{d=1}^{D} q(\boldsymbol{\theta}_d) \cdot \prod_{k=1}^{K} q(\boldsymbol{\phi}_k)$, the lower bound of $\log p(\boldsymbol{x})$ (cf. Eq. (3)) is obtained as follows:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\Lambda}) \\
&= \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di}) q(\boldsymbol{\phi}_{z_{di}})} \big[\log p(x_{di}|z_{di}, \boldsymbol{\phi}_{z_{di}}) \big] + \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di}) q(\boldsymbol{\theta}_d)} \big[\log p(z_{di}|\boldsymbol{\theta}_d) \big] \\
&+ \sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{\theta}_d)} \big[\log p(\boldsymbol{\theta}_d|\alpha) \big] + \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\phi}_k)} \big[\log p(\boldsymbol{\phi}_k|\beta) \big] \\
&- \sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{\theta}_d)} \big[\log q(\boldsymbol{\theta}_d) \big] - \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\phi}_k)} \big[\log q(\boldsymbol{\phi}_k) \big] - \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di})} \big[\log q(z_{di}) \big]. \qquad (6)
\end{aligned}
$$

In the standard VB [4], we obtain the approximate posteriors by a functional derivative method after using the mean field approximation given above. The

result is that the posterior $q(\boldsymbol{\theta}_d)$ for each d and the posterior $q(\boldsymbol{\phi}_k)$ for each k are a Dirichlet distribution. However, it is one thing that approximate posteriors can be found analytically by a functional derivative method, and it is a different thing that such approximate posteriors lead to a good evaluation result in terms of test set perplexity. Therefore, we can choose a distribution other than the Dirichlet for approximating the true posterior.

In this paper, we propose to use the logistic normal distribution [1] for approximating the true posterior. We define θ_{dk} and ϕ_{kv} with the samples $\epsilon_{\theta,dk}$ and $\epsilon_{\phi,kv}$ from the standard normal distribution $\mathcal{N}(0,1)$ as follows:

$$\theta_{dk} \equiv \frac{\exp(\epsilon_{\theta,dk}\sigma_{\theta,dk} + \mu_{\theta,dk})}{\sum_{k'=1}^{K} \exp(\epsilon_{\theta,dk'}\sigma_{\theta,dk'} + \mu_{\theta,dk'})} \text{ and}$$
$$\phi_{kv} \equiv \frac{\exp(\epsilon_{\phi,kv}\sigma_{\phi,kv} + \mu_{\phi,kv})}{\sum_{v'=1}^{V} \exp(\epsilon_{\phi,kv'}\sigma_{\phi,kv'} + \mu_{\phi,kv'})}. \tag{7}$$

Note that $\epsilon\sigma + \mu \sim \mathcal{N}(\mu, \sigma)$ when $\epsilon \sim \mathcal{N}(0,1)$. μ and σ in Eq. (7) are the mean and standard deviation parameters of the logistic normal. Equation (7) gives the reparameterization trick [9] in our case, where we assume that the covariance matrix of the logistic normal is diagonal to make the inference simple.

We can draw $\boldsymbol{\theta}_d \sim \text{LogitNorm}(\boldsymbol{\mu}_{\theta,d}, \boldsymbol{\sigma}_{\theta,d})$ and $\boldsymbol{\phi}_k \sim \text{LogitNorm}(\boldsymbol{\mu}_{\phi,k}, \boldsymbol{\sigma}_{\phi,k})$ efficiently based on Eq. (7). Therefore, the first requirement given in Sect. 2 is met. For the approximate posterior $q(\boldsymbol{z})$, we assume as in the standard VB that we have a different discrete distribution $\text{Discrete}(\boldsymbol{\gamma}_{di})$ for each word token x_{di}, where γ_{dik} is the probability that $z_{di} = k$ holds, i.e., the probability that the ith token of the dth document is assigned to the kth topic.

