Pattern Analysis and Applications (2024) 27:20
https://doi.org/10.1007/s10044-024-01213-y

THEORETICAL ADVANCES

# Big topic modeling based on a two-level hierarchical latent Beta-Liouville allocation for large-scale data and parameter streaming

Koffi Eddy Ihou[1] · Nizar Bouguila[1]

Received: 13 June 2022 / Accepted: 4 December 2023 / Published online: 28 February 2024

## Abstract

As an extension to the standard symmetric latent Dirichlet allocation topic model, we implement asymmetric Beta-Liouville as a conjugate prior to the multinomial and therefore propose the maximum a posteriori for latent Beta-Liouville allocation as an alternative to maximum likelihood estimator for models such as probabilistic latent semantic indexing, unigrams, and mixture of unigrams. Since most Bayesian posteriors, for complex models, are intractable in general, we propose a point estimate (the mode) that offers a much tractable solution. The maximum a posteriori hypotheses using point estimates are much easier than full Bayesian analysis that integrates over the entire parameter space. We show that the proposed maximum a posteriori reduces the three-level hierarchical latent Beta-Liouville allocation to two-level topic mixture as we marginalize out the latent variables. In each document, the maximum a posteriori provides a soft assignment and constructs dense expectation–maximization probabilities over each word (responsibilities) for accurate estimates. For simplicity, we present a stochastic at word-level online expectation–maximization algorithm as an optimization method for maximum a posteriori latent Beta-Liouville allocation estimation whose unnormalized reparameterization is equivalent to a stochastic collapsed variational Bayes. This implicit connection between the collapsed space and expectation–maximization-based maximum a posteriori latent Beta-Liouville allocation shows its flexibility and helps in providing alternative to model selection. We characterize efficiency in the proposed approach for its ability to simultaneously stream both large-scale data and parameters seamlessly. The performance of the model using predictive perplexities as evaluation method shows the robustness of the proposed technique with text document datasets.

**Keywords** Topic model · Maximum a posteriori · Beta-Liouville · Conjugate prior · Expectation–maximization · Stochastic optimization · Parameter streaming · Text modeling

## 1 Introduction

In topic modeling literature, the classical maximum likelihood (ML) estimator has been applied to several classic topic models including PLSA (probabilistic latent semantic analysis), unigrams, and mixture of unigrams. Because of its frequentist nature, it is very limited in predictive modeling as it does not consider prior information. Reduced to multinomial distributions with no prior information, these classic topic models fundamentally carry the limitations of multinomials: using only frequencies as ways to represent probabilities often leads to poor estimates. In a highly sparse dataset, without any smoothing, frequencies are more likely to assign zero probability for unseen or rare events. Moreover, and very often, multinomials do not capture efficiently the words burstiness because of the lack of priors [1–3]. The integration of prior information has become fundamental for the flexibility of topic models such as latent Dirichlet allocation (LDA) over classical frequentist approaches. In other words, the limitations of classic frequentist models led to the emergence of the LDA and its variants. LDA is a latent generative probabilistic graphical model that assumes that words are generated from a mixture of topics [4]. The topics are themselves distributions over the vocabulary words. The topic proportions vary from one document to the other and exhibit how documents are organized, summarized

✉ Nizar Bouguila
nizar.bouguila@concordia.ca

Koffi Eddy Ihou
k_ihou@encs.concordia.ca

[1] Concordia Institute for Information Systems Engineering, Concordia University, 1455 de Maisonneuve Blvd, Montreal, QC H3G 1M8, Canada

according to the global topics. LDA allows documents to exhibit multiple topics. The success of the LDA model has reinforced its use in a wide variety of applications mainly in text document analysis [5–7] and computer vision [8, 9]. Compared to LDA, in a unigram model, words in a document are drawn from a single multinomial distribution (the word simplex). The mixture of unigrams is an augmented version of the unigram model with a discrete topic (latent) variable. With the mixture of unigrams, a document is now generated from only a single topic [10]. The PLSA is almost similar to LDA topic model but with no prior information [11]. It is a relaxation of the mixture of unigrams assumption as it allows a document to exhibit multiple topics. As presented earlier, the lack of priors in PLSA makes the model unfit for prediction and often suffers from overfitting problems. The LDA topic model provides a solution to the probabilistic latent semantic indexing (PLSI) or PLSA, unigram, and mixture of unigrams by including prior information as it treats topic proportions as random variables.

Since LDA and its current variants rely extensively on prior information, it is natural to perform parameter estimation where the logarithm of the priors offers a possibility to act as a regularizer of ML estimates. This ultimately introduces the flexibility of the maximum a posteriori (MAP) framework. The MAP estimates are point estimates, whereas full Bayesian analysis often characterizes the posterior mean instead of a single estimate (mode) [12–14]. However, point estimates are often preferred because posterior means require computationally expensive methods and often lead to intractable solutions. The MAP framework models directly the posterior distribution of the parameters. Due to its prior information, the MAP is robust to outliers. In topic modeling, these advantages could present a possible MAP technique with standard expectation-maximization (EM) algorithm in online fashion as alternatives to complex methods such as variational Bayes (VB) [10, 15], MCMC (Markov chain Monte Carlo) using CGS (collapsed Gibbs sampling) [16], and EP (expectation propagation) [17]. Although the MAP is not invariant to reparameterization as it requires the Jacobian information to relocate the mode, we can use unnormalized reparameterization to simply seek for equivalent models that could help in characterizing efficiently its online framework [11, 18, 19]. We can also observe from the literature that online LDA topic models such as stochastic collapsed variational Bayes (SCVB) [18], online CGS (OGS) [20, 21], SVB (stochastic variational Bayes) [22, 23], online VB (OVB) [23], generally implement stochastic optimizations [24] from batch LDA models (VB [10, 15], collapsed variational Bayes (CVB) [25] and CGS [12, 26, 27]). Furthermore, as these models only focus on large-scale data processing while ignoring parameter streaming, the work in [19] recently implemented an online LDA topic called fast online EM that accommodates parameter streaming to large-scale data modeling. Modeling dependency between hidden variables has been ignored in standard variational Bayes that assumes that joint variational distributions variables are all independent from each other. In addition, relaxing the mean-field assumption can become extremely challenging in the latent update equations when using empirical Bayes framework (the log marginal distribution) to set the lower bound as shown in [25] with its symmetric LDA.

The first critics to all these methods always point to the use of Dirichlet (Dir) prior in LDA for inference. LDA assumes that its topic components are all independent when using the Dir prior. Furthermore, one of the limitations within the multinomial assumption is that often the poor estimates are results of the fact that events are supposed independent, which is not always the case [28–30]. By choosing flexible priors instead, we could characterize efficiently dependency between events which are translated into dependency between documents and topic components (topic proportions). LDA is not the right model when it comes to characterizing dependency since it systematically prohibits such interpretation because documents simply cannot be dependent under the LDA topic model. As suggested in [4] using Dir leads to an unrealistic way to explore unstructured collections of documents because in real life scenario, there is always a high probability of existence of a topic correlation setting in a large collection. This drawback in LDA promoted the use of logistic normal distribution as an alternative to the Dirichlet prior in topic correlation [31–38]. Another major problem and setback in finite mixture topic modeling is the lack of effective model selection criteria [4, 10, 19, 39–41] especially with LDA which relies on cross-validation solutions. For large-scale applications cross-validation methods are not efficient. Since LDA is too restrictive due to the Dir distribution while non-conjugate priors such as logistic normal distributions often led to very complex deterministic (VB, CVB, and EP) and MCMC using CGS inferences, we propose a very simple algorithm that performs a MAP estimate on the latent BL allocation (LBLA) where the conjugate prior to the multinomial is the asymmetric Beta-Liouville (BL) prior. The flexibility of the prior allows us to model dependency between documents. Indeed, the BL offers a more general covariance structure than the Dirichlet as well as more degrees of freedom as deeply discussed in [42–47]. In attempt to induce dependence, the CVB marginalizes out the parameters, while leaving the latent variables; on the other hand, the proposed maximum a posteriori latent Beta-Liouville allocation (MAP-LBLA) integrates out the latent variables instead and even reduces the three-level hierarchical LBLA topic model to just two levels. This ultimately simplifies computation. We proposed a stochastic at word-level online EM algorithm for MAP-LBLA as an alternative to online

LDA in [19] to which we provide a refined model selection including data and parameter streaming for fast inference. Our model outperforms the LDA-based topic models and shows the robustness of the scheme in producing very accurate predictive distributions and perplexities. In our method, because implementing a word-level processing, documents parameters and topics are global parameters. This is in contrast to the standard stochastic VB (SVB) approach that supports a document-level processing where the only global parameter is the topic. We show that our stochastic algorithm using online EM has connections within the collapsed variational Bayesian inferences through unnormalized reparameterization of the MAP [11, 18]. Under this reparameterization it is clear that our technique could follow a minibatch of size one as we will show later in the coming sections. This allows the model to manage the vocabulary size easily. As we implement a stochastic method that favors small samples at a time in a document, the MAP can effectively regularize maximum likelihood estimator (MLE) estimates and performs better than frequentist estimators. To each word accessed, the E-step provides a sample (EM responsibility vector) from the posterior distribution; but no longer stores it within our stochastic method as in the batch EM. All these flexibilities make our approach more robust and accurate over extremely fast methods that could escape many critical steps (during processing) that are required for a good modeling. We finally demonstrated that our model while using unnormalized update equations is flexible due to the asymmetric BL prior that generalizes the Dir distribution. The main contributions of our proposed parametric topic model are:

- We provide alternative to the MAP-LDA and its variants including stochastic and online versions. We selected the BL prior to estimate very heterogeneous topics that enhance predictive models and perplexities.
- The simplicity of inference with the standard EM algorithm over complex methods such as variationals and EP including MCMC methods such as CGS and CVB allows to model dependence in exact manner which leads to much accurate parameters estimates.
- We successfully provide a solution (alternative) to model selection problem within finite mixture topic model setting, which is a very challenging concept due to the lack of criteria for model selection in topic modeling in general as our model stochastically favors small samples which are regularized by the prior information within the proposed MAP framework: our approach uses its equivalent models to efficiently propose model selection

The paper is organized as follows: Section 2 presents the related work and background, while Sect. 3 introduces the proposed online EM-based MAP-LBLA approach. Section 4 illustrates the experimental results, and finally, Sect. 5 provides future work and conclusion.

## 2 Related work and background

LDA is a generative probabilistic graphical model that summarizes documents (texts, images) as mixtures over topics. Topics are distributions over vocabulary words [15]. Under its generative process, LDA assumes that a word is generated from a mixture of topics [4]. Many inferences support the LDA architecture and make it the most recognized topic model in the literature. The main inferences include VB and CVB which describe the variational approaches while GS (Gibbs sampling) and CGS (collapsed GS) which are MCMC methods [25]. The CVB and CGS are based on collapsed representation where the parameters are marginalized out: CVB is variational method in the collapsed space; therefore, a deterministic approach where CGS is an MCMC method or stochastic in the collapsed space. The CGS provides a hard assignment technique, while CVB favors a soft clustering method resulting in a K-dimensional variational distribution being associated with each word or token [12]. One of the advantages of the collapsed representation was to characterize a dependence structure in topic modeling as a way of relaxing the independency assumption in mean-field variational methods. It also provides a better lower bound to the log marginal likelihood for accurate predictive distributions showing parameters estimated in exact way [25]. The VB and GS are inferences in uncollapsed spaces. The work in [48] has constructed a partially collapsed space where documents proportions are marginalized out leaving the latent variables and the topics. The majority of these batch inferences have been extended for online processing leading to OVB [23], SCVB [18], OGS [20, 21], SVB [22, 23], etc. For a direct modeling of the parameters, the MAP marginalizes out the latent variables and optimizes an EM lower bound on the posterior distribution of the parameters in M-step. The E-step follows a stochastic expectation that computes unnormalized expected sufficient statistics (for exponential family distributions) also called EM statistics [18]. In latent topic models, we can observe that the MAP integrates out the latent variables while the CVB inference marginalizes out the parameters. Authors in [11] have tried to show the connection between these inferences for LDA through hyperparameter analysis. MAP reduces the three-level topic model to two levels and introduces a mixture model setting. Other main characteristics and challenges of LDA model include the problem of a robust model selection [10, 19, 39–41], and correlation between topics [4, 37, 38, 49]. The problem with these inferences is that the majority are LDA-based approaches. Furthermore, LDA could not characterize dependence structure because it is one of its intrinsic limitations. Under the Dirichlet its random variable components are independent, so correlation between topics could not be emphasized with efficiency within the LDA. The model selection framework in finite topic modeling is very challenging. For instance, the multitopic technique [40] is efficient for batch learning but not for online one. Its limitation is due to the relevance feedback from a user. It means it cannot perform without human intervention. The VI (variation of information) method [39] operates within the uniform probability measure setting which we believe could not be ideal for the MAP technique because estimates with uniform priors are equivalent to ML estimates. The fast online LDA topic model in [19] provided a model selection that uses accumulated residuals (to select the number of topics and vocabulary size) combined with a buffering system that facilitates easy transfer of data between the PC (personal computer) memory and its external storage. Its sorting mechanism based on residuals for model selection is complicated because both the time and memory (space) complexities rely on the number of topics $K$ and vocabulary size $V$. Despite the fact that the updating and normalization steps of the responsibility vector benefit from time complexity, it is really difficult to understand how the framework became invariant to the number of topics at some points when analyzing the time complexity.

