Online Sparse Collapsed Hybrid Variational-Gibbs Algorithm for Hierarchical Dirichlet Process Topic Models

Sophie Burkhardt and Stefan Kramer

Johannes Gutenberg-Universität Mainz
{burkhardt,kramer}@informatik.uni-mainz.de

Abstract. Topic models for text analysis are most commonly trained using either Gibbs sampling or variational Bayes. Recently, hybrid variational-Gibbs algorithms have been found to combine the best of both worlds. Variational algorithms are fast to converge and more efficient for inference on new documents. Gibbs sampling enables sparse updates since each token is only associated with one topic instead of a distribution over all topics. Additionally, Gibbs sampling is unbiased. Although Gibbs sampling takes longer to converge, it is guaranteed to arrive at the true posterior after infinitely many iterations. By combining the two methods it is possible to reduce the bias of variational methods while simultaneously speeding up variational updates. This idea has previously been applied to standard latent Dirichlet allocation (LDA). We propose a new sampling method that enables the application of the idea to the nonparametric version of LDA, hierarchical Dirichlet process topic models. Our fast sampling method leads to a significant speedup of variational updates as compared to other sampling methods. Experiments show that training of our topic model converges to a better log-likelihood than previously existing variational methods and converges faster than Gibbs sampling in the batch setting.

1 Introduction

Topic models based on latent Dirichlet allocation (LDA) are a common tool for analyzing large collections of text. They are used for extracting common themes and provide a probabilistic clustering of documents. The topics, usually represented by word clouds of frequent words, can be interpreted and used to understand the content of a text corpus. Since scalability is an important factor when modeling large datasets, various online algorithms have been developed to handle streams of documents. A generalization of LDA that allows for asymmetric topic priors and an unbounded number of topics is provided by nonparametric topic models. These are based on hierarchical Dirichlet processes (HDPs) [16].

The two main algorithms for training LDA topic models are Gibbs sampling and variational Bayes. While Gibbs sampling is an unbiased method, it takes very long on large datasets to converge. To make Gibbs sampling online, one has

to use particle samplers which are rather inefficient. Variational Bayes on the other hand may be combined with stochastic gradients to be trained online.

Hybrid methods have gained popularity in recent years, especially in deep neural networks where black box variational inference is a very efficient training algorithm [11]. Sampling can be used to approximate the gradient in variational Bayes which leads to a second source of stochasticity (in addition to random choices of data subsets). In previous work this was applied to parametric topic models [9] and a very efficient variant of this algorithm was recently proposed that takes advantage of the sparsity in topic distributions during sampling [14]. Hybrid algorithms have the combined advantages of a reduced bias of the variational method and a faster convergence as compared to pure Gibbs sampling, as well as the possibility of online training through stochastic gradient estimation.

We will first introduce topic models and the two main training algorithms, Gibbs sampling and variational Bayes. Following this, we briefly introduce nonparametric topic models and present an efficient sampling method. We then show how this sampling method can be used to construct a hybrid Variational-Gibbs method. To further speed up our algorithm, we propose a more efficient sampling algorithm. Our experiments show that our method converges better than the purely variational topic modeling method as well as the Gibbs sampler. Supplementary material for this paper is available at `https://www.datamining.informatik.uni-mainz.de/files/2015/06/supplement.pdf`.

2 Background

In this section, we will first provide the necessary background in latent Dirichlet allocation (LDA). Second, we will introduce the nonparametric version, hierarchical Dirichlet processes (HDP).

2.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative model of document collections where each document is modeled as a mixture of latent topics. Each topic $k \in 1, \ldots, K$ is represented by a multinomial distribution ϕ_k over words that is assumed to be drawn from a Dirichlet distribution $Dir(\beta)$. The dth document is generated by drawing a distribution over topics from a Dirichlet $\theta_d \sim Dir(\alpha)$, and for the nth word token in the document, first drawing a topic indicator $z_{dn} \sim \theta_d$ and finally drawing a word $w_{dn} \sim \phi_{z_{dn}}$.

To learn a model over an observed document collection W, we need to estimate the posterior distribution over the latent variables z, θ, and ϕ.

$$p(\phi, \theta, z|W, \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|\phi_{z_{dn}}) \quad (1)$$

The most popular training algorithms are variational Bayesian inference and Gibbs sampling and will briefly be introduced in the following sections.

2.2 Variational Bayesian Inference

In variational Bayesian inference a variational distribution is introduced to approximate the posterior by minimizing the KL divergence between the variational distribution and the true posterior. Usually, a fully factorized variational distribution is chosen:

$$q(\phi, \theta, z|\tilde{\beta}, \tilde{\alpha}, \tilde{\theta}) = \prod_{k}^{K} q(\phi_k|\tilde{\beta}_k) \prod_{d}^{D} q(\theta_d|\tilde{\alpha}_d) \prod_{n}^{N_d} q(z_{dn}|\tilde{\theta}_{dn}) \tag{2}$$

The evidence lower bound (ELBO) that is to be maximized is given as follows:

$$\log p(W) \geq \mathcal{L}(\tilde{\beta}, \tilde{\alpha}, \tilde{\theta}) \triangleq \mathbb{E}_q[\log p(\phi, \theta, z, W)] + \mathcal{H}(q(\phi, \theta, z)) \ , \tag{3}$$

where $\mathcal{H}$ denotes the entropy.