However, the algebraic manipulation of the expectation with respect to the logistic normal distribution is highly complicated. Here SGVB has an advantage, because it estimates the expectations with respect to approximate posteriors by the Monte Carlo method. $\mathcal{L}(\boldsymbol{\Lambda})$ in Eq. (6) is estimated with L samples $\boldsymbol{\theta}_d^{(l)} \sim \text{LogitNorm}(\boldsymbol{\mu}_{\theta,d}, \boldsymbol{\sigma}_{\theta,d})$ and $\boldsymbol{\phi}_k^{(l)} \sim \text{LogitNorm}(\boldsymbol{\mu}_{\phi,k}, \boldsymbol{\sigma}_{\phi,k})$ for $l = 1, \ldots, L$ as

$$\begin{aligned}
\hat{\mathcal{L}}(\boldsymbol{\Lambda}) = {} & \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di})}\left[\log p(x_{di}|z_{di}, \boldsymbol{\phi}_{z_{di}}^{(l)})\right] \\
& + \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di})}\left[\log p(z_{di}|\boldsymbol{\theta}_d^{(l)})\right] - \sum_{d=1}^{D}\sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di})}\left[\log q(z_{di})\right] \\
& + \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{\theta}_d)}\left[\log p(\boldsymbol{\theta}_d^{(l)}|\alpha) - \log q(\boldsymbol{\theta}_d^{(l)})\right] \\
& + \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\phi}_k)}\left[\log p(\boldsymbol{\phi}_k^{(l)}|\beta) - \log q(\boldsymbol{\phi}_k^{(l)})\right].
\end{aligned} \tag{8}$$

We maximize the above estimate, denoted as $\hat{\mathcal{L}}(\boldsymbol{\Lambda})$, in place of $\mathcal{L}(\boldsymbol{\Lambda})$ in Eq. (6).

Due to the limit of space, we only discuss the first two expectation terms of the right hand side of Eq. (8). The first term can be rewritten as follows:

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathbb{E}_{q(z_{di})}\left[\log p(x_{di}|z_{di},\boldsymbol{\phi}_{z_{di}}^{(l)})\right] = \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}\log\phi_{kx_{di}}^{(l)}$$
$$= \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}\log\left\{\frac{\exp(\epsilon_{\phi,kx_{di}}^{(l)}\sigma_{\phi,kx_{di}}+\mu_{\phi,kx_{di}})}{\sum_{v'=1}^{V}\exp(\epsilon_{\phi,kv'}^{(l)}\sigma_{\phi,kv'}+\mu_{\phi,kv'})}\right\} \quad \text{(cf. Eq. 7)}$$
$$= \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}(\epsilon_{\phi,kx_{di}}^{(l)}\sigma_{\phi,kx_{di}}+\mu_{\phi,kx_{di}})$$
$$-\frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}\log\left\{\sum_{v=1}^{V}\exp\left(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv}\right)\right\}, \tag{9}$$

where we use the definition of ϕ_{kv} in Eq. (7). The summation term on the last line in Eq. (9) can be upper bounded by using the Taylor expansion [3]:

$$\log\sum_{v}\exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv}) \leq \frac{\sum_{v}\exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv})}{\eta_{\phi,k}^{(l)}} - 1 + \log\eta_{\phi,k}^{(l)}, \tag{10}$$

where we have introduced a new variational parameter $\eta_{\phi,k}^{(l)}$. Consequently, we can lower bound Eq. (9) as follows:

$$\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\gamma})}\left[\log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\phi}^{(l)})\right] \geq \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}(\epsilon_{\phi,kx_{di}}^{(l)}\sigma_{\phi,kx_{di}}+\mu_{\phi,kx_{di}})$$
$$-\frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}\left\{\frac{\sum_{v}\exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv})}{\eta_{\phi,k}^{(l)}}+\log\eta_{\phi,k}^{(l)}-1\right\}. \tag{11}$$

Let us define $N_{kv} \equiv \sum_{d=1}^{D}\sum_{i=1}^{N_d}\gamma_{dik}\delta(x_{di}=v)$, where $\delta(\cdot)$ is an indicator function that evaluates to 1 if the condition in parentheses holds and to 0 otherwise. N_{kv} means how many tokens of the vth vocabulary word are assigned to the kth topic in expectation. Further, we define $N_k \equiv \sum_{v=1}^{V}N_{kv}$. Then Eq. (11) can be presented more neatly:

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathbb{E}_{q(z_{di})}\left[\log p(x_{di}|z_{di},\boldsymbol{\phi}_{z_{di}}^{(l)})\right] \geq \sum_{k=1}^{K}\sum_{v=1}^{V}N_{kv}(\sigma_{\phi,kv}\bar{\epsilon}_{\phi,kv}+\mu_{\phi,kv})$$
$$-\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}N_k\left\{\frac{\sum_{v}\exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv})}{\eta_{\phi,k}^{(l)}}+\log\eta_{\phi,k}^{(l)}-1\right\}, \tag{12}$$

where we define $\bar{\epsilon}_{\phi,kv} \equiv \frac{1}{L}\sum_{l=1}^{L}\epsilon_{\phi,kv}^{(l)}$.