Due to these difficulties, we propose an alternative with more improvement: we implement an online EM method for MAP estimation with LBLA topic model, a generalization of the LDA. The proposed approach uses a BL prior [50–52] as an alternative to the Dir distribution. The BL has ability for topic correlation [52] framework similar to work in [4, 37, 49]. We emphasized on a word-level stochastic online EM approach whose unnormalized parameterization connects with stochastic inferences in the collapsed space. Our proposed method uses its internal structure to reduce the number of topics and vocabulary size and allows for flawless data and parameter streaming. Importantly, compared to other methods that use computationally expensive resources for model selection, our proposed model selection technique does not require too much computation. Its advantage is that it promotes small samples processing (reasonable minibatch sizes), which encourages the use of small number of topics and vocabulary sizes. This reason explains our stochastic method which can implement a minibatch setting of size one for small samples. It constantly processes and updates, for the global topic matrix of size $K \times V$, only its $v$th column using a reduced number of topics. Small samples are appropriate for our MAP-LBLA method because of the presence of the prior to regularize or correct estimates.

## 3 Proposed approach

In this section, we propose a standard EM algorithm for MAP estimation of LBLA topic model. We show that it is an alternative to the CVB algorithm [25]. Moreover, we show that the complexity of the VB approach (when characterizing dependency between latent variables and parameters) ultimately led to the implementation of our simple and standard EM algorithm for MAP estimation: modeling dependence in latent topic models is a way of characterizing accurate parameter estimation from the work in [25]. However, the variational method in the collapsed space can be extremely complex despite its flexibility. We demonstrate that in spite of the simplicity of the proposed EM algorithm for MAP-LBLA, it is implicitly connected to the CVB inference. Furthermore, we cover the generative equation of the MAP-LBLA that allows us to formulate through a coordinate ascent framework the EM-based batch and online algorithms for MAP-LBLA. The accuracy in the expectations also depends on the proposed unnormalized representation.

### 3.1 Modeling dependency between hidden variables

One of the central themes in CVB inference is the possibility to reach accurate parameters estimates by relaxing the independence assumption in mean-field variational methods. This relaxation introduces dependency between latent variables and models parameters. To be more specific, from Table 1, with the LBLA hyperparameters $\varepsilon, \zeta$, and hidden variables $Z, \theta$, and $\varphi$, let's consider the case where we lower bound the log marginal likelihood $\log p(X|\varepsilon, \zeta)$ using variational distributions $q$:

$$\begin{aligned}\log p(X|\varepsilon,\zeta) &= \log \int_{\theta}\int_{\varphi}\sum_{Z} p(X,Z,\theta,\varphi|\varepsilon,\zeta)\mathrm{d}\theta\mathrm{d}\varphi \\ &= \log \mathbb{E}_{q(\theta,\varphi,Z)}\left(\frac{p(X,Z,\theta,\varphi|\varepsilon,\zeta)}{q(\theta,\varphi,Z)}\right) \\ &\geq \mathbb{E}_{q(\theta,\varphi,Z)}\log\left(\frac{p(X,Z,\theta,\varphi|\varepsilon,\zeta)}{q(\theta,\varphi,Z)}\right)\end{aligned}$$

This is also equivalent to:

$$\begin{aligned}\log p(X|\varepsilon,\zeta) \geq\ & \mathbb{E}_{q(\theta,\varphi,Z)}\log(p(X,Z,\theta,\varphi|\varepsilon,\zeta)) \\ & - \mathbb{E}_{q(\theta,\varphi,Z)}\log(q(\theta,\varphi,Z))\end{aligned}$$

such that:

$$\begin{aligned}\log p(X|\varepsilon,\zeta) =\ & \mathscr{F}(q,\theta,\varphi,Z) \\ & + KL(q(\theta,\varphi,Z)||p(\theta,\varphi,Z|X,\varepsilon,\zeta))\end{aligned} \tag{1}$$

where

**Table 1** Variables and definitions

| Model variables and acronyms | Definitions |
|---|---|
| $\mathscr{D}$ | Total number of documents |
| $\mathscr{W}$ | Total number of words in the corpus |
| $\mathscr{V}$ | Minibatch size |
| $\mathscr{W}_j$ | Total number of words in a document $j$ |
| $K$ | Total number of topics |
| $V$ | Vocabulary size |
| $(i, j)$ | The $i$th word or topic assignment in a document $j$ |
| $k$ | The $k$th topic |
| $X = \{x_{ij}\}$ | Observed words |
| $Z = \{z_{ij}\}$ | Latent variables |
| $\theta_j = \{\theta_{jk}\}$ | Topic proportions |
| $\varphi_k = \{\varphi_{kv}\}$ | Corpus parameters or global topics |
| $\mathrm{BL}(\varepsilon)$ | Beta-Liouville distribution with parameter $\varepsilon$ |
| $\theta_{jk}/\varepsilon \sim \mathrm{BL}(\varepsilon)$ | $\theta_{jk}/\varepsilon$ drawn from $\mathrm{BL}(\varepsilon)$ |
| $\varphi_{kv}/\zeta \sim \mathrm{BL}(\zeta)$ | $\varphi_{kv}/\zeta$ drawn from $\mathrm{BL}(\zeta)$ |
| $\mathrm{Mult}(\theta_{jk})$ | Multinomial distribution with parameter $(\theta_{jk})$ |
| $z_{ik}/\theta_{jk} \sim \mathrm{Mult}(\theta_{jk})$ | $z_{ik}/\theta_{jk}$ drawn from $\mathrm{Multinomial}(\theta_{jk})$ |
| $x_i/z_{ik}, \varphi_{kv} \sim \mathrm{Mult}(\varphi_{z_{ik}})$ | $x_i/z_{ik}, \varphi_{kv}$ drawn from $\mathrm{Multinomial}(\varphi_{z_{ik}})$ |
| $\psi_{ijk}$ | Responsibility of component $k$ for word $x_{ij}$ in document $j$ |
| $\mathscr{F}(\psi_{ijk}, \theta, \varphi)$ | EM lower bound to the log likelihood |
| $\mathscr{L}(\psi_{ijk}, \theta, \varphi)$ | EM lower bound for MAP (maximum a posteriori) |
| $\mathscr{N}_{-ij}$ | Expected Count excluding $z_{ij}$ |

$$\log p(X|\varepsilon, \zeta) \geq \mathscr{F}(q, \theta, \varphi, Z) \tag{2}$$

In the joint space, the variational distribution $q(\theta, \varphi, Z)$ using the mean-field variational is:

$$q(\theta, \varphi, Z) = q(\theta)q(\varphi)q(Z) \tag{3}$$

The variational distribution in (3) characterizes the independence assumption in standard VB inference. In the collapsed space, the variational distribution in (4) follows dependency between latent variables and parameters:

$$q(\theta, \varphi, Z) = q(\theta, \varphi|Z)q(Z) \tag{4}$$

Using the lower bound, we reach the maximum at $q(\theta, \varphi|Z) = p(\theta, \varphi|X, Z)$ where the functional $\mathscr{F}$ (lower bound) now becomes $\mathscr{F}(q, Z)$. From the work in [25, 53], we obtain:

$\mathscr{F}(q, Z) = \mathbb{E}_{q(Z)} \log(p(X, Z|\varepsilon, \zeta)) - \mathbb{E}_{q(Z)} \log(q(Z))$ and $\log q(Z_j) = \mathbb{E}_{i \neq j} q(Z) \log(p(X, Z|\varepsilon, \zeta)) + C$ with $C$ being a constant. It leads to:

$$q(Z_j) = \frac{\exp \mathbb{E}_{i \neq j} q(Z) \log \left(p(X, Z|\varepsilon, \zeta)\right)}{\sum_z \exp \mathbb{E}_{i \neq j} q(Z) \log \left(p(X, Z|\varepsilon, \zeta)\right)} \tag{5}$$

which is also equivalent to:

$$q(Z_j = k) = \frac{\exp\{\mathbb{E}_{q(Z_{-j})}\left[\log p(X, Z_{-j}, Z_j = k|\varepsilon, \zeta)\right]\}}{\sum_{i=1}^{K} \exp \mathbb{E}_{q(Z_{-j})} \log \left(p(X, Z_{-j}, Z_j = i|\varepsilon, \zeta)\right)} \tag{6}$$

where $q(Z_j)$ is the update equation for CVB algorithm as illustrated in [25, 41, 52]. This update equation is really complex and requires the Gaussian approximation along with second order Taylor expansion. The CVB when modeling dependence structure makes the joint variational distributions for the parameters conditioned on the latent variables in (4).

### 3.1.1 The space of parameters and MAP-LBLA

We marginalize out the latent variables, leaving the model (LBLA) parameters. This is a reverse setting of the collapsed representation that integrates out the parameters. A way of modeling dependency (between hidden variables) in MAP estimation is to make the multinomial variational distribution conditioned on the parameters as shown in (7). In this section, we show that using variational methods makes the MAP update equation extremely complex as well; which ultimately leads to a much simpler method using standard EM algorithm. We have the following variational joint distribution:

$$q(\theta, \varphi, Z) = q(\theta, \varphi)q(Z|\theta, \varphi) \tag{7}$$

Here, we get the maximum when $q(Z|\theta, \varphi) = p(Z|\theta, \varphi, X)$ leading to a lower bound functional:

$$\begin{aligned}
\mathscr{F}(q, \theta, \varphi) &= \mathbb{E}_{q(\theta,\varphi)} \log p(X, \varphi, \theta|\varepsilon, \zeta) \\
&\quad - \mathbb{E}_{q(\theta,\varphi)} \log q(\theta, \varphi) \\
\log q(Z|\varphi, \theta) &= \mathbb{E}_{q(\varphi,\theta)} \log p(X, Z, \theta, \varphi|\varepsilon, \zeta) \\
&\quad + C \propto \mathbb{E}_{q(\varphi,\theta)} \left[\log(p(X, Z|\theta, \varphi)p(\theta, \varphi|\varepsilon, \zeta))\right] \\
&\propto \mathbb{E}_{q(\varphi,\theta)} \left[\log(p(X, Z|\theta, \varphi)p(\theta|\varepsilon)p(\varphi|\zeta))\right]
\end{aligned}$$

We obtain the following update equations for MAP-LBLA:

$$\begin{aligned}
\log q(Z|\varphi, \theta) \propto\ & \mathbb{E}_{q(\varphi,\theta)} \left[\log p(X|Z, \varphi) + \log p(Z|\theta)\right] \\
& + \mathbb{E}_{q(\varphi,\theta)} \left[\log p(\theta|\varepsilon) + \log p(\varphi|\zeta)\right]
\end{aligned} \tag{8}$$

$$\begin{aligned}
\log q(\varphi, \theta) &= \mathbb{E}_{q(Z)} \left[\log p(X, Z, \theta, \varphi|\varepsilon, \zeta)\right] + C \\
\log q(\varphi, \theta) &\propto \mathbb{E}_{q(Z)} \left[\log p(X|Z, \varphi) + \log p(Z|\theta)\right] \\
&\quad + \mathbb{E}_{q(Z)} \left[\log p(\theta|\varepsilon) + \log p(\varphi|\zeta)\right]
\end{aligned} \tag{9}$$

With the Jensen's inequality, providing a lower bound to the log marginal likelihood function $p(X|\varepsilon, \zeta)$ in [25] makes the variational update equation in (9) for MAP intractable because of the coupling between the corpus and document parameters. Even the posterior variational distribution for latent $Z$ in (8) is intractable due to the same coupling. Using the Jensen's inequality, we therefore propose a lower bound to the log likelihood function instead. Then, we derive the MAP lower bound from the log likelihood's lower bound by adding the log of the priors distributions to the log likelihood's lower bound.