By calculating the gradient of the ELBO with respect to the variational parameters, the parameters can be updated until convergence. The local/document-level update equations for collapsed variational Bayes (CVB0 [1, 6]), holding global variational parameter $\tilde{\beta}$ fixed, are:

$$\tilde{\alpha}_{dk} = \alpha + \sum_{n=1}^{N_d} \tilde{\theta}_{dnk} \tag{4}$$

$$\tilde{\theta}_{dnk} \propto \frac{\tilde{\beta}_{wk} + \beta}{\sum_v (\tilde{\beta}_{vk} + \beta_v)} (\tilde{\alpha}_{dk} + \alpha) \ , \tag{5}$$

where α and β are hyperparameters and N_d is the number of words in document d. Based on the local variational parameters $\tilde{\theta}$, the global parameter $\tilde{\beta}$ can be updated as follows:

$$\tilde{\beta}_{vk} = \beta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \tilde{\theta}_{dnk} \mathbb{1}[w_{dn} = v] \ , \tag{6}$$

where D is the number of documents. $\mathbb{1}[w_{dn} = v]$ is one if word $w_{dn} = v$ and zero otherwise.

2.3 Gibbs Sampling

Gibbs sampling does not have to resort to a factorized variational distribution, which is why the method is unbiased. Through integrating out the latent variables ϕ and θ, the model can be efficiently trained. Convergence is slower for Gibbs sampling since updates only involve a sampled topic instead of the full distribution over topics as in variational Bayes.

The conditional probabilities for training an LDA topic model are [7]:

$$p(z_i = k|z_{-i}, d, w) \propto \frac{n_{wk} + \beta}{\sum_v n_{vk} + \beta_v} (n_{dk} + \alpha) \ , \tag{7}$$

where n_{wk} and n_{dk} are the respective counts of topics k with words w or in documents d and α and β are hyperparameters as before. z_{-i} are all topic indicators except the one for token i.

2.4 Hierarchical Dirichlet Processes

For hierarchical Dirichlet process (HDP) topic models [16], the multinomial distribution θ from LDA is drawn from an HDP instead of a Dirichlet distribution: $\theta \sim DP(G_0, b_1), G_0 \sim DP(H, b_0)$. The base distribution G_0 of the first Dirichlet process (DP) is again drawn from a DP with base distribution H. This is why it is called a *hierarchical* DP.

A DP is a prior for a multinomial with a potentially unbounded number of topics. Drawing different multinomials from a DP results in multinomials of different sizes. Because the prior is hierarchical, there is a local topic distribution θ for each document and a global topic distribution G_0 which is shared among all documents. The advantage of this global topic distribution is that it allows topics of widely varying frequencies whereas in standard LDA with a symmetric prior α, all topics are expected to have the same frequency. The asymmetric prior of HDP usually leads to a better representation and higher log-likelihood of the dataset [16].

2.5 Sampling for HDP

Sampling methods for HDPs are mostly based on the Chinese restaurant process metaphor. Each word token is assumed to be a customer entering a restaurant, and sitting down at a certain table where a specific dish is served. Each table is associated with one dish which corresponds to a topic in a topic model. The probability for a customer to sit down at a certain table is proportional to the number of customers already sitting at that table. With a certain probability α, the customer sits down at a new table. In this case a topic is sampled from the base distribution. For an HDP topic model, each document corresponds to a restaurant. The topics in each document-restaurant are drawn from a global restaurant. Because all documents share the same global restaurant, the topics are shared. If a new table is added to a document restaurant, a pseudo customer enters the global restaurant. If a new table is opened in the global restaurant, a new topic is added to the topic model.

In terms of the statistics that need to be kept, in the basic version we need to store for each word not only the sampled topic, but also the table it is associated with. Also, we need to store the corresponding topic for each table.

Three basic sampling methods were introduced in Teh *et al.* [16], two methods are directly based on the Chinese restaurant representation, the third is the direct assignment sampler. The first two methods sample a table for each word and a topic for each table. This can be very slow and requires to store separate counts for each table. The direct assignment sampler does not sample an individual table but assigns a topic to each word token directly, and instead of keeping the statistics for each table separately, it simply samples the number of tables that are associated with a certain topic. While this sampler has improved convergence over the other sampling methods, it needs to sample $s_k \in \{1, \ldots, N_k\}$, the number of tables for topic k which can be very inefficient when the number of customers per topic N_k is very large. A further improved version was therefore

introduced by Chen *et al.* [5]. In this version another auxiliary variable u is introduced which is sampled for each customer and determines whether or not the customer sits down at a new table or an existing one. This is then used to update the table count s_k. u itself does not have to be kept in memory but can be sampled when needed. This way, the memory requirements are similar to Teh's auxiliary variable sampler, the sampling process itself is more efficient, and convergence is improved. For this reason we base our sampler on the sampling method by Chen *et al.*. We will only give the sampling equations here and refer to the original publication for more details.