In a similar manner, we can lower bound the second expectation term of the right hand side in Eq. (8) as follows:

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathbb{E}_{q(z_{di})}\left[\log p(z_{di}|\boldsymbol{\theta}_d^{(l)})\right] \geq \sum_{d=1}^{D}\sum_{k=1}^{K} N_{dk}(\sigma_{\theta,dk}\bar{\epsilon}_{\theta,dk}+\mu_{\theta,dk}) - \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D} N_d\left\{\frac{\sum_k \exp(\epsilon_{\theta,dk}^{(l)}\sigma_{\theta,dk}+\mu_{\theta,dk})}{\eta_{\theta,d}^{(l)}} + \log\eta_{\theta,d}^{(l)} - 1\right\}, \tag{13}$$

where $\eta_{\theta,d}$ is a new parameter introduced in a manner similar to Eq. (10). N_{dk} is defined as $\sum_{i=1}^{N_d}\gamma_{dik}$. N_{dk} means how many word tokens of the dth document are assigned to the kth topic in expectation.

We skip the explanation for other expectation terms in Eq. (8) and only show the final result. We can lower bound $\hat{\mathcal{L}}(\boldsymbol{\Lambda})$ as follows:

$$\begin{aligned}\hat{\mathcal{L}}(\boldsymbol{\Lambda}) \geq & \sum_{d,k}(N_{dk}+\alpha)(\sigma_{\theta,dk}\bar{\epsilon}_{\theta,dk}+\mu_{\theta,dk}) + \sum_{k,v}(N_{kv}+\beta)(\sigma_{\phi,kv}\bar{\epsilon}_{\phi,kv}+\mu_{\phi,kv}) \\ & - \frac{1}{L}\sum_{l=1}^{L}\sum_{d=1}^{D}(N_d+K\alpha)\left\{\frac{\sum_k \exp(\epsilon_{\theta,dk}^{(l)}\sigma_{\theta,dk}+\mu_{\theta,dk})}{\eta_{\theta,d}^{(l)}} + \log\eta_{\theta,d}^{(l)} - 1\right\} \\ & - \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}(N_k+V\beta)\left\{\frac{\sum_v \exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv})}{\eta_{\phi,k}^{(l)}} + \log\eta_{\phi,k}^{(l)} - 1\right\} \\ & + \sum_{k=1}^{K}\sum_{v=1}^{V}\log\sigma_{\phi,kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\log\sigma_{\theta,dk} - \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{dik}\log\gamma_{dik} \\ & + D\log\Gamma(K\alpha) - DK\log\Gamma(\alpha) + K\log\Gamma(V\beta) - KV\log\Gamma(\beta),\end{aligned} \tag{14}$$

where the constant term is omitted. We refer to the right hand side of Eq. (14) by $\tilde{\mathcal{L}}(\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ denotes the posterior parameters $\{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{\gamma}\}$. In our version of SGVB for LDA, we maximize $\tilde{\mathcal{L}}(\boldsymbol{\Lambda})$. Note that $\tilde{\mathcal{L}}(\boldsymbol{\Lambda})$ is differentiable with respect to all relevant variables owing to the parameterization trick in Eq. (7). Therefore, the second requirement given in Sect. 2 is met.