### 3.2 Unnormalized parameterization

The stochastic variational inference randomly draws a data point (a word or a document) and then learns its local parameters to update the global parameters following a natural gradient update approach [22, 23]. Let's suppose, for instance, that we are sampling one document at a time. Following the stochastic variational method at document level, we compute the noisy estimate of the natural gradient of the objective function corresponding to $\mathscr{D}$ copies of document $j$ which are then used to update the global parameters. To allow $\mathscr{D}$ copies of the objective function (ELBO), we take the corpus-wide terms [23, 54] in the variational lower bound of a single document $j$ and normalize them by $\mathscr{D}$ (the total number of documents in the corpus) so that lower bound becomes:

$$\mathscr{L} = \sum_j \mathscr{L}_j = \mathbb{E}_j[\mathscr{D}\mathscr{L}_j] \tag{10}$$

where $\mathscr{D}\mathscr{L}_j$ is the variational lower bound (ELBO) with $\mathscr{D}$ copies of document $j$. Similarly, in MAP estimate as we follow this time a stochastic framework at word level, we need to operate on unnormalized parameterization in order to compute unnormalized expected sufficient statistics during the stochastic expectation step as in MAP-LDA. This is because in online EM algorithm as proposed in [55], the likelihood function and the sufficient statistics are normalized by the total number of words $\mathscr{W}$ in the corpus. Using $\mathscr{W}$ copies of the proposed EM lower bound for each word leads to an unnormalized expected sufficient statistics during E-step and provides the appropriate scale between the normalized ML estimates and the prior distribution that summarizes the posterior probability of the parameters. This shows the correspondence between the proposed approach for MAP where we estimate sufficient statistics within unnormalized representation and the stochastic variational inferences as they use noisy estimates of natural gradient of the ELBO to update the global parameters. We compute the unnormalized expected sufficient statistics as MAP estimates for the parameters using online averages as alternatives. While performing in unnormalized parameterization of LDA, one of the advantages is the fact that MAP-LDA's update equation and the one for CVB0 (zero-order approximation of LDA) are analytically identical if we adjust their hyperparameters by one [11]. This ultimately connects the CVB0-LDA to MAP-LDA and stochastic CVB0-LDA (SCVB0-LDA) to online EM for MAP-LDA, and it introduces the EM statistics and responsibilities to CVB0 statistics and variational distributions (responsibilities). This connection allows the SCVB0-LDA as unnormalized stochastic MAP-LDA with minibatch of size one scheme when assessing one data point at a time (from its recursive update equation). In this paper, we are also performing in unnormalized parameterization of LBLA where we hope to show its connection to the collapsed space representation. The connection originates from the fact that both MAP and SCVB0 operate on unnormalized parameterization of the LDA. Furthermore, the SCVB0-LDA's update equation is also similar to that of MAP-LDA [18]. We implement a MAP-LBLA estimation with stochasticity at word-level that is connected to SCVB0-LDA inference. In MAP, from the hidden variables, we marginalize out the latent variables from the corpus while leaving only the parameters. On the other hand, CVB inference integrate out the parameters from the hidden variables.

### 3.3 Generative process of the MAP-LBLA

LDA [10, 15] is generally a three-level hierarchical model. The corpus level includes the global topics and their hyperparameters and document hyperparameters. The document level is characterized by the topic proportions, and finally, the word level includes the topic assignments and the words [10]. By marginalizing out the parameters, we get a two-level hierarchical LDA (corpus to document and document to word). As based on the LDA architecture, LBLA in this condition also follows a two-level topic model, and as a result, generates documents within the MAP framework as:

Choose a global topic $\varphi_k|\zeta \sim \mathrm{BL}(\zeta)$ where $k \in \{1, ..., K\}$
For each document $j$
the topic proportion $\theta_j|\varepsilon \sim \mathrm{BL}\ (\varepsilon)$
For $i \in \{1, 2, ..., \mathscr{W}\}$ in document $j$
Choose word $x_i|\theta_j, \varphi_{1:K} \sim \mathrm{Mult}\left(\sum_{i=1}^{K} \theta_{ji}\varphi_i\right)$

The variables $x_i$, $w_i$, and $v_i$ could be used interchangeably to denote a word in the vocabulary. Table 1 summarizes the relevant variables for the MAP-LBLA topic model.

### 3.4 The two-level LBLA topic mixture model

In general from the hidden variables and observed data, the three-level generative equation is:

$$p(X, Z, \theta, \varphi|\varepsilon, \zeta) = \prod_{k=1}^{K} p(\varphi_k|\zeta) \prod_{j=1}^{\mathscr{D}} p(\theta_j|\varepsilon) \times \prod_{i=1}^{N} p(z_{ij}|\theta_j)p(x_{ij}|\varphi, z_{ij}) \tag{11}$$

We then compute the joint posterior distribution:

$$p(Z, \theta, \varphi|X, \varepsilon, \zeta) = \frac{p(X, Z, \theta, \varphi|\varepsilon, \zeta)}{p(X|\varepsilon, \zeta)}$$

When we marginalize out the latent variables, the two-level LBLA posterior distribution becomes:

$$p(\theta, \varphi|X, \varepsilon, \zeta) = \frac{p(X, \theta, \varphi|\varepsilon, \zeta)}{p(X|\varepsilon, \zeta)} \propto p(X, \theta, \varphi|\varepsilon, \zeta) \tag{12}$$

where

$$p(X, \theta, \varphi|\varepsilon, \zeta) = \sum_{Z} p(X, Z|\theta, \varphi)p(\theta, \varphi|\varepsilon, \zeta)$$

$$\sum p(X, Z|\theta, \varphi)p(\theta, \varphi|\epsilon, \zeta) = p(\theta|\varepsilon)p(\varphi|\zeta) \sum_{Z} p(X, Z|\theta, \varphi) \tag{13}$$

$p(\theta_j|\varepsilon)$ and $p(\varphi_k|\zeta)$ are BL priors; $\varepsilon = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ and $\zeta = (\lambda_{k1}, ..., \lambda_{kV}, \lambda, \eta)$ are their respective parameters. The document BL prior $p(\theta_j|\varepsilon)$ is defined as:

$$\begin{aligned} p(\theta_j|\varepsilon) &= BL(\alpha_1, ..., \alpha_K, \alpha, \beta) \\ &= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &\times \prod_{k=1}^{K} \frac{\theta_{jk}^{\alpha_k - 1}}{\Gamma\left(\alpha_k\right)} \left(\sum_{k=1}^{K} \theta_{jk}\right)^{\alpha - \sum_{k=1}^{K} \alpha_k} \\ &\left(1 - \sum_{k=1}^{K} \theta_{jk}\right)^{\beta - 1} \end{aligned} \tag{14}$$

To show the sufficient statistics and natural parameters of the BL priors for the corpus and documents parameters, we represent them in exponential family form:

$$\begin{aligned} p(\theta_d|\varepsilon) = \exp \Bigg\{ &\left(\sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{jk}\right) \\ &+ \left(\alpha - \sum_{k=1}^{K} \alpha_k\right) \log \left(\sum_{k=1}^{K} \theta_{jk}\right) \\ &+ (\beta - 1) \log \left(1 - \sum_{k=1}^{K} \theta_{jk}\right) + \log \Gamma\left(\sum_{k=1}^{K} \alpha_k\right) \\ &+ \log(\alpha + \beta) - \log \Gamma(\alpha) - \log \Gamma(\beta) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) \Bigg\} \end{aligned} \tag{15}$$

From (14) and (15), we use similar steps for the corpus BL prior $p(\varphi_k|\zeta)$. We define the joint distribution $p(X, \theta, \varphi|\varepsilon, \zeta)$ such that: $p(X, \theta, \varphi|\varepsilon, \zeta) = p(X|\theta, \varphi)p(\theta|\varepsilon)p(\varphi|\zeta)$

$$p(X, \theta, \varphi|\varepsilon, \zeta) = \sum_{Z} p(X, Z, \theta, \varphi|\varepsilon, \zeta) \tag{16}$$

$$= \left(\sum_{Z} p(X, Z|\theta, \varphi)\right) p(\theta|\varepsilon)p(\varphi|\zeta) \tag{17}$$

$$= \left(\sum_{Z} p(X|Z, \varphi)p(Z|\theta)\right) p(\theta|\varepsilon)p(\varphi|\zeta) \tag{18}$$

From (16) to (18), we saw that when the latent variables are marginalized out, the three-level LBLA topic model is reduced to a two-level LBLA model similar to LDA. The distribution $p(X|\varphi, \theta) = \sum_{Z} p(X|Z, \varphi)p(Z|\theta)$ is reminiscent of a mixture topic model (PLSI or mixture of unigrams) [10]. Given priors information, we can define:

$$p(X|\varepsilon, \zeta) = \int \int p(\theta|\varepsilon)p(\varphi|\zeta) \left(\prod_{i=1}^{N} p(x_i|\theta, \varphi)\right) \mathrm{d}\theta \mathrm{d}\varphi \tag{19}$$

This marginal distribution (19) of a document is a (continuous) mixture distribution whose mixture weights are $(p(\theta|\varepsilon) \times p(\varphi|\zeta))$ and components $p(x_i|\theta, \varphi)$ [10].

### 3.5 The EM lower bound for the MAP-LBLA

We first define the log likelihood $\log P(X|\theta, \varphi)$ as:

$$\log P(X|\theta, \varphi) = \log \sum_Z p(X, Z|\theta, \varphi) \tag{20}$$

We introduce the distribution $q(Z)$ over the latent variables $Z$. Instead of log marginal distribution, we provide an EM lower bound to the log likelihood which allows us to include the prior in the EM lower bound for MAP-LBLA in (28).

$$\log p(X|\theta, \varphi) = \mathscr{F}(q, \theta, \varphi) + \mathrm{KL}(q||p) \geq \mathscr{F}(q, \theta, \varphi) \tag{21}$$

$$\mathscr{F}(q, \theta, \varphi) = \sum_Z q(Z) \log \frac{p(X, Z|\theta, \varphi)}{q(Z)} \tag{22}$$

$$\mathrm{KL}(q||p) = -\sum_Z q(Z) \log \frac{p(Z|X, \theta, \varphi)}{q(Z)} \tag{23}$$

If $q(Z) = p(Z|X, \theta^0, \varphi^0)$, then $KL(q||p) = 0$, then we have:

$$\mathscr{F}(q, \theta, \varphi) = \sum_Z q(Z) \log p(X, Z|\theta, \varphi) - \sum_Z q(Z) \log q(Z) \tag{24}$$

$$\begin{aligned}\mathscr{F}(q, \theta, \varphi) = &\sum_Z p(Z|X, \theta^0, \varphi^0) \log p(X, Z|\theta, \varphi) \\ &- \sum_Z p(Z|X, \theta^0, \varphi^0) \log p(Z|X, \theta^0, \varphi^0)\end{aligned} \tag{25}$$

$$\mathscr{F}(q, \theta, \varphi) = Q(\theta, \varphi, \theta^0, \varphi^0) + C \tag{26}$$

$$= \sum_Z p(Z|X, \theta^0, \varphi^0) \log p(X, Z|\theta, \varphi) \tag{27}$$

The functional $\mathscr{F}$ represents the standard EM lower bound for MLE as illustrated in Table 1. Now using Bayes' theorem, we can derive an EM lower bound for MAP-LBLA:

$$\begin{aligned}\log p(\theta, \varphi|X) &= \log p(\theta, \varphi, X) - \log p(X) \\ &= \log p(X|\theta, \varphi) + \log p(\theta, \varphi) - \log p(X) \\ &= \mathscr{F}(q, \theta, \varphi) + KL(q||p) + \log p(\theta, \varphi) + C \\ &\geq \mathscr{F}(q, \theta, \varphi) + \log p(\theta, \varphi) + C \\ &\geq \mathscr{F}(q, \theta, \varphi) + \log p(\theta) + \log p(\varphi) + C\end{aligned}$$

Here $KL(q||p) \geq 0$ and $\log p(X)$ is a constant $C$. Since $q = q(Z) = p(Z|X, \theta^0, \varphi^0) = \psi_{ijk}$ which is our EM responsibility vector, similar to a variational responsibility, then the EM lower bound for MAP is:

$$\mathscr{L}(\psi_{ijk}, \theta, \varphi|\varepsilon, \zeta) = \mathscr{F}(\psi_{ijk}, \theta, \varphi) + \log p(\theta|\varepsilon) + \log p(\varphi|\zeta) \tag{28}$$

This shows that at the E-step, the MAP lower bound will be identical or reduced to the MLE one if we compute the latent $\psi_{ijk}$ and then the M-step will require both the MLE lower bound and the priors information to estimate the parameters [53]. We have in our case a topic mixture model where its parameters are drawn from their respective conjugate priors. We showed that when $q(Z) = p(Z|X, \theta^0, \varphi^0) = \psi_{ijk}$ which is the complete conditional distribution of the latent variables given the samples and model parameters, the variational case and the standard mixture model technique coincide. Below, from (29) to (34), we show the MAP steps for its point estimate from its EM lower bound with LBLA.

$$\begin{aligned}\mathscr{L}(\psi_{ijk}, \theta, \varphi) \propto &\left(\sum_k \psi_{ijk} \sum_{i,j,v} \log p(X_i|Z_{ij}, \varphi_{kv}) p(Z_{ij}|\theta_{jk})\right) \\ &+ \left(\sum_{j,k} \log p(\theta_{jk}|\varepsilon) + \sum_{k,v} \log p(\varphi_{kv}|\zeta)\right)\end{aligned} \tag{29}$$

From the lower bound in (29) and (1), we derive the coordinate ascent method that is used to compute the model point estimate $\theta$ and $\varphi$ from (2) to (9). Then we formulate the MAP-LBLA update equation as a function of $\theta$ and $\varphi$ using (9), (7), (30), (31), (32), and (33).

$$\psi_{ijk} \propto (\theta_k)(\varphi_k)(\varphi_{k(V+1)}) \tag{30}$$

$$\psi_{ij(K+1)} \propto \left(\theta_{j(K+1)}\right) \tag{31}$$

$$\begin{aligned}\psi_{ijk} \propto &\left[\frac{\left(\mathscr{N}_\theta^{jk} + \alpha_k - 1\right)}{\left(\sum_k \alpha_k - 1\right) + \left(\sum_{k=1}^K \mathscr{N}_\theta^{jk}\right)}\right] \\ &\times \left[\frac{\left(\mathscr{N}_\varphi^{v_{ij}k} + \lambda_{kv} - 1\right)}{\left(\sum_v \lambda_{kv} - 1\right) + \left(\sum_{v=1}^V \mathscr{N}_\varphi^{v_{ij}k}\right)}\right] \\ &\times \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)(\varphi_{k(V+1)})\end{aligned} \tag{32}$$

such that:

$$\mathbb{U} = \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)(\varphi_{k(V+1)}) \tag{33}$$

with:

$$\begin{cases} 1 - \theta_{d(K+1)} = \sum_{k=1}^K \theta_{dk} < 1 \\ 1 - \varphi_{k(V+1)} = \sum_{v=1}^V \varphi_{kv} < 1 \\ \varphi_{k(V+1)} = 1 - \sum_{v=1}^V \varphi_{kv} < 1 \end{cases} \tag{34}$$