$P(z_i = k|rest)$, the conditional probability of assigning topic k to token i in document j, is given as follows:

If the topic is new for the root restaurant (table indicator is zero):

$$P(z_i = k_{new}, u_i = 0|rest) \propto \frac{b_0 b_1}{(b_0 + M_\cdot)(b_1 + N_j)} \frac{N_{wk} + \beta}{\sum_{w'} (N_{w'k} + \beta)} \tag{8}$$

If the topic is new for the base restaurant (e.g. a document), but not for the root restaurant (table indicator is one):

$$P(z_i = k, u_i = 1|rest) \propto \frac{b_1 * M_k^2}{(M_k + 1)(b_0 + M_\cdot)(b_1 + N_j)} \frac{N_{wk} + \beta}{\sum_{w'} (N_{w'k} + \beta)} \tag{9}$$

If the topic exists at the base restaurant and a new table is opened (table indicator is one):

$$P(z_i = k, u_i = 1|rest) \propto \frac{b_1}{b_1 + N_j} \frac{S_{m_{jk}+1}^{n_{jk}+1}}{S_{m_{jk}}^{n_{jk}}} \frac{m_{jk} + 1}{n_{jk} + 1} \frac{M_k^2}{(b_0 + M_\cdot)(M_k + 1)} \frac{N_{wk} + \beta}{\sum_{w'} (N_{w'k} + \beta)} \tag{10}$$

If the topic exists at the base restaurant and an old table is chosen (table indicator is two):

$$P(z_i = k, u_i = 2|rest) \propto \frac{S_{m_{jk}}^{n_{jk}+1}}{S_{m_{jk}}^{n_{jk}}} \frac{n_{jk} - m_{jk} + 1}{(n_{jk} + 1)(b_1 + N_j)} \frac{N_{wk} + \beta}{\sum_{w'} (N_{w'k} + \beta)} \tag{11}$$

In the above equations, b_0 and b_1 are hyperparameters, M_k is the overall number of tables for topic k, N_{wk} is the overall number of customers for topic k and word w, N_j is the overall number of customers for restaurant j, n_{jk} is the number of customers in restaurant j for topic k, and m_{jk} is the number of tables in restaurant j for topic k. S_m^n are generalized Stirling numbers that can be efficiently precomputed and retrieved in $O(1)$ [2].

3 Proposed Method – Hybrid Variational-Gibbs

We will now describe how the sampling method above can be used to construct a hybrid Variational-Gibbs training algorithm for the HDP. Our algorithm is

online since it is based on stochastic gradient ascent [10, 9, 6]. This means our model can be continuously updated with new batches of data.

Taking up Sect. 2.2 on variational inference and following Hoffman *et al.* [9], the natural gradient of the ELBO with respect to $\tilde{\beta}$ is defined as

$$\mathbb{E}_q[N_{dkw}] + \frac{1}{D}(\beta - \tilde{\beta}_{kw}) \tag{12}$$

To evaluate the expectation in this equation we would need to evaluate all possible topic configurations for each document. For using stochastic gradient ascent however, an approximation is sufficient. This is where Gibbs sampling comes in. By taking samples from the distribution $q^\star$ we can approximate the expectation in the above equation.

$$q^\star(z_{di} = k|z_{-i}) \propto \exp\{\mathbb{E}_{q(\neg z_d)} \log(p(z_d|b_1, G_0)p(w_d|z_d, \phi))\} \ , \tag{13}$$

where $\neg z_d$ denotes all topic indicators z except the ones for document d. This distribution is difficult to normalize since we would have to consider all possible topic configurations z_d. However, we can easily sample from it and estimate the variational Dirichlet parameters as follows [9]:

$$\tilde{\beta}_{kv} = \beta + \sum_d \sum_i \mathbb{E}_q[\mathbb{1}[z_{di} = k]\mathbb{1}[w_{di} = v]] \ , \tag{14}$$

where the expectation is approximated by the samples from $q^\star$.

In contrast to Hoffman *et al.*, we have an additional variational distribution over the topics G_0. This is the global topic prior. The global variational distribution for G_0 and the mixture components ϕ is

$$q(G_0, \phi|\tilde{\gamma}, \tilde{\beta}) = \prod_k q(G_{0_k}|\tilde{\gamma})q(\phi_k|\tilde{\beta}_k) \ , \tag{15}$$

where $\tilde{\gamma}$ and $\tilde{\beta}$ are Dirichlet parameters.

The variational Dirichlet parameter for the global topic distribution is analogously estimated as follows:

$$\tilde{\gamma}_k = \gamma + \sum_d \sum_i \mathbb{E}_q[\mathbb{1}[z_{di} = k]\mathbb{1}[u_{di} = 1|u_{di} = 0]] \ , \tag{16}$$

where γ is a hyperparameter, and $\mathbb{1}[u = 1|u = 0]$ is one if the table indicator u is either zero or one, which means that a new table is being opened, and otherwise zero.