3.2 Maximization of Lower Bound

We maximize $\tilde{\mathcal{L}}(\boldsymbol{\Lambda})$, i.e., the right hand side of Eq. (14), by differentiating it with respect to the relevant variables. With respect to $\eta_{\theta,d}^{(l)}$, we obtain the derivative:

$$\frac{\partial\hat{L}(\boldsymbol{\Lambda})}{\partial\eta_{\theta,d}^{(l)}} = \frac{1}{L}(N_d+K\alpha)\left\{\frac{1}{\eta_{\theta,d}^{(l)}} - \frac{1}{(\eta_{\theta,d}^{(l)})^2}\sum_{k=1}^{K}\exp\left(\epsilon_{\theta,dk}^{(l)}\sigma_{\theta,dk}+\mu_{\theta,dk}\right)\right\}. \tag{15}$$

The equation $\frac{\partial\hat{L}(\boldsymbol{\Lambda})}{\partial\eta_{\theta,d}^{(l)}} = 0$ gives the solution: $\eta_{\theta,d}^{(l)} = \sum_{k=1}^{K}\exp(\epsilon_{\theta,dk}^{(l)}\sigma_{\theta,dk}+\mu_{\theta,dk})$. Similarly, $\frac{\partial\hat{L}(\boldsymbol{\Lambda})}{\partial\eta_{\phi,k}^{(l)}} = 0$ gives the solution: $\eta_{\phi,k}^{(l)} = \sum_{v=1}^{V}\exp(\epsilon_{\phi,kv}^{(l)}\sigma_{\phi,kv}+\mu_{\phi,kv})$. Note that each of these solutions appears as the denominator in Eq. (7).

Further, with respect to $\mu_{\theta,dk}$, we obtain the derivative:

$$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \mu_{\theta,dk}} = (N_{dk} + \alpha) - \frac{\exp(\mu_{\theta,dk})}{L}(N_d + K\alpha) \sum_{l=1}^{L} \frac{\exp(\epsilon_{\theta,dk}^{(l)} \sigma_{\theta,dk})}{\eta_{\theta,d}^{(l)}}. \tag{16}$$

Therefore, $\exp(\mu_{\theta,dk})$ is estimated as $\frac{N_{dk}+\alpha}{N_d+K\alpha} \cdot \frac{L}{\sum_l \exp(\epsilon_{\theta,dk}^{(l)} \sigma_{\theta,dk})/\eta_{\theta,d}^{(l)}}$. Based on the definition of $\theta_{dk}^{(l)}$ in Eq. (7), we obtain the following update for $\exp(\mu_{\theta,dk})$:

$$\exp(\mu_{\theta,dk}) \leftarrow \exp(\mu_{\theta,dk}) \cdot \left(\frac{N_{dk} + \alpha}{N_d + K\alpha}\right) \Big/ \left(\frac{\sum_l \theta_{dk}^{(l)}}{L}\right). \tag{17}$$

Similarly, $\exp(\mu_{\phi,kv})$ is updated as follows:

$$\exp(\mu_{\phi,kv}) \leftarrow \exp(\mu_{\phi,kv}) \cdot \left(\frac{N_{kv} + \beta}{N_k + V\beta}\right) \Big/ \left(\frac{\sum_l \phi_{kv}^{(l)}}{L}\right). \tag{18}$$

We can give an intuitive explanation to the update in Eq. (17). $\frac{N_{dk}+\alpha}{N_d+K\alpha}$ is an estimate of the per-document topic probability based on the word count expectation N_{dk}. In contrast, $\frac{\sum_l \theta_{dk}^{(l)}}{L}$ is an estimate of the same probability based on the logistic normal samples $\theta_{dk}^{(1)}, \ldots, \theta_{dk}^{(L)}$. We adjust $\exp(\mu_{\theta,dk})$ based on to what extent the latter estimate deviates from the former. A similar explanation can be given to the update in Eq. (18).

For the standard deviation parameters $\sigma_{\theta,dk}$ and $\sigma_{\phi,kv}$, we cannot obtain any closed form update. Therefore, we perform a gradient-based optimization. For numerical reasons, we change variables as $\tau = \log(\sigma^2)$. Then the derivatives with respect to $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ are obtained as follows:

$$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \tau_{\theta,dk}} = \frac{1}{2} + \frac{1}{2}\exp\left(\frac{\tau_{\theta,dk}}{2}\right) \frac{\sum_{l=1}^{L} \epsilon_{\theta,dk}^{(l)} \{(N_{dk} + \alpha) - (N_d + K\alpha)\theta_{dk}^{(l)}\}}{L}, \tag{19}$$

$$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \tau_{\phi,kv}} = \frac{1}{2} + \frac{1}{2}\exp\left(\frac{\tau_{\phi,kv}}{2}\right) \frac{\sum_{l=1}^{L} \epsilon_{\phi,kv}^{(l)} \{(N_{kv} + \beta) - (N_k + V\beta)\phi_{kv}^{(l)}\}}{L}. \tag{20}$$