The Beta-distributed random variables in (33) make the MAP-LBLA (32) irreducible to MAP-LDA due to the constraints in (34) which prohibit the factor (33) to be equal to one: as a result, the MAP-LBLA and MAP-LDA do not have the same update equation. However, under some conditions, we could observe that MAP-LBLA update equation in (32) is proportional to that of MAP-LDA in [11, 18], and [19] when using the unnormalized representation. In that case, the EM responsibility vector becomes:

$$\psi_{ijk} \propto \left[\frac{\left(\mathscr{N}_{\theta}^{jk} + \alpha_k - 1\right)}{\left(\sum_k \alpha_k - 1\right) + \left(\sum_{k=1}^{K} \mathscr{N}_{\theta}^{jk}\right)}\right] \times \left[\frac{\left(\mathscr{N}_{\varphi}^{v_{ij}k} + \lambda_{kv} - 1\right)}{\left(\sum_v \lambda_{kv} - 1\right) + \left(\sum_{v=1}^{V} \mathscr{N}_{\varphi}^{v_{ij}k}\right)}\right] \tag{35}$$

From (35), the EM algorithm for MAP-LBLA could be identified with MAP-LDA. We notice that BL prior in (36) contains Beta distribution (38) (the generating density function) that is related to the density generator in (37) [50]:

$$p(\theta_j|\alpha_1, ..., \alpha_K) = \mathscr{G}(\gamma) \prod_{k=1}^{K} \frac{\theta_{jk}^{\alpha_k - 1}}{\Gamma(\alpha_k)} \tag{36}$$

The density generator $\mathscr{G}(.)$ of BL gives:

$$\mathscr{G}(\gamma) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\gamma^{\sum_{k=1}^{K} \alpha_k - 1}} \mathscr{J}(\gamma) \tag{37}$$

Below is the representation of the Beta distribution in BL given its hyperparameters $\alpha$ and $\beta$.

$$\mathscr{J}(\gamma|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \gamma^{\alpha-1}(1-\gamma)^{\beta-1} \tag{38}$$

with $\gamma = \sum_{k=1}^{K} \theta_{jk} < 1$. From (38), we can use (33) and (32) to show that:

$$\left\{ \varphi_{k(V+1)}(1 - \varphi_{k(V+1)}) \rightarrow \sum_{v=1}^{V} \varphi_{kv} \sim Beta(2, 2) \right. \tag{39}$$

since $\mathscr{J}(\gamma|2, 2) \propto \gamma(1-\gamma)$ for $\gamma = \sum_{v=1}^{V} \varphi_{kv}$. Then, we identify the Beta parameters from (39) and (38):

$$\lambda = 2 \quad \eta = 2 \tag{40}$$

for the corpus BL prior. From (30) to (32), we can observe that $\mathscr{J}(\gamma|2, 2) \propto \gamma(1-\gamma)$ for $\gamma = \sum_{v=1}^{V} \theta_{jk}$ as well; so for the document BL prior, we have:

$$\alpha = 2 \quad \beta = 2 \tag{41}$$

As we identify the hyperparameters of the generating density function or the Beta distribution (38), the corpus BL is then defined as BL$(\lambda_{k1}, ..., \lambda_{kV}, 2, 2)$, while the document BL is still BL$(\alpha_1, ..., \alpha_K, 2, 2)$ from (36), (37), (38), (40), and (41). Importantly, during initializations, the only unknown hyperparameters in the MAP-LBLA are the Liouville distribution document parameters $(\alpha_k)_{k=1}^{K}$ and Liouville corpus parameters $(\lambda_{kv})_{v=1}^{V}$. The EM lower bound to MAP estimation therefore simplifies the LBLA structure, which has been reduced from a three-level hierarchical model to two levels. This ultimately suggests that while the formulation of MAP-LBLA in (32) is proportional to the update equation in (35) which bears some resemblance with MAP-LDA [11, 18, 19], we can primarily identify (35) as a Liouville family distribution that turns out to be proportional to the Dirichlet. Since the Liouville family distribution of the second kind is proportional to Dirichlet, then both their update equations under a topic modeling framework would be proportional. This is the case because in (35) by proportionality, the Beta prior defined in (39) acts as a uniform prior. As previously mentioned, when considering proportionality, the MAP-LBLA's update equations could be equivalent to those from MAP-LDA. Instead of using EM statistics in a form of $\mathscr{N}_{\varphi}$, $\mathscr{N}_{\theta}^{j}$, and $\mathscr{N}_{Z}$, we could also represent the EM algorithm for LBLA point estimates in terms of unnormalized counts of the EM responsibilities as:

$$\theta_{jk} \propto \left(\mathscr{N}_{\theta}^{jk} + \alpha_k - 1\right)(1 - \theta_{j(K+1)}) \tag{42}$$

$$\propto \left(\mathscr{N}_{\theta}^{jk} + \alpha_k - 1\right) = \sum_i \psi_{ijk} + \alpha_k - 1 \tag{43}$$

$$\varphi_{kv} \propto \left(\mathscr{N}_{\varphi}^{v_{ij}k} + \lambda_{kv} - 1\right)(1 - \varphi_{(V+1)k}) \tag{44}$$

$$\propto \left(\mathscr{N}_{\varphi}^{v_{ij}k} + \lambda_{kv} - 1\right) = \sum_{ij} \psi_{ijk} + \lambda_{kv} - 1 \tag{45}$$

where the EM statistics for LBLA are:

$$\begin{cases} \mathscr{N}_{\theta}^{jk} = \sum_i \psi_{ijk} \\ \mathscr{N}_{\varphi}^{kv} = \sum_j \psi_{ijk} \\ \mathscr{N}_{Z}^{k} = \sum_{ij} \psi_{ijk} \end{cases} \tag{46}$$

We just showed that with unnormalized count, the LBLA using EM for MAP, and LDA share similar parameters. So we combine unnormalized count method to parameterization to connect the MAP estimation to other inferences such as stochastic variational inference for the LDA architecture. We will show that our proposed approach could be in alignment with the work in [11]. The batch algorithm for MAP using EM for LBLA (BEM-LBLA) follows (7), (9), and (32). It requires an extensive amount of memory because it stores on each word, in the corpus, an EM responsibility vector.

It constantly needs access to all the available data at every iteration before providing an update which is not efficient. We first aim for a fast batch method (similar to CVB0) from which we can build a stochastic EM algorithm for MAP estimation.

### 3.6 Fast batch algorithm for EM-LBLA

It is a refined version of the original batch EM for MAP-LBLA. For time and memory complexity, it is directly faster and provides a good performance over the original batch EM because it excludes the current posterior from its sufficient statistics. It excludes current value of the responsibility for the word $x$. The counts in (47) are then used for batch processing. In the collapsed space, it is equivalent to CVB0. We summarize its expected counts:

$$\begin{cases} \mathscr{N}_{\theta-ij}^{jk} = \sum_{-i} \psi_{ijk} \\ \mathscr{N}_{\varphi-ij}^{v_{ij}k} = \sum_{-j} \psi_{ijk} \\ \mathscr{N}_{Z-ij}^{k} = \sum_{-(i,j)} \psi_{ijk} \end{cases} \tag{47}$$

where the responsibility update equation is defined as:

$$\begin{aligned} \psi_{ijk} \propto & \left[ \frac{\left( \mathscr{N}_{\theta-v_{ij}}^{jk} + \alpha_k - 1 \right)}{\left( \sum_{k=1}^{K} \alpha_k - 1 \right) + \left( \sum_{k=1}^{K} \mathscr{N}_{\theta-v_{ij}}^{jk} \right)} \right] \\ & \times \left[ \frac{\left( \mathscr{N}_{\varphi-v_{ij}}^{v_{ij}k} + \lambda_{kv} - 1 \right)}{\left( \sum_{v} \lambda_{kv} - 1 \right) + \left( \sum_{v=1}^{V} \mathscr{N}_{\varphi-v_{ij}}^{v_{ij}k} \right)} \right] \\ & \times \left(1 - \theta_{j(K+1)}\right)\left(1 - \varphi_{k(V+1)}\right)\left(\varphi_{k(V+1)}\right) \end{aligned} \tag{48}$$

with $ij$ meaning the $i$th word in document $j$; $ijk$ is the $i$th word in document $j$ in topic $k$; $i$ also represents the number of latent variables in the document $j$. From the work in [11] and [18], it shows that SCVB0-LDA is equivalent to MAP-LDA through unnormalized parameterization because the MAP-LDA update equation is identical to the SCVB0 except that their hyperparameters must be offset by one. This equivalent relationship between the MAP-LDA and SCVB0-LDA characterizes the similarity between the EM statistics and responsibilities with CVB0-LDA statistics and variational responsibilities (distributions). The SCVB0-LDA is the unnormalized MAP-LDA using standard EM. Because it implements a stochastic method at word-level using the variational distribution as a local parameter, the SCVB0-LDA is the stochastic unnormalized MAP-LDA using minibatch scheme of size one. In EM, the MAP estimates unnormalized expected sufficient statistics to scale properly its prior distributions for a normalized likelihood function. In this paper and using online stochastic CVB0 for LBLA that we previously proposed in [52], we can observe that there is no equivalent relationship between the currently proposed MAP-LBLA and the SCVB0-LBLA [52] as their respective update equations are different. This is in contrast to the MAP-LDA and the SCVB0-LDA that share some equivalent relationship as shown in [18, 19]. However, our EM-based MAP-LBLA shares some equivalent relationship with SCVB0-LDA under unnormalized parameterization in (32) from [18]. As the current MAP-LBLA's update equation is proportional to the one in MAP-LDA, it is connected to SCVB0-LDA. With unnormalized representation, the EM-based MAP-LBLA associates its EM statistics and responsibilities to those of CVB0-LDA: the EM algorithm for MAP-LBLA therefore operates on unnormalized parameterization of LDA. We have just illustrated that both MAP-LBLA and MAP-LDA operate on unnormalized parameterization of LDA. Therefore, the SCVB0-LDA could characterize a MAP estimation for LBLA as well. Connecting the MAP-LBLA to CVB0-LDA and SCVB0-LDA will help in providing an alternative to a model selection as we show in Sect. 3.8. The proposed MAP-LBLA ultimately optimizes an EM lower bound on the posterior probability of the parameters using EM responsibilities and EM statistics.

### 3.7 Stochastic EM algorithm for MAP-LBLA model

We call this method SEM-LBLA which stands for stochastic EM algorithm for MAP estimation for LBLA topic model. This method does not require all the available samples for an update as in the original batch. It follows a stochastic technique within a minibatch scheme that also refines the fast batch in Sect. 3.6. The standard batch method is slow. This approach provides two update equations for its global parameters: we have the update equation for the document global parameter and the one for the global topics. In our proposed method, SEM-LBLA operates as follow: In minibatch, SEM-LBLA accesses one word $x$ in a corpus (a random and uniform draw), then from that sample, it updates its parameters. In the E-step, it computes unnormalized expected sufficient statistics and evaluates the EM responsibility $\psi_{ij}$ associated to the word $x = x_{ij}$. In the M-step, it evaluates the intermediate global parameters (topics in terms of expected counts) in the corpus as it optimizes the EM lower bound. To do that, it creates $\mathscr{W}$ copies of the intermediate global parameters associated to $x$ in the minibatch and then average them using $\mathscr{V}$. The average estimate of the intermediate global parameters in a minibatch (of size $\mathscr{V}$) scheme using $\mathscr{W}$ copies is generally given as:

$$\hat{\mathscr{N}}_{\varphi} = \frac{\mathscr{W}}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \mathscr{A}^{(i,j)} \tag{49}$$

where $\mathscr{A}^{(i,j)}$ is the word-topic expected $K \times V$ count matrix which $v$th column contains the responsibility

$$\psi_{ij} = \sum_k \psi_{ijk} \tag{50}$$

So for $\mathscr{V} = 1$, we have a minibatch of size one. When $\mathscr{V} > 1$ we have a standard minibatch scheme which draws a subset of samples from the corpus at each iteration. When $\mathscr{W} = \mathscr{V}$ and $\kappa = 0$, we have a batch EM for MAP. However, for a minibatch of size one, the estimate accesses $\mathscr{W}$ copies of the distribution on word $x$; so the estimate becomes:

$$\hat{\mathscr{N}}_\varphi = \mathscr{W} \sum_{v_{ij} \in \mathscr{V}} \mathscr{A}^{(i,j)} = \mathscr{W} \psi_{ij}[v_{ij} = v] \tag{51}$$

This simply counts the number of times the word $v$ appears in the corpus (of size $\mathscr{W}$) as the global intermediate estimate for selecting a random word $v$ in the corpus. Then, this intermediate global parameter estimate is then used to update the global topic parameter as shown in:

$$\mathscr{N}_\varphi[t+1] = (1 - \rho_t)\mathscr{N}_\varphi[t] + \rho_t \hat{\mathscr{N}}_\varphi \tag{52}$$

where $\rho_t = (\tau_0 + t)^{-\kappa}$ is the step size. The variable $\tau_0$ is the number of minibatches (predefined), $t$ is the minibatch index, and $\kappa \in (0.5 \;\; 1]$ is provided by the users. Similarly, in the document $j$, the random and uniform draw of a word in a corpus creates an intermediate global estimate of $\hat{\mathscr{N}}^j_\theta = \mathscr{W}_j \psi_{ij}$ (for $\mathscr{V} = 1$) leading to an update equation of:

$$\mathscr{N}^j_\theta[t+1] = (1 - \rho_t)\mathscr{N}^j_\theta[t] + \rho_t \hat{\mathscr{N}}^j_\theta \tag{53}$$

When $\mathscr{V} > 1$, we use:

$$\begin{cases} \hat{\mathscr{N}}^j_\theta = \frac{\mathscr{W}_j}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \psi_{ij} \\ \hat{\mathscr{N}}_Z = \frac{\mathscr{W}}{\mathscr{V}} \sum_{v_{ij} \in \mathscr{V}} \psi_{ij} \end{cases} \tag{54}$$

We also estimate the expected count $\hat{\mathscr{N}}_Z = \mathscr{W}\psi_{ij}$ (when $\mathscr{V} = 1$) and then summarize its online average equation as:

$$\mathscr{N}_Z[t+1] = (1 - \rho_t)\mathscr{N}_Z[t] + \rho_t \hat{\mathscr{N}}_Z \tag{55}$$

From [18], the SEM-LBLA will converge to the stationary point of the MAP objective function since $0 < \rho_t \leq 1 \, \forall \, t$ and $\sum_t^\infty \rho_t = \infty$ and $\lim_{\infty} \rho_t = 0$. These expectations explain the EM statistics for the MAP-LBLA along with the responsibility vector $\psi_{ijk}$. The parameter estimates at M-step are identical to the expected sufficient statistics from E-step.