The expectations in Equations 14 and 16 can be estimated by sampling from $q^\star$ which is given by the following set of equations (compare to Equations 8–11, differences are highlighted in bold):

If the topic is new for the root restaurant (table indicator is zero):

$$q^\star(z_{di} = k, u = 0|z_{-i}) \propto \frac{b_0 b_1}{(b_0 + \boldsymbol{\sum_{k'} \tilde{\gamma}_{k'}})(b_1 + N_j)} \boldsymbol{\exp\left(\mathbb{E}[\log_{wk}]\right)} \tag{17}$$

If the topic is new for the base restaurant (e.g. a document), but not for the root restaurant (table indicator is one):

$$q^{\star}(z_{di} = k, u = 1 | z_{-i}) \propto \frac{b_1 * \tilde{\boldsymbol{\gamma}}_{\boldsymbol{k}}^2}{(\tilde{\boldsymbol{\gamma}}_{\boldsymbol{k}} + 1)\,(\sum_{\boldsymbol{k'}} \tilde{\boldsymbol{\gamma}}_{\boldsymbol{k'}} + b_0)\,(b_1 + N_j)} \exp\left(\mathbb{E}[\log_{\boldsymbol{wk}}]\right) \tag{18}$$

If the topic exists at the base restaurant and a new table is opened ($u = 1$):

$$\begin{aligned} q^{\star}(z_{di} = k, u = 1 | z_{-i}) \propto \frac{b_1}{b_1 + N_j} \frac{S_{m_{jk}+1}^{n_{jk}+1}}{S_{m_{jk}}^{n_{jk}}} \frac{m_{jk} + 1}{n_{jk} + 1} \\ \frac{\tilde{\boldsymbol{\gamma}}_{\boldsymbol{k}}^2}{(\sum_{\boldsymbol{k'}} \tilde{\boldsymbol{\gamma}}_{\boldsymbol{k'}} + b_0)(\tilde{\boldsymbol{\gamma}}_{\boldsymbol{k}} + 1)} \exp\left(\mathbb{E}[\log_{\boldsymbol{wk}}]\right) \end{aligned} \tag{19}$$

If the topic exists at the base restaurant and an old table is chosen ($u = 2$):

$$q^{\star}(z_{di} = k, u = 2 | z_{-i}) \propto \frac{S_{m_{jk}}^{n_{jk}+1}}{S_{m_{jk}}^{n_{jk}}} \frac{n_{jk} - m_{jk} + 1}{n_{jk} + 1} \exp\left(\mathbb{E}[\log_{\boldsymbol{wk}}]\right) \tag{20}$$

In the above equations, the number of tables M_k (Equations 8 and 9) is substituted by the global variational parameter $\tilde{\gamma}$. $\exp(\mathbb{E}[\log \phi_{wk}])$ is expensive to compute, since $log(\phi_{wk}) = \psi(\tilde{\beta}_{wk}) - \psi(\sum_w \tilde{\beta}_{wk})$, where $\psi(\cdot)$ is the digamma function. Following Wang *et al.* [18] and Li *et al.* [14] we use $\frac{\tilde{\beta}_{wk}+\beta}{\sum_{w'}(\tilde{\beta}_{w'k}+\beta)}$ instead. The remaining variables are the local counts equivalent to the counts in Equations 8 to 11.

Updating variational parameters $\tilde{\beta}$ and $\tilde{\gamma}$ for one minibatch M is done as follows, where the counts for one minibatch are scaled by $\frac{|D|}{B|M|}$ for B burn-in iterations, to arrive at the expectation for the whole corpus and ρ is a parameter between zero and one.

$$\tilde{\beta}_{kw} = (1 - \rho_t)\tilde{\beta}_{kw} + \rho_t \left(\beta + \frac{|D|}{B|M|} \sum_{d \in M} N_{dkw}\right) \tag{21}$$

$$\tilde{\gamma}_k = (1 - \rho_t)\tilde{\gamma}_k + \rho_t \left(\gamma + \frac{|D|}{B|M|} \sum_{d \in M} \sum_{n \in d} \mathbb{1}[u_{dn} = 1 | u_{dn} = 0]\right) \tag{22}$$

Summing up this section, we make use of the table indicators from Chen *et al.*'s sampling method for the HDP to be able to approximate the global topic distribution from minibatch samples. This yields an online algorithm for the HDP topic model.

3.1 Doubly Sparse Sampling for HDP

Having introduced our hybrid variational algorithm based on the table indicator sampling scheme, we will now introduce a doubly sparse sampling method for the nonparametric topic model [4]. This is similar to Li *et al.*'s [14] method for

Algorithm 1: Train Topic Model

Input: Dataset D

repeat
  $M \leftarrow$ get minibatch from D
  compute q^e (Equations 23,17,18) and $Q = \sum q^e$
  $A_w \leftarrow$ computeAliasTable($\frac{q^e_w}{Q_w}$) for each word w (see supplement)
  for *document* $d \in M$ **do**
    $z_d \leftarrow$ initialize randomly
    for *iteration* $i = 1, \ldots, S + B$ **do**
      for *token* $n = 1, \ldots, N_d$ **do**
        $z_{dn}, u_{dn} \leftarrow \text{Sample}(A, w_{dn})$ (Algorithm 2)
      if $i > B$ **then**
        Save sample
  update $\tilde{\beta}$ and $\tilde{\gamma}$ (Equations 21 and 22)
until *convergence*