With respect to γ_{dik}, we obtain the following derivative:

$$\begin{aligned}\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \gamma_{dik}} &= \frac{1}{L}\sum_{l=1}^{L}(\epsilon_{\theta,dk}^{(l)} \sigma_{\theta,dk} + \mu_{\theta,dk}) + \frac{1}{L}\sum_{l=1}^{L}(\epsilon_{\phi,kx_{di}}^{(l)} \sigma_{\phi,kx_{di}} + \mu_{\phi,kx_{di}}) \\ &- \frac{1}{L}\sum_{l=1}^{L}\left\{\frac{\sum_v \exp(\epsilon_{\phi,kv}^{(l)} \sigma_{\phi,kv} + \mu_{\phi,kv})}{\eta_{\phi,k}^{(l)}} + \log \eta_{\phi,k}^{(l)} - 1\right\} - \log \gamma_{dn,k} - 1.\end{aligned} \tag{21}$$

$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \gamma_{dik}} = 0$ gives the following update:

$$\gamma_{dik} \propto \left(\prod_{l=1}^{L} \theta_{dk}^{(l)}\right)^{\frac{1}{L}} \cdot \left(\prod_{l=1}^{L} \phi_{kx_{dn}}^{(l)}\right)^{\frac{1}{L}}. \tag{22}$$

Algorithm 1. An SGVB for LDA with logistic normal

```
 1: Split the document set into small batches
 2: for each iteration do
 3:     for each small batch do
 4:         for each topic k do
 5:             Draw φ_k^(l) ~ LogitNorm(μ_{φ,k}, σ_{φ,k}) for l = 1, ..., L
 6:             Update exp(μ_{φ,kv}) based on Eq. (18)
 7:             Update τ_{φ,kv} based on the gradient in Eq. (20)
 8:         end for
 9:         for each document d do
10:             Draw θ_d^(l) ~ LogitNorm(μ_{θ,d}, σ_{θ,d}) for l = 1, ..., L
11:             for i = 1, ..., N_d do
12:                 Update γ_{dik} based on Eq. (22)
13:                 Update N_{dk}, N_{kv}, and N_k
14:             end for
15:             Update exp(μ_{θ,dk}) based on Eq. (17)
16:             Update τ_{θ,dk} based on the gradient in Eq. (19)
17:         end for
18:     end for
19: end for
```

The right hand side of Eq. (22) is the product of the geometric mean of the sampled per-document topic probabilities $\theta_{dk}^{(l)}$ and the geometric mean of the sampled per-topic word probabilities $\phi_{kx_{dn}}^{(l)}$. Interestingly, other inference methods for LDA also represent the per-token topic probability as the product of the per-document topic probability and the per-topic word probability [2].

Algorithm 1 gives the pseudocode. As in CVB for LDA [16], we only need to maintain one copy of γ_{dik} for each unique document/word pair. Therefore, we can reduce the number of iterations of the loop on line 11 from N_d to the number of different words in the dth document. Consequently, the time complexity of the proposed inference for each scan of the data is $\mathcal{O}(MK)$, where M is the total number of unique document/word pairs. For updating $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ based on the gradients in Eqs. (19) and (20), we use Adam [8] in this paper.

3.3 Estimation Without Sampling

By setting all standard deviation parameters $\boldsymbol{\sigma}$ to 0, we obtain an estimate of the log-evidence lower bound without sampling as a degenerated version of our proposal. In this case, we only update the parameter γ_{dik} by

$$\gamma_{dik} \propto \left(\frac{N_{dk} + \alpha}{N_d + K\alpha}\right) \cdot \left(\frac{N_{kv} + \beta}{N_k + V\beta}\right). \tag{23}$$

This is almost the same with the update of γ_{dik} in CVB0 [2] except that the contribution of γ_{dik} is not subtracted from N_{dk}, N_{kv}, and N_k. In the evaluation experiment presented in the next section, we compared the proposed method also with this degenerated version to clarify the effect of the sampling.