### 3.8 Model selection: small samples, number of topics, and vocabulary size under MAP-LBLA

The MAP favors small samples as it can regularize better ML estimates with its prior information. Because it can perform well on small datasets, we expect it to offer a much improved performance when using, for instance, a minibatch processing (stochastic method) compared to full batch methods. For extremely large samples, the MAP estimate will be close to the posterior means (unbiased estimator [12, 26]). However, it will require expensive computational resources [12]. Large samples can cause an increase in the number of parameters, especially the number of topics and vocabulary size. Increasing the number of topics in a finite, parametric topic mixture model is not ideal because such setting automatically increases the search space for the optimal number of topics and vocabulary size [12, 41]. To efficiently reduce the searching space for model selection, we propose a performance of our MAP algorithm using small samples size along with small number of topics and possibly small vocabulary size as well [19]. From previous sections, we showed that our model, the MAP-LBLA, under unnormalized parameterization is equivalent to SCVB0 which uses CVB0 (the zero-order approximation of CVB) as a fast batch method [18]. The CVB0 itself could be restrictive for large-scale processing because of its memory requirement problems at every iteration [12, 18]. It led to a stochastic CVB0 or SCVB0. The work in [12] showed that CVB0 favors a small set of topics because when the hypothesis grows, the CVB0 is unable to find a global optimum as it often gets stuck in local maxima [12, 22]. The SCVBO uses its stochasticity to escape local optima [12, 22]. SCVB0 can operate in large-scale applications (Big Data), but it has no ability in parameter streaming especially when the vocabulary size and number of topics increase in topic-word matrix. To solve this problem the SCVB0 could operate on a small number of topics, and small minibatches, possibly using minibatch of size one. Since the MAP-LBLA is connected to SCVB0-LDA through MAP-LDA, it can therefore carry such implementation to allow both large-scale data and parameter streaming. This explains our decision to operate on small number of topics and samples sizes for MAP-LBLA topic model. In terms of EM algorithm, under unnormalized counts, the MAP-LDA and MAP-LBLA have similar update equations. We can use these characteristics to assess a model selection for our LBLA model through LDA since MAP-LDA and SCVB0 have identical update equations with only their hyperparameters offset by one [11, 18]. In other words, from SCVB0-LDA to MAP-LBLA, the MAP update equation only adds negative one on its hyperparameters. The SCVB0-LDA is therefore the unnormalized representation of online EM for both MAP-LDA and MAP-LBLA. However, from analysis, SCVB0 uses CVB0 as a fast batch method in a stochastic optimization. Despite its use of large memory, the CVB0 could outperform the unbiased estimator CGS when the number of topics is low [12]. As a deterministic method, this allows it to have fast convergence. Finally, for instance, in multi-label framework [12], when only one sample is required, the CVB0 outperforms both the CGS and its unbiased estimator. Since SCVB0 is a stochastic version of CVB0, it carries all the advantages of

CVB0 while improving the memory requirement of CVB0 for large-scale data processing.

In topic modeling, time and memory complexities are functions of the number of topics and the size of the vocabulary [19, 25]. When the size of global parameters increases, especially the word-topic expected count matrix of size $K \times V$, a model selection that efficiently reduces the variables $K$ and $V$ ultimately improves time and memory complexities. Such model selection scheme would implicitly provide an efficient setting for a parameter streaming. We would like to consider improving solutions provided to SCVB0-LDA in model selection with our proposed EM-based MAP-LBLA topic model. This is because our approach is connected to SCVB0 through the MAP-LDA: from the literature, as SCVB0 (with a minibatch scheme that processes one sample at a time [18]) is equivalent to a stochastic unnormalized MAP for LDA, we can therefore set a minibatch of size one for MAP-LBLA as we suggested it earlier in Sect. 3.7 to accommodate data and parameter streaming. This constitutes a direct alternative to [19] that uses a dynamic scheduling approach based on residuals including a buffer mechanism that provides an alternative to model selection which also improves both time and memory complexities. The approach in [19] first reduces the number of topics and vocabulary size in a parametric finite topic mixture model using LDA. Their buffering technique makes it easy to transfer data between computer's memory and external storages that carry the load (massive data including word-topics expected count matrices). This finally leads to a parameter streaming that fixes the problem of big topic modeling in large-scale applications.

Our proposed alternative to model selection using minibatch scheme of size one or reasonable minibatch sizes is in agreement with the core method that is implemented in [19]. Though, ours is more simpler and also allows us to process documents with almost infinite vocabularies: a minibatch of size one ultimately fixes the problem of vocabulary. This is equivalent to processing or updating only the $v$th column of the $K \times V$ word-topic matrix while the corpus expected count increases by one anytime we access a new vocabulary, for instance. We can summarize our contribution as follows: ultimately, with our scheme supporting a small number of topics and a minibatch method of size one, there is no need for a buffer of size $K \times V$ to connect to external storages. This facilitates flow of data and parameter. In fact, the buffering scheme would have required us to probably implement two buffers: one for the global topic matrix and one for the document parameter (documents expected count matrix), and use both simultaneously in inferences within a stochastic framework at word level which defines the topic and document parameters as global parameters. In contrast to the method in [19] which follows a stochasticity at document-level, our approach does not discard the document parameter after one look. It updates both the corpus (topics) and document parameters. We only used the connection between the MAP-LBLA and MAP-LDA and their equivalent relationship within the collapsed variational Bayes inferences to suggest for an improved alternative to model selection for MAP in order to handle both large-scale data and parameter streaming. Our method is not computationally expensive when we compare it to [19] that supports expensive dynamic scheduling and buffering. In our proposed method, despite being stochastic, we also prioritize accurate estimates over extremely fast methods that could miss important processing steps and negatively affect overall results. For a regular minibatch (49), with reasonable small samples, we can use the proposed $|K| \leq 150$ as almost similar to the setting in [12], and for a minibatch scheme of size one (51), we can set $|K| = 10$ for every word as in [19]. We combine both processes in our framework where we use regular minibatch when the parameters and data are manageable or we switch to a minibatch of size one for extremely large vocabulary size in the data and parameters.

**Table 2** Text document datasets

| | $\mathscr{D}$train | $\mathscr{D}$test | $\mathscr{W}$ | V | $\mathscr{D}$ |
|---|---|---|---|---|---|
| NIPS | 1256 | 419 | 2,166,029 | 12,419 | 1675 |
| KOS | 2573 | 857 | 4,67,714 | 6909 | 3430 |
| ENRON | 29,896 | 9965 | 6,400,000 | 28,102 | 39,861 |

# 4 Experimental results and settings

## 4.1 Datasets

We consider three challenging text document datasets: ENRON,[1] NIPS text documents,[2] and KOS blog entries[3] as shown in Table 2. ENRON dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains emails from about 150 users, mostly senior management of Enron, organized into folders and it has a total corpus of $\mathscr{D} = 39,861$ documents. With a vocabulary size $V = 28,102$, it provides a total of $\mathscr{W} = 6,400,000$ words. NIPS data set represents a collection from scientific papers from the proceedings of NIPS database. It has a corpus around 2484 papers. The corpus contains $\mathscr{D} = 1740$ documents for a total $V = 12,419$. It also carries a total of $\mathscr{W} = 2,166,029$ words and $M = 8,36,644$ unique word-document pairs. KOS is from a report blog website (online). It has a total of

[1] http://www.cs.cmu.edu/~enron/

[2] https://cs.nyu.edu/~roweis/data.html

[3] https://archive.ics.uci.edu/dataset/164/bag+of+words

$\mathscr{D} = 3430$ documents, a vocabulary size of $V = 6909$, and a total of $\mathscr{W} = 4,67,714$ words and $M = 3,60,664$ unique word-document pairs.

## 4.2 Implementation

This is a stochastic EM algorithm for MAP estimation using the LBLA topic model. As we perform a stochastic at word-level method in our proposed approach, we have two global parameters to estimate instead of one global parameter as in case of a stochasticity at document level. Our global parameters include the topic-word parameters and the document parameters. We estimate these parameters in terms of unnormalized expected counts which define our EM statistics for the stochastic EM-based LBLA model for MAP estimation. The M-step optimizes the EM lower bound with respect to the parameters, while the E-step provides the unnormalized expected sufficient statistics as we use here exponential family distributions. The proposed approach requires initializations on the hyperparameters. We usually set them randomly. However, for the BL hyperparameters, we also provide initializations as follows: For BL prior on the document multinomial parameter, we choose $\alpha_{jk} = \frac{1}{k}$ where $k \in \{1, 2, ..., K\}$ to characterize asymmetric BL prior. We set $\alpha_j = 2$ based on (41), and we choose $\alpha_{jk}$ such that $\alpha_j - \sum_{k=1}^{K} \alpha_{jk} \neq 0$. Then, we choose $\beta_j = 2$. For the BL on the corpus multinomial parameter, we are setting values for $\lambda_{kv}$ with $v \in \{1, 2, ..., V\}$ (similar to the document BL) and $\lambda = 2$ (where $\lambda - \sum_{v=1}^{V} \lambda_{kv} \neq 0$) and $\eta = 2$ from (40) for every $k$. We use a stochasticity at word-level where we randomly sample one word at a time from which we estimate its EM responsibility vector (local parameter) $\psi_{ijk}$ that allows us to obtain estimates of the model parameters in terms of expected counts.

We implement a minibatch method of size one to process one sample at a time. We use regular minibatch (multiple samples) when the parameters and data are manageable or we can also switch to a minibatch of size one for extremely large vocabulary size in the data. This illustrates the flexibility of our framework to large-scale applications. At convergence, the global parameters are approximated as point estimates. The method still favors much smaller batch size so that the prior regularizes estimates. We set the minibatch sizes as: $\mathscr{V} = \{10, 40, 60, 80, 100\}$. The set of topics is: $K = \{10, 20, 40, 60, 80, 100, 120, 150\}$. We provided a learning rate $\rho_t$ at iteration $t$ such that:

$$\rho_t = (t + \tau_0)^{-\kappa} \tag{56}$$

The forgetting rate $\kappa \in (0.5, 1]$ actively controls how quickly previously estimated data are forgotten, during successive iterations. With EM algorithm, we can always reach a local optimum of the EM lower bound of the posterior distribution of the parameters. We maintain $\tau_0 = 1$ and $\kappa = 0.7$.

### 4.2.1 Evaluation method using perplexity

Each of the three datasets selected for this experiment went through similar process. In each dataset (a collection of text documents), we randomly divide the data into training and testing sets. We compute the corpus parameters $\varphi$ during the training phase. Then, in the test document, we randomly divide it into a ratio of 90% and 10% as each subset contains word tokens. As we fix $\varphi$, we estimate the document topic proportions $\theta$ on the 90% of the test set and then calculate the predictive perplexity on the rest 10% of the subset using (57) in [19].

A low value of the predictive perplexity or a high predictive log likelihood suggests a better model.

$$\text{perplexity} = \exp\left\{ -\frac{\sum_{i,j} x_{ij} \log[\sum_k \psi_{ijk}]}{\sum_{ij} x_{ij}} \right\} \tag{57}$$

The variables $x_{ij}$ and $[\sum_k \psi_{ijk}]$ represent the data and responsibility at 10%, respectively. We compute the responsibility vector $\psi_{ijk}$ using (32) from our EM statistics while $\varphi$ is maintained fixed. Since time and memory complexities are functions of the parameters such as $K$ and $V$, and the size of the dataset $\mathscr{D}$ [19, 25], when $K$, $V$, and $\mathscr{D}$ become extremely small, they can significantly improve the memory requirement (with stochastic method) and the time complexity. The possibility in our case to carry extremely small samples makes it a better approach over the LDA. It also makes online method efficient over batch techniques.