Algorithm 2: Sample(A, w)

compute $\tilde{p}$,$\tilde{P} = \sum \tilde{p}$,$\Delta$ (Equations 24,25), $i = -1$, $u \leftarrow 1$
sample $r \sim \text{Uniform}(0, \tilde{P} + \tilde{Q})$
if $r < \tilde{P}$ **then**
  while $r > 0$ **do**
    $i \leftarrow i + 1$, $t \leftarrow i/2$, $u \leftarrow 2 - (i \mod 2)$, $r \leftarrow r - \tilde{p}_{j,w}(t, u)$
else
  repeat
    $t \leftarrow$ sample from Alias A_w
  until *t is new in document*
return u, t

the parametric topic model and would not be possible for the direct assignment sampler. Hereby, we take advantage of the fact that the number of topics that occur in one document K_d is usually much lower than the total number of topics K. Furthermore, we improve this sampling method to make it more memory efficient.

Making Table Indicator Sampling Sparse. To obtain sparsity, the topic distribution can be divided into three parts according to the table indicators [4]:

$$q^\star(z = k|z_{-i}) = q^\star(z = k, u = 0|z_{-i}) + q^\star(z = k, u = 1|z_{-i}) + q^\star(z = k, u = 2|z_{-i}).$$

The last part is sparse since it is only nonzero for the topics that occur with the document. Therefore, we can use alias-sampling to save the distribution for

the dense part (The algorithm is provided in the supplementary material.) and subsequently draw samples from it in $O(1)$, whereas the sparse part can be computed in $O(K_j)$ since it is only necessary to iterate over those topics that occur within the jth document.

Formally, we rewrite the topic distribution $q^\star(k)$ over topics k as a combination of a dense distribution q_{jw} and a sparse distribution p_{jw}, where w is a word, and j is a document-restaurant. The normalization terms are given by $P_{jw} = \sum_k p_{jw}(k)$ and $Q_{jw} = \sum_k q_{jw}(k)$. The resulting distribution is given by:

$$q^\star(k) := \frac{p_{jw}(k) + q_{jw}(k)}{P_{jw} + Q_{jw}}$$

We define the stale distribution q_{jw} as the distribution over all topics and a table indicator of 0 or 1:

$$q_{jw}(k) := q^\star(z = k, u = 0|rest) + q^\star(z = k, u = 1|rest) \tag{23}$$

The fresh distribution p_{jw} is defined as the distribution over all topics that exist in restaurant j and a table indicator of 2:

$$p_{jw}(k) := q^\star(z = k, u = 2|rest)$$

Improving the sparse sampler. As can be inferred from the above equations and Equations 18 and 19, q_{jw} depends on document j. If topic k exists in document j, the probability is given by Equation 19, otherwise by Equation 18. This means we would have to save topic distributions for every single document. It would be more appealing to have one topic distribution that represents the global topic distribution and can be used for all documents. The solution we propose to improve the method described above is as follows:

We simply assume for each topic that it does not exist in the document and save the resulting distribution q_w^e for an empty pseudo document e. This is equivalent to replacing Equation 19 with Equation 18. Now we only need to add a subsequent rejection step for the case where we sample a topic from this distribution that exists in the current document. If this happens, we simply discard it and draw a new sample. Since each sample is drawn in $O(1)$ this does not significantly increase the complexity as long as the basic assumption holds that $K_d \ll K$. The sparse distribution is now over table indicators one and two instead of just two, to account for the case where a new table is opened for an existing topic.

$$\tilde{p}_{jw}(k, u') := q^\star(z = k, u = u'|rest)\mathbb{1}[n_{jk} > 0] \ , \tag{24}$$

where $\mathbb{1}[n_{jk} > 0]$ is one if the number of tokens in document-restaurant j associated with topic k is at least one and zero otherwise. Accordingly, the normalization sum is $\tilde{P}_{jw} = \sum_k \sum_u \tilde{p}_{jw}(k, u)$.

We need to subtract an amount Δ_j from the normalization sum Q_w which is different for each document j and accounts for the topics that are present in

document j and would be rejected if drawn from distribution q. We call it the discard mass Δ and it is defined as follows:

$$\Delta_j := \sum q_k^e \mathbb{1}[n_{jk} > 0] \tag{25}$$

Since Δ_j can be computed in $O(K_j)$ time, it does not add to the overall computational complexity. Following this, the normalization sum is now given by $\tilde{Q}_{jw} = Q_w - \Delta_j$, where $Q_w = \sum q_w^e$.

Note that since the distribution does not change during the gradient estimation for one minibatch, i.e. we sample from the exact distribution, we do not need to add a Metropolis-Hastings acceptance step as in Li *et al.* [13].

The whole algorithm is summed up in Algorithm 1. For each minibatch, the dense distributions q^e are computed for each word that occurs in the minibatch. Alias tables are computed for these distributions (see the supplementary material) to be able to sample from them in $O(1)$. For each document d in the minibatch, the topics z_d are sampled using Algorithm 2 and stored after a burn-in period of B iterations. The stored samples are finally used to update the global variational distributions.