Table 1. Specifications of the document sets

	# documents (D)	# vocabulary words (V)	# word tokens ($\sum_d N_d$)	Average document length ($\sum_d N_d/D$)
NYT	99,932	46,263	34,458,469	344.8
MOVIE	27,859	62,408	12,788,477	459.0
NSF	128,818	21,471	14,681,181	114.0
MED	125,490	42,830	17,610,749	140.3

4 Experiment

4.1 Data Sets

In the evaluation experiment, we used the four English document sets in Table 1. NYT is a part of the NYTimes news articles in "Bag of Words Data Set" of the UCI Machine Learning Repository.[2] We reduced the number of documents to one third of its original number due to the limit of the main memory. MOVIE is the set of movie reviews known as "Movie Review Data."[3] NSF is "NSF Research Award Abstracts 1990–2003 Data Set" of the UCI Machine Learning Repository. MED is a subset of the paper abstracts of the MEDLINE®/PUBMED®, a database of the U.S. National Library of Medicine.[4] For all document sets, we applied the Porter stemming algorithm and removed highly frequent words and extremely rare words. The average document lengths, i.e., $\sum_d N_d/D$, of NYT and MOVIE are 344.6 and 459.0, respectively. In contrast, those of NSF and MED are 114.0 and 140.3, respectively. This difference comes from the fact that NSF and MED consist of abstracts.

4.2 Evaluation Method

By using the above four data sets, we compared our proposal with the following three inference methods for LDA: the standard VB [4], CGS [6], and the degenerated version described in Sect. 3.3. The evaluation measure is the test set perplexity. We ran each of the compared inference methods on the 90 % word tokens randomly selected from each document and used the estimated parameters for computing the test set perplexity on the other 10 % word tokens as follows:

$$perplexity \equiv \exp\left\{ -\frac{1}{N_{\text{test}}} \sum_{d=1}^{D} \sum_{i \in \mathcal{I}_d} \log\left(\sum_{k=1}^{K} \theta_{dk} \phi_{k x_{di}} \right) \right\}, \tag{24}$$

where $\mathcal{I}_d$ is the set of the indices of the test word tokens in the dth document, and N_{test} is the total number of the test tokens. For K, i.e., the number of topics, we tested the following three settings: $K = 50, 100,$ and 200.

[2] https://archive.ics.uci.edu/ml/datasets.html.
[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip.
[4] We used the XML files from medline14n0770.xml to medline14n0774.xml.

4.3 Inference Settings

The proposed inference was run on each data set in the following manner. We tuned the free parameters on a random training/test split that was prepared only for validation. On lines 7 and 16 in Algorithm 1, $\tau_{\phi,kv}$ and $\tau_{\theta,dk}$ are updated by using the gradients. For this optimization, we used Adam [8]. The stepsize parameter in Adam was chosen from $\{0.01, 0.001, 0.0001\}$, though the other parameters are used with their default settings. The common initial value of the parameters $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ was chosen from $\{-10.0, -1.0, -0.5\}$. The sample size L was one, because larger sample sizes gave comparable or worse results.

Based on the discussion in [2], we tuned the hyperparameters α and β of the symmetric Dirichlet priors by a grid search. Each of α and β was chosen from {0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5}. The same grid search was used for the Dirichlet hyperparameters of the compared methods. The number of small batches in Algorithm 1 was set to 20. The number of iterations, i.e., how many times we scanned the entire document set, was chosen as 500. We computed ten test set perplexities based on the ten different random splits prepared for evaluation. The test set perplexity in Eq. (24) was computed at the 500th iteration.

4.4 Evaluation Results

Figs. 1 and 2 present the evaluation results. We split the results into two figures, because the two data sets NYT and MOVIE are widely different from the other two NSF and MED in their average document lengths. This difference may be part of the reason why we obtained different evaluation results. The horizontal axis of the charts in Figs. 1 and 2 gives the three settings for the number of topics: $K = 50$, 100, and 200. The vertical axis gives the test set perplexity averaged over the ten different random splits. The standard deviation of the ten test set perplexities is presented by the error bar. LNV and DEG are the labels for our proposal and its degenerated version, respectively. VB and CGS refer to the standard VB [4] and the collapsed Gibbs sampling [6], respectively.