### 4.2.2 Time and memory complexities

The proposed online EM-based-MAP-LBLA has similar time and memory complexity to LDA topic model in general [19, 25]. Especially, the work in [19] has provided an extensive detail on LDA's time and memory complexities. Though, the main difference between the LDA and LBLA's time and memory complexities is the flexibility of the BL that allows the model to perform many tasks: its covariance structure offers possibility to model selection easier than the one in LDA when analyzing topic structure based on probability masses (topic proportions) associated to global topics in LBLA. The LDA has no ability to topic correlation analysis as we mentioned earlier. Therefore, our model is much faster because it can handle more tasks than LDA including performing topic correlation analysis; all these tasks within the same time of LDA. This suggests that online EM-based MAP-LBLA is faster at performing each task and therefore has a much improved time complexity compared to its LDA counterpart per task. Furthermore, with flexible priors such as BL, it means we do not need too much samples including the number of topics to achieve better estimates as the MAP improves and regularizes our point estimates.

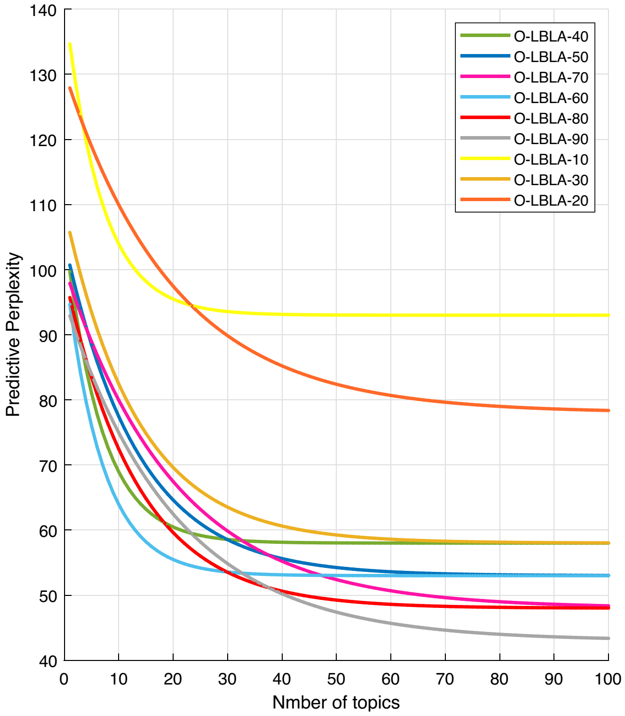


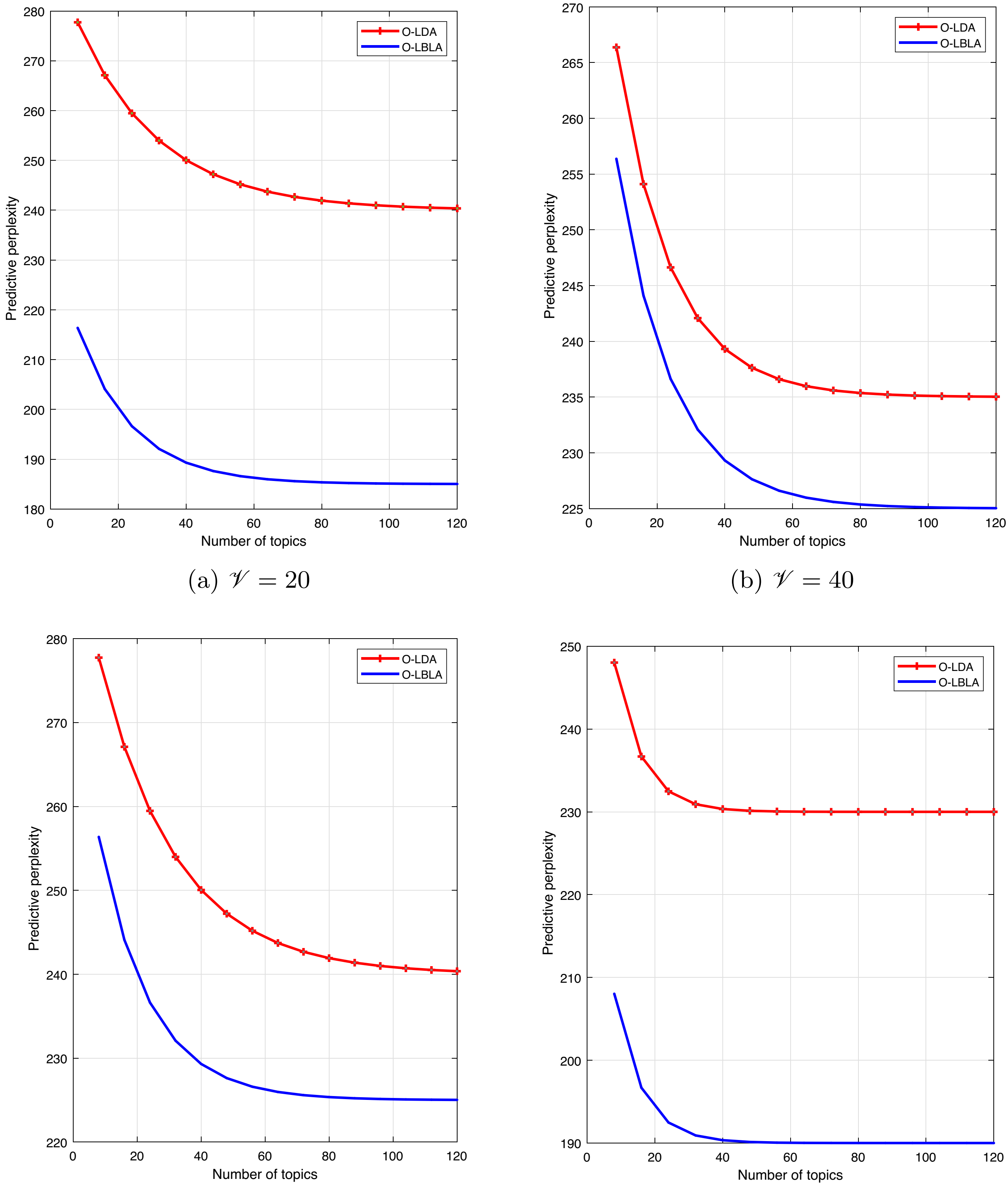
**Fig. 1** Online MAP-LBLA and NIPS batch sizes

### 4.3 Results

The use of prior distributions for MAP estimation makes PLSA and mixture of unigrams unfit for comparison because the work in [10, 15] even used the simple symmetric LDA to show the limitations of the PLSA and mixture of unigrams as they lack prior information. In this experiment, we mainly focus on topic models that could characterize a Bayesian framework. We compare the performance of our LBLA topic model directly to LDA for MAP estimation. We then use the predictive perplexity to evaluate the online EM algorithm for MAP-LBLA and MAP-LDA under a variety of situations: in each dataset, we monitor the influence of the number of topics and batch size in the predictive perplexity. In each dataset we observe that online EM for MAP-LBLA is faster than online EM for MAP-LDA because of its ability to summarize relevant topics faster than symmetric LDA. Importantly, we observe that the predictive perplexity favors a small number of topics as we assess the first topic values to which the perplexity remains constant while being at its lowest values. The online EM for MAP-LBLA constantly outperforms online EM for MAP-LDA in terms of predictive perplexity. Figures 1, 2, and 3 show the performance of the LBLA over the symmetric LDA in each of our proposed datasets. The flexibility of the BL prior in LBLA also plays a central role in the predictive distributions and perplexity: a topic model in general has a fixed multinomial distribution as likelihood function. Its robustness relies on the choice of priors such as BL. The symmetric prior with a uniform base measure does not offer a variability in the set of topics while the asymmetric BL prior provides heterogeneity in the topics that speeds up the search for most relevant topics. In addition, the use of uniform priors such as symmetric Dir, while it simplifies computation, reduces the MAP framework to MLE. Within the MAP-LBLA topic models, we also observe that providing a reasonable batch size ultimately enhances the predictive performance in our datasets as shown in Figs. 4, 5, and 6 where different batch sizes varying from 10 to 90 have been considered. This is because a reasonable size of samples could benefit from the contribution of prior information to provide a better modeling and then an enhanced perplexity, yet obviously better perplexities are obtained the beginning when starting with relatively larger batch sizes as shown in the figures. In this case, the MAP could act as regularizer through the prior for small sample sizes. These characteristics in the proposed approach improve point estimates and contribute to a much robust perplexity framework. It is also important to mention that in many occasions, the predictive perplexity of the MAP-LDA is almost close to that of the MAP-LBLA as shown in Figs. 2b, 6a, and 6c. This could be explained by the hyperparameter setting in LBLA. The LBLA is a generalization of LDA which means under some conditions (hyperparameter initialization) the LBLA could be reduced to LDA topic model. The MAP-LBLA favoring a small number of topics and a relatively reasonable batch size show its equivalent relationship with CVB0 that also favors small number of topics [12].

## 5 Conclusion

In this paper, for parameter estimation in topic modeling, we provide an alternative to the collapsed variational Bayes and collapsed Gibbs inferences by proposing a simple MAP estimation technique based on standard EM algorithm. The method optimizes an EM lower bound on the posterior distribution of the parameters in the M-step. In the E-step, it updates exponential family sufficient statistics using online averages. Our main parameters are the unnormalized expected counts (EM statistics) that summarize the MAP-LBLA's update equation. The CVB and CGS, the collapsed space inferences, marginalize out the parameters while leaving the latent variables. On the other hand, the MAP estimation method integrates out the latent variables leaving only the parameters. It also reduces the three-level hierarchical structure in topic models to two levels in the hierarchy. We implement the MAP-LBLA using online EM algorithm and then compare its performance (predictive perplexity) against the MAP-LDA that is with equipped symmetric Dir. We show that the update equation of MAP-LBLA could be proportional to that of MAP-LDA. The MAP-LDA is