4 Experiments

4.1 Algorithms

We compare three different topic models:

1. Our method OSCHVB-HDP (Online Sparse Collapsed Hybrid Variational Bayes): The implementation is done in Java.
2. Wang *et al.*'s stochastic mean-field variational HDP (SMF-HDP) [19]: The python implementation provided by the author was used. Unfortunately this does not allow doing a fair runtime comparison. We generally observed our method to be faster, however, this could be due to the different programming languages.
3. The Gibbs sampling algorithm by Chen *et al.* [5]: Our own Java implementation was used to make the runtime comparable to our hybrid algorithm; the code only differs in the implementation of the sampling step.

4.2 Parameter Settings

The Dirichlet parameter for the topic-word distributions β is always set to 0.01, a standard default value.

The number of sampling iterations S and the number of burn-in iterations B are parameters in the hybrid algorithm. In accordance with the existing literature, we chose $S = B = 5$. The same parameters were used during evaluation.

We used the same update parameter $\rho_t = \frac{1}{(1+n)^{0.6}}$ for all algorithms, where n is the number of batches that have been processed up to time t.

For the mean-field variational algorithm by Wang *et al.* [19] we used the public python implementation provided by the authors. As parameter settings we chose the default settings that were also used by Wang and Blei [18], the second level truncation was set to 20. A batch size of 100 was chosen as our default batch size. For our nonparametric method we used the same hyperparameters $b_1 = b_0 = 1.0$.

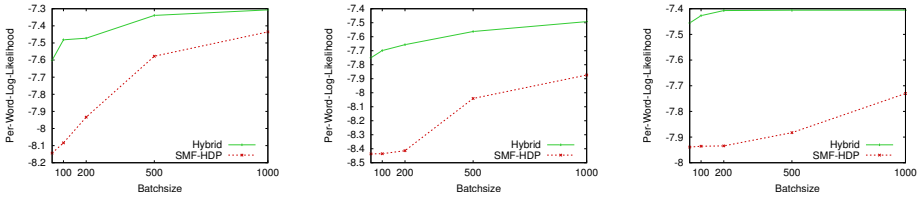
4.3 Datasets

Table 1: Statistics for the datasets used in our experiments, $|D|$ train and test: the number of documents in the train- and testset, respectively, $|V|$: size of vocabulary

Dataset	$\|D\|$ train	$\|D\|$ test	$\|V\|$
Enron	507,401	10,000	20,000
BioASQ	500,000	10,000	20,000
NIPS	4,811	1,000	11,463
KOS	2930	500	6906

We used four publicly available datasets and preprocessed them as follows:

1. `BioASQ`:
   This dataset consists of paper abstracts from the PubMed database. It was made available for the BioASQ competition, a large-scale semantic indexing challenge [17]. We separated the 500,000 most recent documents plus 10,000 documents as a separate test dataset. After stopword removal we kept the 20,000 most frequent features.
2. `Enron`:
   The `Enron` dataset consists of ca. 500,000 emails and is available at `https://www.cs.cmu.edu/~./enron/`[12]. We removed the header, tokenized the emails, removed stopwords and kept the 20,000 most frequent features. We randomly separated 10,000 documents as a testing dataset.
3. `NIPS`:
   This dataset is available in a preprocessed format from the UCI Machine Learning Repository [15]. It has 5,812 documents and 11,463 features and is the second smallest dataset that we used. It consists of NIPS conference papers published between 1987 and 2015. In comparison to the other datasets, the individual documents are large. 1000 documents were separated as a testset.
4. `KOS`:
   The 3430 blog entries of this dataset were originally extracted from `http://www.dailykos.com/`, the dataset is available in the UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/datasets/Bag+of+Words`. The number of features is 6906.



(a) Enron (b) BioASQ (c) NIPS

Fig. 1: Effect of batch size on the log-likelihood after 1000 updates. The truncation was set to 100 topics.

4.4 Evaluation

We evaluated the models on the per-word-log-likelihood according to Heinrich [8]:

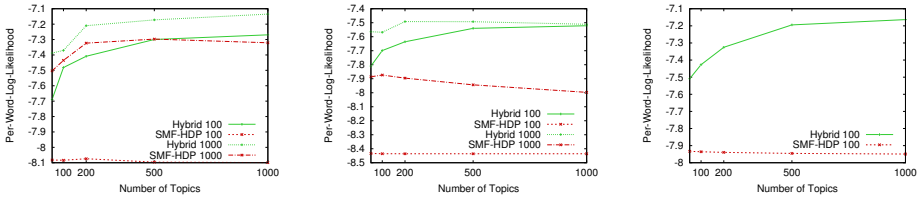
$$\frac{\log p(w|M)}{|n_d|} = \left(\sum_{w \in d} n_{dw} \log \left(\sum_{k=1}^{K} \phi_{kw} \theta_{kw} \right) \right) / |n_d| \ , \tag{26}$$

where n_{dw} is the number of times word w occurs in document d, K is the overall number of topics, and ϕ are the model parameters. θ are the document specific parameters that need to be estimated using the model. In our case, we run the sampler with a fixed point estimate of parameters ϕ and estimate θ analogously to the training procedure with S samples that are saved after B burn-in iterations.