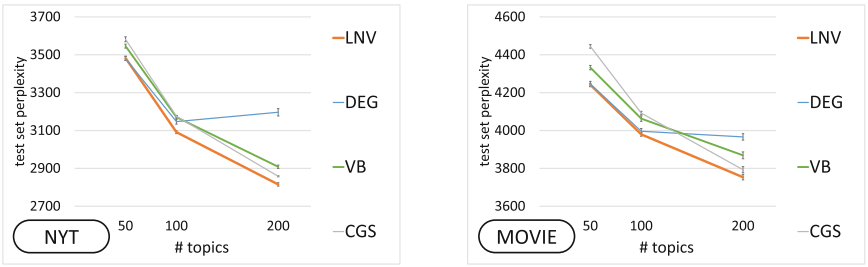

Fig. 1. Evaluation results in terms of test set perplexity for the two document sets NYT (left) and MOVIE (right), whose average document lengths are relatively long.

Figure 1 presents the results for the two data sets NYT and MOVIE, which consist of relatively long documents. Our proposal LNV led to the best perplexity for four cases among the six cases given in Fig. 1. When we set $K = 50$, DEG could give almost the same test set perplexity with LNV for both of the NYT and MOVIE data sets. However, the differences were not statistically significant, because the p-values of the two-tailed t-test were 0.463 and 0.211, respectively. For the other four cases, LNV could give the best perplexity. The differences for all these four cases were statistically significant. For example, when we set $K = 100$ for the MOVIE data set, the p-value of the two-tailed t-test where we compare LNV and DEG was 0.00134. It can be said that our proposal worked effectively for these two document sets.

Figure 2 shows the results for the two data sets NSF and MED, which consist of relatively short documents. Our proposal LNV could provide the best perplexity for the following three cases: $K = 50$ for the NSF data set, $K = 50$ for the MED data set, and $K = 100$ for the MED data set. Note that the difference between LNV and CGS when we set $K = 50$ for the NSF data set was statistically significant, because the p-value of the two-tailed t-test was 0.000121. For the other three cases, CGS gave the best perplexity. For these two data sets, our proposal worked only when the number of topics was not large.

Note that LNV was superior to VB for all settings presented in Figs. 1 and 2. This proves empirically that the Dirichlet distribution is not necessarily the best choice for approximating the true posterior even when we use the same mean field approximation with the standard VB. In sum, we can draw the following conclusion. When LNV is available, there may be no reason to use VB, DEG, or CGS for relatively long documents. Also for relatively short documents, our proposal may be adopted when the number of latent topics is small.

However, in terms of computation time, LNV has a disadvantage. For example, when we set $K = 200$ for the NYT data set, it took 43 h for finishing 500 iterations with LNV, though it took 14 h with CGS and 23 h with VB. However, inferences for LDA are generally easy to parallelize, e.g. by using GPU [18,19]. It may be an advantage for parallelization that each sample $\epsilon \sim \mathcal{N}(0, 1)$ in the proposed method can be drawn independently.

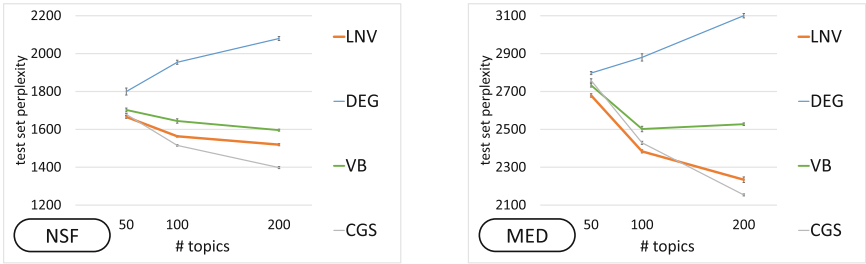

Fig. 2. Evaluation results in terms of test set perplexity for the two document sets NSF (left) and MED (right), whose average document lengths are relatively short.

4.5 Effect of Sampling

The degenerated version of our proposal gave a comparable perplexity only for the limited number of cases in our experiment. When the number of latent topics was large, or when the average document length was short, the degenerated version led to quite a poor perplexity. Especially in the two charts of Fig. 2, it seems that the degenerated version exhibits substantial overfitting. Therefore, sampling is indispensable. Recall that L, i.e., the number of the samples from the standard normal, was set to 1, because larger numbers of samples did not improve the test set perplexity. However, a single random sample could work as a kind of *perturbation* for the update of the corresponding parameter. Without this perturbation, the inference tended to get trapped in local minima as shown in Fig. 2. A single sample can change the course of inference through perturbation. This may be the reason why our proposal gave better perplexities than its degenerated version in many of the situations consulted in the experiment.