(a) $\mathscr{V} = 20$

(b) $\mathscr{V} = 40$

(c) $\mathscr{V} = 60$

(d) $\mathscr{V} = 80$

**Fig. 2** Online EM-based MAP-LBLA versus online EM-based MAP-LDA at different minibatch sizes (NIPS dataset)

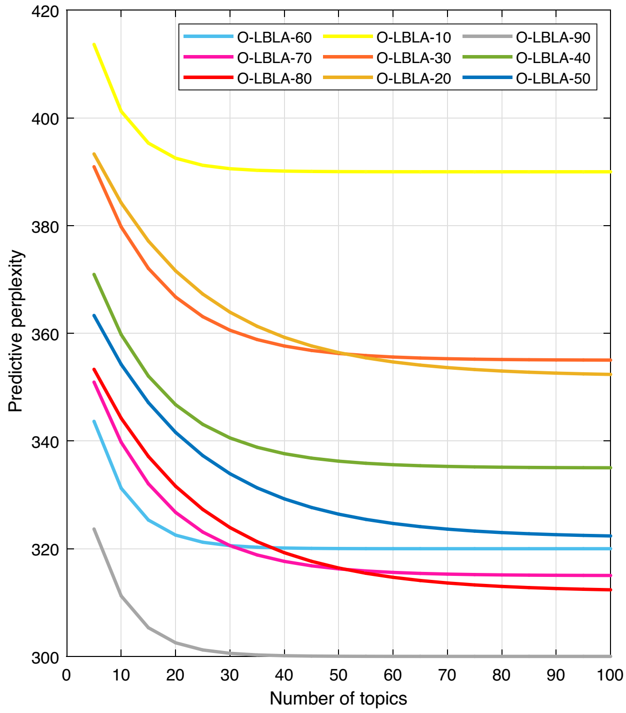


**Fig. 3** Online MAP-LBLA and KOS batch sizes

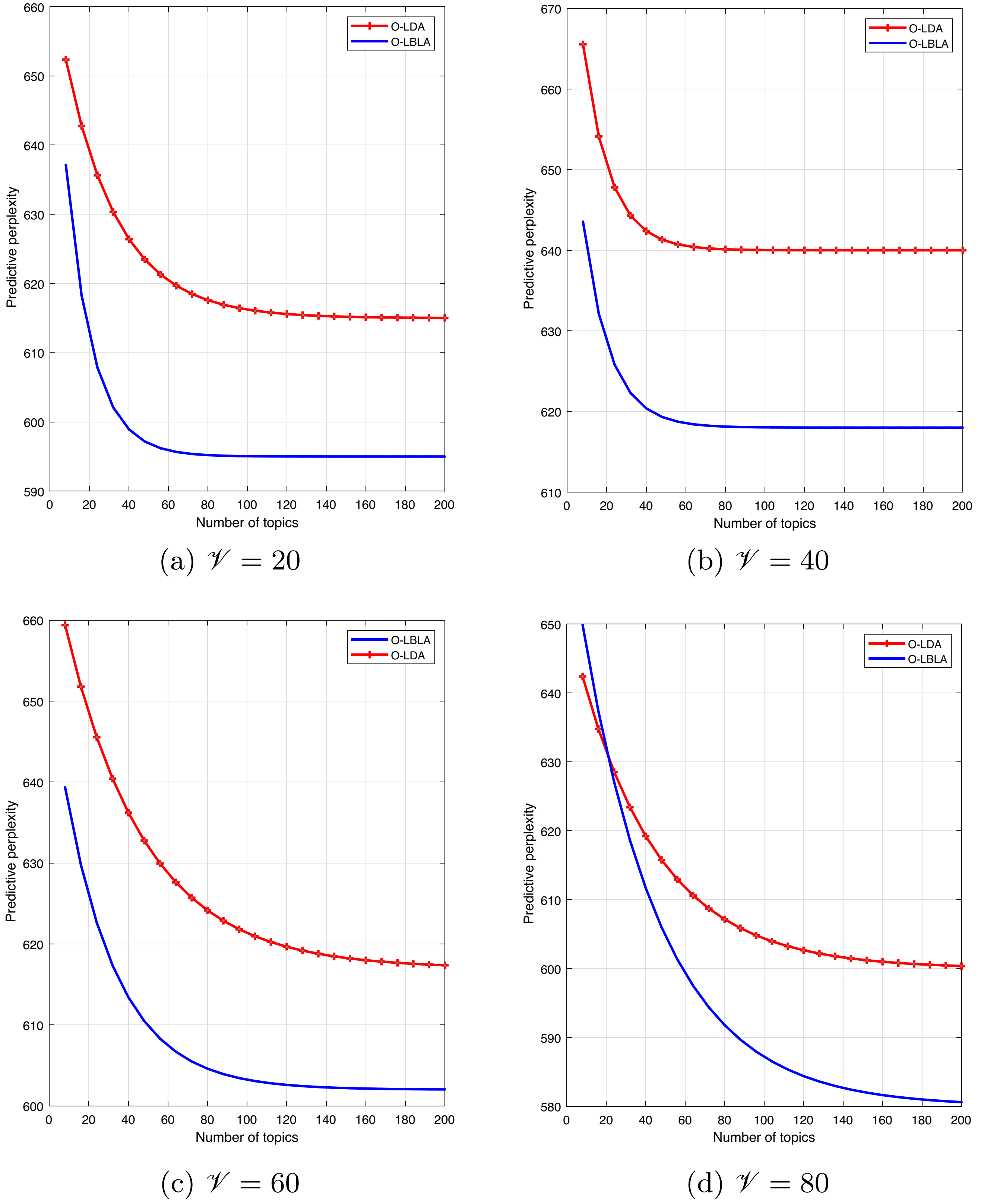
connected to CVB0 because they have identical update equations with only their hyperparameters adjusted or offset by one. The CVB0 favors a small number of topics. The stochastic CVB0 (SCVB0) allows large-scale data modeling but could not handle parameter streaming due to the size of vocabulary and number of topics as they increase in large-scale processing. The MAP-LBLA (which is connected to MAP-LDA) aims to improve the capability of SCVB0-LDA that has an equivalent relationship with MAP-LDA: under unnormalized parameterization, the SCVB0-LDA is equivalent to MAP-LDA. Furthermore, using reasonable samples sizes in the minibatch scheme ultimately fixes the problem related to large parameter matrices especially the word-topic expected count matrix during inferences. We manage the data and parameter streaming by creating a framework where we use regular minibatch when the parameters and data are manageable or we switch to a minibatch of size one for extremely large vocabulary sizes in the data. Because the number of topics and vocabulary size are reduced in this way, the memory and time complexities are much improved in the proposed approach. We also think that the efficiency in the predictive perplexities is due to the flexibility of the BL prior in LBLA compared to the Dir distribution in LDA. Its ability to model dependency between documents through topic correlation characterizes a much robust compression algorithm and predictive models. It is still important to recognize that, in general, one of the problems in parametric finite topic mixture models is the parameters initializations, especially the number of topics. In addition, these models seem to have a much reduced hypothesis space that do not allow them to cope with extremely large number of topics. For future work, we could investigate the performance of the topic model when using other flexible conjugate priors such as generalized Dirichlet based on hyperparameter estimation. Similarly, we could also implement non-conjugate priors using, for instance, logistic normal distributions. Another alternative to finite mixture topic models would be to implement nonparametric models.

## Appendix

We formulate the EM lower bound for MAP-LBLA where the priors $\mathscr{L}(\psi_{ijk}, \theta, \varphi)$ are BL distributions.

$$
\begin{aligned}
\mathscr{L}(\psi_{ijk}, \theta, \varphi) \propto \sum_{k,i,j,v} & \psi_{ijk} \log \varphi_{kv} + \psi_{ijk} \log \theta_{jk} \\
& + \psi_{ij(K+1)} \log(\theta_{j(K+1)}) + \psi_{ijk} \log(\varphi_{k(V+1)}) \\
& + \Bigg\{ \Bigg( \sum_{j,k} (\alpha_k - 1) \log \theta_{jk} \Bigg) \\
& + \Bigg( \alpha - \sum_{k} \alpha_k \Bigg) \log \Bigg( \sum_{j,k} \theta_{jk} \Bigg) \\
& + (\beta - 1) \log \Bigg( 1 - \sum_{j,k} \theta_{jk} \Bigg) + \log \Gamma \Bigg( \sum_{k=1} \alpha_k \Bigg) \\
& + \log \Gamma(\alpha + \beta) \\
& - \log \Gamma(\alpha) - \log \Gamma(\beta) - \sum_{k} \log \Gamma(\alpha_k) \Bigg\} \\
& + \Bigg\{ \Bigg( \sum_{k,v} (\lambda_{kv} - 1) \log \varphi_{kv} \Bigg) \\
& + \Bigg( \lambda - \sum_{v} \lambda_{kv} \Bigg) \log \Bigg( \sum_{k,v} \varphi_{kv} \Bigg) \\
& + (\eta - 1) \log \Bigg( 1 - \sum_{k,v} \varphi_{kv} \Bigg) \\
& + \log \Gamma \Bigg( \sum_{v} \lambda_{kv} \Bigg) + \log \Gamma(\lambda + \eta) - \log \Gamma(\lambda) \\
& - \log \Gamma(\eta) - \sum_{v} \log \Gamma(\lambda_{kv}) \Bigg\}
\end{aligned}
\tag{1}
$$

We perform a coordinate ascent method to obtain the parameter update equations: we characterize the lower bound associated to each parameter, compute the corresponding derivative and set it equal to zero. We added the Lagrangian term to the lower bound to include the optimizations constraints for the parameters before derivation.



(a) $\mathscr{V} = 20$

(b) $\mathscr{V} = 40$

(c) $\mathscr{V} = 60$

(d) $\mathscr{V} = 80$

**Fig. 4** Online EM-based MAP-LBLA versus online EM-based MAP-LDA at different minibatch sizes (KOS dataset)

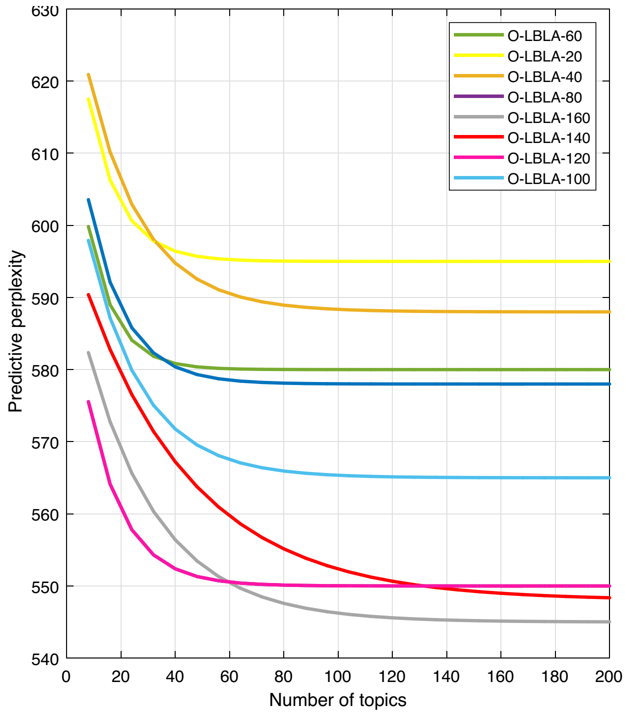


**Fig. 5** Online MAP-LBLA and ENRON batch sizes

$$\begin{aligned}\mathscr{L}(\theta) = &\sum_{k,i,j,v} \psi_{ijk}\left(\log \theta_{jk}\right) \\ &+ \left\{\left(\sum_{j,k}(\alpha_k - 1)\log \theta_{jk}\right)\right\} \\ &+ \left(\alpha - \sum_k \alpha_k\right)\log\left(\sum_{j,k}\theta_{dk}\right) \\ &+ (\beta - 1)\log\left(1 - \sum_{j,k}\theta_{jk}\right) \\ &+ \xi\left(\theta_{j(K+1)} + \sum_{k=1}^{K}\theta_{jk}\right)\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\partial}{\partial \theta_{jk}}\mathscr{L}(\theta) = &\frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} + \frac{\alpha - \sum_k \alpha_k}{\sum_{k,j}\theta_{jk}} \\ &+ \frac{1-\beta}{1 - \sum_{k,j}\theta_{jk}} + \xi\end{aligned} \tag{3}$$

Let $\mathscr{T}$ be defined as: $\mathscr{T} = \frac{\alpha - \sum_k \alpha_k}{\sum_{k,d}\theta_{dk}} + \frac{1-\beta}{1-\sum_{k,j}\theta_{jk}}$, so we can see that $\mathscr{T}$ is not defined when $\sum_{k,j}\theta_{jk} = 0$ and $1 - \sum_{k,j}\theta_{jk} = 0$; $\sum_{k,j}\theta_{jk} \neq 0$ and $(1 - \sum_{k,j}\theta_{jk}) = \theta_{j(K+1)} \neq 0$, $\mathscr{T} = 0$ means

$$\alpha = \sum_k \alpha_k \quad \beta = 1 \tag{4}$$

So we have:

$$\frac{\partial}{\partial \theta_{jk}}\mathscr{L}(\theta) = \frac{\sum_n \psi_{ijk} + \alpha_k - 1}{\theta_{jk}} + \mathscr{T} + \xi \tag{5}$$

Now making the derivative equal to zero gives $\frac{\sum_n \psi_{ijk}+\alpha_k-1}{\theta_{jk}} + \mathscr{T} + \xi = 0$ or $\frac{\sum_n \psi_{ijk}+\alpha_k-1}{\theta_{jk}} = -\mathscr{T} - \xi$; so $\theta_{dk} = \frac{\sum_n \psi_{ijk}+\alpha_k-1}{-C-\xi}$where$\sum_k \theta_{jk} = \sum_k \frac{\sum_n \psi_{ijk}+\alpha_k-1}{-\mathscr{T}-\xi} = 1 - \theta_{j(K+1)}$; $-\xi = \frac{\sum_k(\sum_i \psi_{ijk}+\alpha_k-1)+\mathscr{T}(1-\theta_{j(K+1)})}{1-\theta_{j(K+1)}}$;

$\theta_{jk} = \frac{\sum_i \psi_{ijk}+\alpha_k-1}{\frac{\sum_k(\sum_i \psi_{ijk}+\alpha_k-1)+\mathscr{T}(1-\theta_{j(K+1)})}{1-\theta_{j(K+1)}} - \mathscr{T}}$ or

$$\theta_{jk} = \frac{\sum_i \psi_{ijk} + \alpha_k - 1}{\frac{\sum_k\left(\sum_n \psi_{ijk}+\alpha_k-1\right)+\mathscr{T}(1-\theta_{j(K+1)})-\mathscr{T}(1-\theta_{j(K+1)})}{1-\theta_{j(K+1)}}} \tag{6}$$

$\theta_{jk} = \frac{\sum_n \psi_{ijk}+\alpha_k-1}{\frac{\sum_k(\sum_n \psi_{ijk}+\alpha_k-1)}{1-\theta_{j(K+1)}}} = \frac{\sum_n \psi_{ijk}+\alpha_k-1}{\sum_k \sum_i \psi_{ijk}+\alpha_k-1}(1 - \theta_{j(K+1)})$

For $\mathscr{N}_\theta^{jk} = \sum_i \psi_{ijk}$

$$\theta_{jk} = \frac{\left(\mathscr{N}_\theta^{jk} + \alpha_k - 1\right)}{\left(\sum_k \alpha_k - 1\right) + \left(\sum_k \mathscr{N}_\theta^{jk}\right)}(1 - \theta_{j(K+1)}) \tag{7}$$