4.5 Experimental Results

Mean-field vs. Hybrid approach. First of all we compare our hybrid approach to Wang *et al.*'s SMF-HDP [19] on the three largest datasets. We notice that the performance of SMF-HDP heavily depends on the batch size. Small batch sizes lead to a much worse performance (see Fig. 1). The same observation was made by Wang and Blei [18]. SMF-HDP starts with the maximum topic number and then reduces the number of topics. In our experiments, often only a handful of topics remained for small batch sizes. Wang and Blei hypothesized that this is due to the algorithm being strongly dependent on the initialization and not being able to add topics occurring in later batches that had not been present from the start. This behavior is problematic, especially in the online setting where it is not guaranteed that the first batch contains all the topics. Our method is more robust and better suited to settings where small batch sizes are a requirement.

Second, we compare the two algorithm for different settings of the truncation for the number of topics (50, 100, 200, 500, 1000). The results are shown in Fig. 2. While the truncation does not seem to influence the performance of SMF-HDP for small batch sizes, our method has an improved performance for higher



(a) `Enron` (b) `BioASQ` (c) `NIPS`

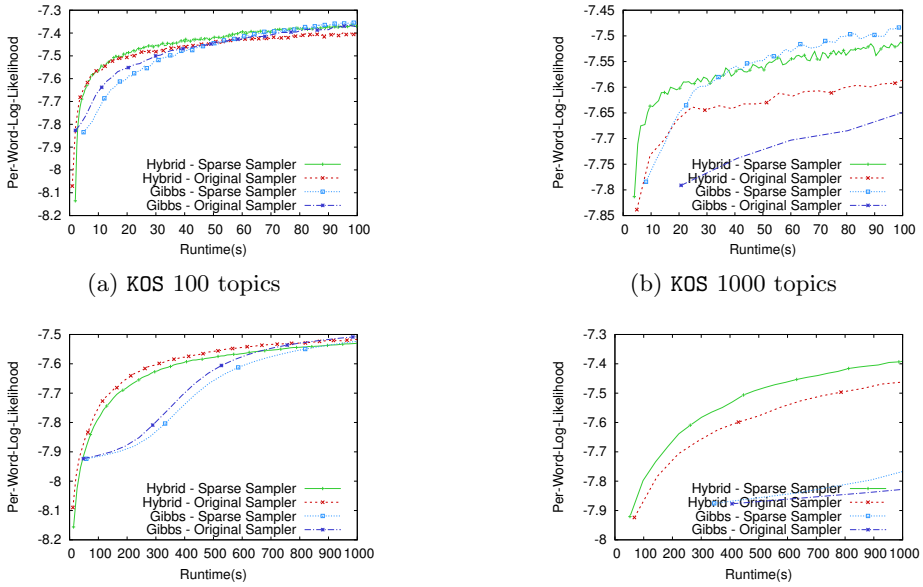
Fig. 2: Effect of truncation on the log-likelihood after 1000 updates. The batch size was set to 100 and 1000 documents as mentioned in the plot labels. For `Enron` we only used a batch size of 100 since larger batch sizes lead to a substantial increase in runtime. Our method has a higher log-likelihood for all settings.

topic numbers on all three datasets. We see therefore, that it is not necessary to start with one topic and add more topics subsequently as was suggested by Wang and Blei [18]. Agreeing with the observations in previous work [3], we find that it is beneficial to start out with the maximum number of topics and subsequently reduce it. Overall, our method has a higher log-likelihood for all truncation settings.

Gibbs Sampling vs. Hybrid Approach. Since training and evaluation of the Gibbs sampler take too long on the large datasets, we also included the smaller `KOS` dataset in our experiments. We compare the convergence of the Gibbs sampler to the convergence of our method by measuring the testset per-word-log-likelihood after each iteration over the full dataset for the Gibbs sampler, and evaluating our method after each batch. Fig. 3a shows the performance of four different methods trained with a truncation to 100 topics. We can see that the two hybrid methods converge much faster initially than the Gibbs samplers. Comparing the sparse and the original sampler, the sparse sampler is worse in the beginning since it does not sample from the true distribution, but manages to catch up and even surpass the original sampler due to its faster sampling[1].

Fig. 3b shows the performance for a truncation to 1000 topics. We can see that here the difference in log-likelihood between the sparse and the original sampling method is much bigger. This is because the number of topics per document K_d does not grow that much when the number of total topics is increased. Therefore, while for small topic numbers the differences might be negligible in practice, for higher topic numbers, our sparse sampling method is preferable. This performance improvement in the experiments shows that sparseness is actually achieved and the assumption $K_d \ll K$ is justified in practice.

[1] Note that our implementation of the hybrid method uses the sparse sampling method, but does not use the sparse updating introduced by Hoffman *et al.* [9]. Therefore, a further speedup is possible.



(a) KOS 100 topics

(b) KOS 1000 topics

(c) NIPS 100 topics

(d) NIPS 1000 topics

Fig. 3: Comparison of runtime with 100 and 1000 topics, respectively. The hybrid methods were trained with a batch size of 100 documents. Performance was evaluated on a separate testset. Our hybrid method with the sparse sampling algorithm converges faster than the other methods.