Based on our experience, it is important to optimize the standard deviation parameters $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ carefully in order to avoid overfitting. When the stepsize parameter of Adam was not tuned, the standard deviation parameters stayed around at their initial values. This made the perplexity almost similar to that of the degenerated version. In addition, the initial values of $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ also needed to be tuned carefully. However, the tuning was not that difficult, because the optimization was almost always successful when the parameters $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ took values widely different from their initial values. Further, we only needed to test several setting for the combination of the stepsize parameter in Adam and the common initial value of $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$. In our experiment, the stepsize parameter was chosen from $\{0.01, 0.001, 0.0001\}$. The common initial value of the parameters $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ was chosen from $\{-10.0, -1.0, -0.5\}$. Therefore, at most nine settings were checked. However, the combination of 0.001 for the stepsize parameter and -0.5 for the initial value of $\tau_{\theta,dk}$ and $\tau_{\phi,kv}$ often worked. Only when this setting did not work, we considered other settings.

5 Conclusion

In this paper, we proposed a new VB-type inference method for LDA. Our method is based on the stochastic gradient variational Bayes [9,13] and approximates the true posterior with the logistic normal distribution. The proposed method was better than the standard VB for all situations consulted in the experiment and was better even than the collapsed Gibbs sampling for not all but many situations. Further, when deprived of sampling, the inference tended to get trapped in local minima. Therefore, sampling worked.

While we use the logistic normal distribution in the proposed inference, we can choose other distributions as long as they meet the two requirements given in Sect. 2. Further, we can propose a similar inference also for other Bayesian probabilistic models. One important merit of SGVB is that the expectation with respect to the approximate posterior for continuous variables is estimated by the Monte Carlo method. Even when the full joint distribution of the target

Bayesian model is complicated, SGVB may make the computation relating to such expectations efficient. Therefore, it is worthwhile future work to provide a new inference for the existing Bayesian models with the distribution that has not been considered due to the complication in handling the expectations.

References

1. Aitchison, J., Shen, S.-M.: Logistic-normal distributions: some properties and uses. Biometrika **67**(2), 261–272 (1980)
2. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: UAI, pp. 27–34 (2009)
3. Blei, D.M., Lafferty, J.D.: Correlated topic models. In: NIPS, pp. 147–154 (2005)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003)
5. Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton (2011)
6. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS **101**(Suppl 1), 5228–5235 (2004)
7. Kang, J.-H., Lerman, K., Getoor, L.: LA-LDA: a limited attention topic model for social recommendation. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 211–220. Springer, Heidelberg (2013)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
9. Kingma, D.P., Welling, M.: Stochastic gradient VB and the variational auto-encoder. In: ICLR (2014)
10. Lin, T.-Y., Tian, W.-T., Mei, Q.-Z., Cheng, H.: The dual-sparse topic model: mining focused topics and focused terms in short text. In: WWW, pp. 539–550 (2014)
11. Mnih, A., Gregor, K.: Neural variational inference and learning in belief networks. In: ICML, pp. 1791–1799 (2014)
12. O'Connor, B., Stewart, B.M., Smith, N.A.: Learning to extract international relations from political context. In: ACL, pp. 1094–1104 (2013)
13. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML, pp. 1278–1286 (2014)
14. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004)
15. Sasaki, K., Yoshikawa, T., Furuhashi, T.: Online topic model for Twitter considering dynamics of user interests and topic trends. In: EMNLP, pp. 1977–1985 (2014)
16. Teh, Y.-W., Newman, D., Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: NIPS, pp. 1353–1360 (2007)
17. Vosecky, J., Leung, KW.-T., Ng, W.: Collaborative personalized Twitter search with topic-language models. In: SIGIR, pp. 53–62 (2014)
18. Yan, F., Xu, N.-Y., Qi, Y.: Parallel inference for latent Dirichlet allocation on graphics processing units. In: NIPS, pp. 2134–2142 (2009)
19. Zhao, H.-S., Jiang, B.-Y., Canny, J.F., Jaros, B.: SAME but different: fast and high quality Gibbs parameter estimation. In: KDD, pp. 1495–1502 (2015)