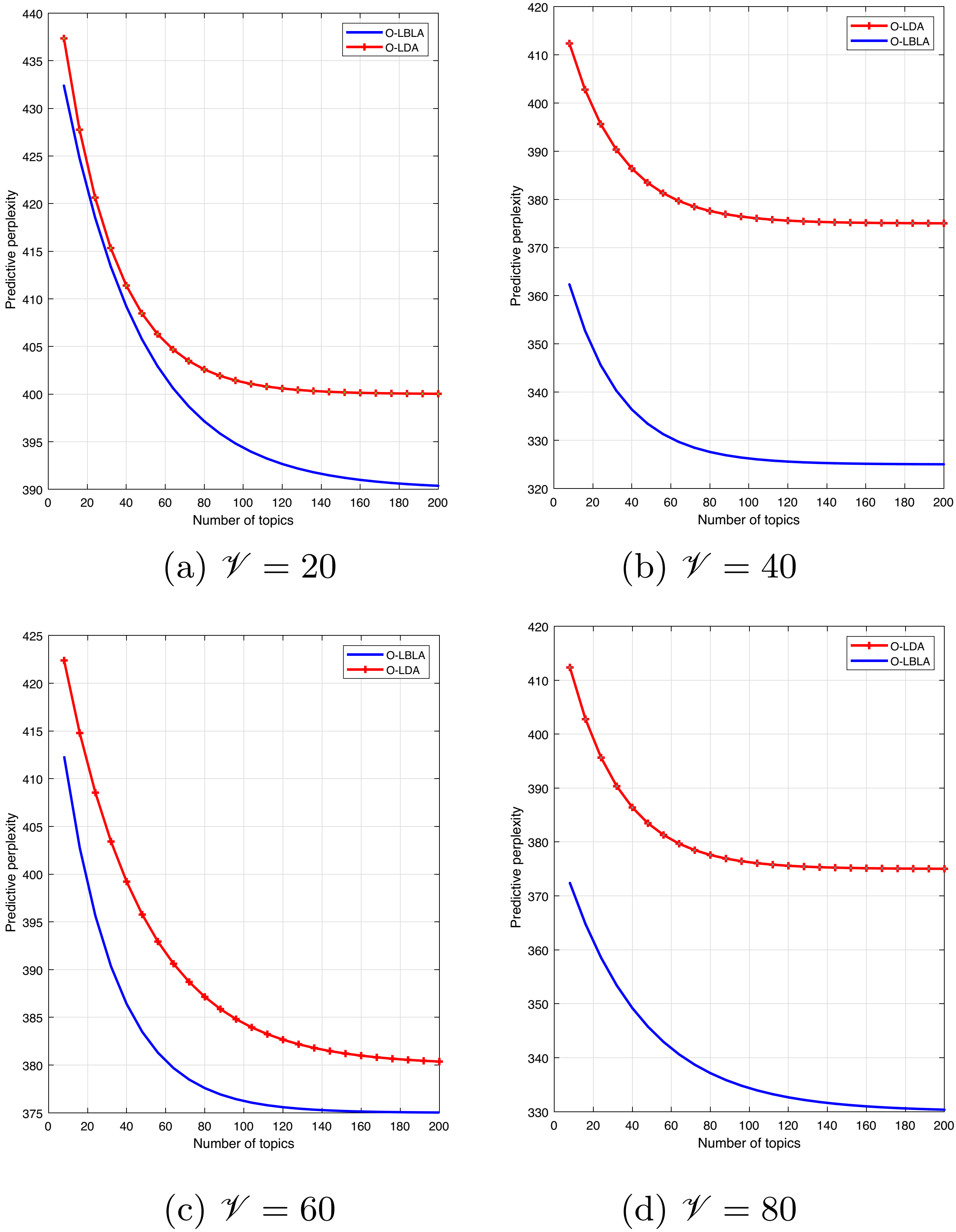
We have also

$$\begin{aligned}\mathscr{L}(\varphi) = &\sum_{k,i,j,v}\psi_{ijk}\left(\log \varphi_{kv}\right) \\ &+ \left(\sum_{k,v}(\lambda_{kv} - 1)\log \varphi_{kv}\right) \\ &+ \left(\lambda - \sum_v \lambda_{kv}\right)\log\left(\sum_{k,v}\varphi_{kv}\right) \\ &+ (\eta - 1)\log\left(1 - \sum_{k,v}\varphi_{kv}\right) + \varrho\left(\varphi_{(V+1)k} + \sum_{v=1}^{V}\varphi_{kv}\right)\end{aligned} \tag{8}$$

Similarly for $\varphi_{kv}$ using $\mathscr{L}(\varphi)$, we have:

$$\varphi_{kv} = \frac{\left(\mathscr{N}_\varphi^{v_{ij}k} + \lambda_{kv} - 1\right)}{\left(\sum_v \lambda_{kv} - 1\right) + \left(\sum_v \mathscr{N}_\varphi^{v_{ij}k}\right)}(1 - \varphi_{k(V+1)}) \tag{9}$$

We define $\Omega$ similar to $\mathscr{T}$ as $\Omega = \frac{\lambda - \sum_v \lambda_{kv}}{\sum_{k,v}\varphi_{kv}} + \frac{1-\eta}{1-\sum_{k,v}\varphi_{kv}}$ with $\mathscr{N}_\varphi^{vjk} = \sum_{(ij)=v}\psi_{ijk}$ where the $i$th word is $v$ with $1 - \theta_{j(K+1)} = \sum_{k=1}^{K}\theta_{jk}$ and $1 - \varphi_{k(V+1)} = \sum_{v=1}^{V}\varphi_{kv}$



(a) $\mathscr{V} = 20$

(b) $\mathscr{V} = 40$

(c) $\mathscr{V} = 60$

(d) $\mathscr{V} = 80$

**Fig. 6** Online EM-based MAP-LBLA versus online EM-based MAP-LDA at different minibatch sizes (ENRON dataset)

**Data Availability** Data could be made available on reasonable request.

## Declarations

**Conflict of interest** No conflict of interest to declare

## References

1. Elkan C (2006) Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In: Proceedings of the 23rd international conference on machine learning, pp 289–296
2. Bouguila N (2008) Clustering of count data using generalized Dirichlet multinomial distributions. IEEE Trans Knowl Data Eng 20(4):462–474
3. Bouguila N (2011) Count data modeling and classification using finite mixtures of distributions. IEEE Trans Neural Netw 22(2):186–198
4. Blei D, Lafferty J (2006) Correlated topic models. Adv Neural Inf Process Syst 18:147
5. Zheng W, Liu Y, Lu H, Tang H (2017) Discriminative topic sparse representation for text categorization. In: 10th International symposium on computational intelligence and design, ISCID 2017, Hangzhou, China, December 9–10, 2017, vol 1. IEEE, pp 454–457
6. Yang S, Zhang H (2018) Text mining of twitter data using a latent Dirichlet allocation topic model and sentiment analysis. Int J Comput Inf Eng 12(7):525–529
7. Xiong S, Wang K, Ji D, Wang B (2018) A short text sentiment-topic model for product reviews. Neurocomputing 297:94–102
8. Yang Y, Jia J, Zhang S, Wu B, Chen Q, Li J, Xing C, Tang J (2014) How do your friends on social media disclose your emotions? In: Brodley CE, Stone P (eds) Proceedings of the twenty-eighth AAAI conference on artificial intelligence. AAAI Press, pp. 306–312
9. Prasad KR, Mohammed M, Noorullah R (2019) Visual topic models for healthcare data clustering. Evolut Intell 14:1–17
10. Blei DM (2004) Probabilistic models of text and images. PhD thesis, University of California, Berkeley
11. Asuncion AU, Welling M, Smyth P, Teh YW (2009) On smoothing and inference for topic models. In: Bilmes JA, Ng AY (eds) UAI 2009, Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, Montreal, QC, Canada, June 18–21, 2009. AUAI Press, pp 27–34
12. Papanikolaou Y, Foulds JR, Rubin TN, Tsoumakas G (2017) Dense distributions from sparse samples: improved Gibbs sampling parameter estimators for LDA. J Mach Learn Res 18:1–58
13. Bouguila N (2009) A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. IEEE Trans Knowl Data Eng 21(12):1649–1664
14. Ali S, Bouguila N (2022) Maximum a posteriori approximation of hidden Markov models for proportional sequential data modeling with simultaneous feature selection. IEEE Trans Neural Netw Learn Syst 33(10):5590–5601
15. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
16. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235
17. Minka TP, Lafferty JD (2013) Expectation-propogation for the generative aspect model. CoRR arXiv:1301.0588
18. Foulds J, Boyles L, DuBois C, Smyth P, Welling M (2013) Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 446–454
19. Zeng J, Liu Z-Q, Cao X-Q (2015) Fast online EM for big topic modeling. IEEE Trans Knowl Data Eng 28(3):675–688
20. Yao L, Mimno D, McCallum A (2009) Efficient methods for topic model inference on streaming document collections. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp 937–946
21. Gao Y, Chen J, Zhu J (2016) Streaming Gibbs sampling for LDA model. arXiv preprint arXiv:1601.01142
22. Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. J Mach Learn Res 14(1):1303–1347
23. Hoffman M, Bach FR, Blei DM (2010) Online learning for latent Dirichlet allocation. In: Advances in neural information processing systems, pp 856–864
24. Robbins H, Monro S et al (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407
25. Teh YW, Newman D, Welling M (2007) A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Advances in neural information processing systems, pp 1353–1360
26. Burkhardt S, Kramer S (2017) Online sparse collapsed hybrid variational-Gibbs algorithm for hierarchical Dirichlet process topic models. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 189–204
27. Ihou KE, Bouguila N (2018) A smoothed latent generalized Dirichlet allocation model in the collapsed space. In: IEEE 61st International midwest symposium on circuits and systems, MWSCAS, pp 877–880
28. Katz SM (1996) Distribution of content words and phrases in text and language modelling. Nat Lang Eng 2(1):15–59
29. Church KW, Gale WA (1995) Poisson mixtures. Nat Lang Eng 1(2):163–190
30. Bouguila N (2007) Spatial color image databases summarization. In: 2007 IEEE International conference on acoustics, speech and signal processing—ICASSP '07, vol 1, pp 953–956
31. Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 113–120
32. Chi R, Wu B, Wang L (2018) Expert identification based on dynamic LDA topic model. In: 2018 IEEE Third international conference on data science in cyberspace (DSC). IEEE, pp 881–888
33. Dieng AB, Ruiz FJR, Blei DM (2019) The dynamic embedded topic model. CoRR arXiv:1907.05545
34. Espinoza I, Mendoza M, Ortega P, Rivera D, Weiss F (2018) Viscovery: trend tracking in opinion forums based on dynamic topic models. CoRR arXiv:1805.00457
35. Putthividhya DP, Attias HT, Nagarajan S (2009) Independent factor topic models. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 833–840
36. Putthividhya DP (2010) A family of statistical topic models for text and multimedia documents. PhD thesis, University of California at San Diego
37. Liu L, Huang H, Gao Y, Zhang Y, Wei X (2019) Neural variational correlated topic modeling. In: The World Wide Web conference. ACM, pp 1142–1152
38. Xun G, Li Y, Zhao WX, Gao J, Zhang A (2017) A correlated topic model using word embeddings. In: IJCAI, pp 4207–4213
39. Wallach HM, Mimno D, McCallum A (2009) Rethinking LDA: why priors matter. In: Proceedings of the 22nd international conference on neural information processing systems. Curran Associates Inc, pp 1973–1981
40. Leng B, Zeng J, Yao M, Xiong Z (2015) 3D object retrieval with multitopic model combining relevance feedback and LDA model. IEEE Trans Image Process 24(1):94–105

41. Ihou KE, Bouguila N (2019) Variational-based latent generalized Dirichlet allocation model in the collapsed space and applications. Neurocomputing 332:372–395
42. Fan W, Bouguila N (2013) Learning finite Beta-Liouville mixture models via variational Bayes for proportional data clustering. In: Rossi F (ed) IJCAI 2013, Proceedings of the 23rd international joint conference on artificial intelligence, Beijing, China, August 3–9, 2013, pp 1323–1329
43. Bouguila N (2012) Infinite Liouville mixture models with application to text and texture categorization. Pattern Recognit Lett 33(2):103–110
44. Fan W, Bouguila N (2013) Online learning of a Dirichlet process mixture of Beta-Liouville distributions via variational inference. IEEE Trans Neural Netw Learn Syst 24(11):1850–1862
45. Epaillard E, Bouguila N (2016) Proportional data modeling with hidden Markov models based on generalized Dirichlet and Beta-Liouville mixtures applied to anomaly detection in public areas. Pattern Recognit 55:125–136
46. Bouguila N (2013) On the smoothing of multinomial estimates using Liouville mixture models and applications. Pattern Anal Appl 16(3):349–363
47. Rahman MH, Bouguila N (2021) Efficient feature mapping in classifying proportional data. IEEE Access 9:3712–3724
48. Mimno D, Hoffman M, Blei D (2012) Sparse stochastic inference for latent dirichlet allocation. arXiv preprint arXiv:1206.6425
49. Li W, McCallum A (2006) Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 577–584
50. Bouguila N (2012) Hybrid generative/discriminative approaches for proportional data modeling and classification. IEEE Trans Knowl Data Eng 24(12):2184–2202
51. Bakhtiari AS, Bouguila N (2014) Online learning for two novel latent topic models. In: Information and communication technology: second IFIP TC 5/8 international conference, ICT-EurAsia 2014, Bali, Indonesia, April 14–17, 2014, Proceedings, vol 8407. Springer, p 286
52. Ihou KE, Bouguila N (2020) Stochastic topic models for large scale and nonstationary data. Eng Appl Artif Intell 88:103364
53. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
54. Wang C, Paisley J, Blei D (2011) Online variational inference for the hierarchical Dirichlet process. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp 752–760
55. Cappé O, Moulines E (2009) On-line expectation-maximization algorithm for latent data models. J R Stat Soc Ser B (Stat Methodol) 71(3):593–613

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.