For the NIPS dataset, the difference between 100 and 1000 topics is even bigger (Fig. 3c and 3d). NIPS has very long documents which means that it has more topics per document on average. With only 100 topics, it is possible that almost all topics are present in the document. Therefore, the original sampler is faster than the sparse sampler for 100 topics, but not for 1000 topics, where it is the other way around. The Gibbs samplers have barely even started to converge in the first 1000 seconds where the convergence for the hybrid methods is far ahead.

5 Related Work

A hybrid Variational-Gibbs method for parametric LDA topic models was introduced by Hoffman *et al.*[9]. In this work, a sparse update scheme was proposed that allowed to do variational updates for only the topic-word-combinations that were actually sampled. Experiments by Hoffman *et al.* showed that the method is faster especially for large topic numbers. Additionally the convergence is improved as compared to other variational methods since the variational distribu-

tion considered is not completely factorized but considers each document as a unity.

A doubly sparse method was build on top of the sparse hybrid model by Li *et al.* [14]. They used a fast Gibbs sampling method to further speed up sampling for parametric LDA. Thereby they exploited sparseness in the variational updates as well as the document-topic distributions. We do the same, only for nonparametric topic models.

Another extension of the original work by Hoffman *et al.* was proposed by Wang and Blei [18] who developed a similar method for the nonparametric HDP. Unfortunately, we were not able to compare to this method since the code is not publicly available. The main contribution of this work is the development of a truncation-free variational method that allows the number of topics to grow. This is made possible by the sampling step which does not depend on a truncation as in pure variational methods. In contrast to our work, their method builds on Teh's direct assignment sampler [16], whereas our method relies on the more advanced table indicator sampler proposed by Chen *et al.* [5]. Also, our method starts out with the maximum number of topics, and subsequently removes topics (by letting their expected counts approach zero over time). This was found to be beneficial in previous work [3]. Finally, their method is not sparse which is due to the direct assignment sampling scheme which cannot be made sparse as easily as the table indicator sampling scheme.

6 Conclusion

To conclude, we introduced a hybrid sparse Variational-Gibbs nonparametric topic model that can be trained online on large-scale or streaming datasets. Experiments on three large-scale test datasets as well as one smaller dataset were conducted. We found our method to be superior to the purely variational Bayes mean field approach in per-word log-likelihood. Additionally, it is more robust to different settings of the batch size. Compared to the pure Gibbs sampler it converges faster with improved log-likelihood. In the future, we would like to apply our method to hierarchical topic models with more levels.

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. pp. 27–34. UAI '09, AUAI Press, Arlington, Virginia, United States (2009)
2. Buntine, W., Hutter, M.: A Bayesian view of the Poisson-Dirichlet process. arXiv preprint arXiv:1007.0296 (2010)
3. Buntine, W.L., Mishra, S.: Experiments with non-parametric topic models. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 881–890. KDD '14, ACM, New York, NY, USA (2014)

4. Burkhardt, S., Kramer, S.: Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. In: ICBK 2017 - International Conference on Big Knowledge. Hefei, China (2017), to appear
5. Chen, C., Du, L., Buntine, W.: Sampling table configurations for the hierarchical Poisson-Dirichlet process. In: Proc. of ECML-PKDD. pp. 296–311 (2011)
6. Foulds, J., Boyles, L., DuBois, C., Smyth, P., Welling, M.: Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 446–454. KDD '13, ACM, New York, NY, USA (2013)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101, pp. 5228–5235. National Acad. Sciences (2004)
8. Heinrich, G.: Parameter estimation for text analysis. Technical report, Fraunhofer IGD (2004)
9. Hoffman, M., Blei, D.M., Mimno, D.M.: Sparse stochastic inference for latent dirichlet allocation. In: Proceedings of the 29th International Conference on Machine Learning (ICML-12). pp. 1599–1606 (2012)
10. Hoffman, M.D., Blei, D.M., Bach, F.R.: Online learning for latent dirichlet allocation. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23. pp. 856–864. Curran Associates, Inc. (2010)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR abs/1312.6114 (2013)
12. Klimt, B., Yang, Y.: Introducing the enron corpus. In: CEAS 2004 - First Conference on Email and Anti-Spam, July 30-31, 2004, Mountain View, California, USA (2004)
13. Li, A.Q., Ahmed, A., Ravi, S., Smola, A.J.: Reducing the sampling complexity of topic models. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 891–900. KDD '14, ACM, New York, NY, USA (2014)
14. Li, X., OuYang, J., Zhou, X.: Sparse hybrid variational-gibbs algorithm for latent dirichlet allocation. In: Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016. pp. 729–737 (2016)
15. Perrone, V., Jenkins, P.A., Spano, D., Teh, Y.W.: Poisson random fields for dynamic feature models (2016), arXiv e-prints: 1611.07460
16. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101 (2004)
17. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16, 138 (2015)
18. Wang, C., Blei, D.M.: Truncation-free online variational inference for bayesian nonparametric models. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 413–421. Curran Associates, Inc. (2012)
19. Wang, C., Paisley, J.W., Blei, D.M.: Online variational inference for the hierarchical dirichlet process. In: AISTATS. vol. 2, p. 4 (2011)