Bayesian Machine Learning project: Collapsed Variational Bayesian Inference for LDA

Tristan Dot, Jean-Rémy Conti
Ecole Normale Supérieure Paris-Saclay - MVA

Abstract

An analysis of a collapsed variational bayesian inference method for Latent Dirichlet Allocation networks, as presented in [1]. This analysis presents the context of work, points out the main ideas that led to this method, details its mathematical theory and computational properties, and finally tests it on some new data, as a 'toy' example.

Notations

The following variable notations will be applied in this work:

- K: number of latent topics considered
- D: number of documents considered
- W: vocabulary size
- α: parameter of the Dirichlet prior on the per-document topic distributions
- β: parameter of the Dirichlet prior on the per-topic word distribution
- $\theta_j = \{\theta_{jk}\}$: topic distribution for document j, over K topics
- $\phi_k = \{\phi_{kw}\}$: word distribution for topic k, over W words
- z_{ij}: topic for the word i of document j
- x_{ij}: word i of document j
- $\boldsymbol{\theta}$: $(D \times K)$ matrix composed of rows $\theta_1, \ldots, \theta_D$, which are distributions over topics
- $\boldsymbol{\phi}$: $(K \times W)$ matrix composed of rows $\phi_1, \ldots, \phi_K$, which are distributions over words
- $\boldsymbol{z}$: $(W \times D)$ matrix composed of z_{ij}
- $\boldsymbol{x}$: $(W \times D)$ matrix composed of x_{ij}

1 General Introduction

1.1 Latent Dirichlet Allocation

1.1.1 Presentation

Latent Dirichlet allocation (LDA) (and particularly what is called "smoothed" LDA) is a highly popular class of Bayesian networks with discrete random variables that allows sets of observations to be explained by unobserved latent variables. It was introduced in 2003 [2] as a graphical model for topic discovery, and it has found important applications in text modeling [3, 4], but also in computer vision classification tasks (by treating an image as a document, and small patches of the image as words) [5, 6].

Generally speaking, in this model, documents are seen as mixtures of latent topics, which distributions are assumed to have a sparse Dirichlet prior. From a technical point of view, each document is seen as in a bag of words model: words are therefore considered exchangeable, and only their number of occurrences matters.

Moreover, in Bayesian networks with discrete random variables, it is usual to use Dirichlet priors over the probabilisitic models parameters, in order to ease inference computations: Dirichlet distributions are indeed conjugate to the multinomial distributions over the discrete random variables. In the particular LDA case, the use of sparse Dirichlet priors simply reflects the supposed 'sparsity' of the topics / documents and words links. In fact, LDA can be seen as a generalization of probabilistic latent semantic analysis (PLSA) [7].

1.1.2 Model

The ideas behind LDA model are the following ones.

First, we define for each topic a multinomial distribution over all possible words: $z_{ij} \sim \text{Multinomial}(\theta_j)$. Then, x_{ij} will be generated by the distribution over words corresponding to the topic z_{ij}: $x_{ij} \sim \text{Multinomial}(\phi_{z_{ij}})$.

Finally, we give prior distributions for the parameters θ_j and ϕ_k. The multinomial distribution is a generalization of the binomial distribution, and its conjugate prior is a generalization of the beta distribution: the Dirichlet distribution. We can therefore model the data with the following generative model:

1. For $k \in [1, \dots, K]$, $\phi_k \sim \text{Dirichlet}(\beta)$
2. For $j \in [1, \dots, D]$, $\theta_j \sim \text{Dirichlet}(\alpha)$
3. For $(i, j) \in [1, \dots, W] \times [1, \dots, D]$
   - $z_{ij} \sim \text{Multinomial}(\theta_j)$
   - $x_{ij} \sim \text{Multinomial}(\phi_{z_{ij}})$

The graphical model representation for this is given in Fig. 1.

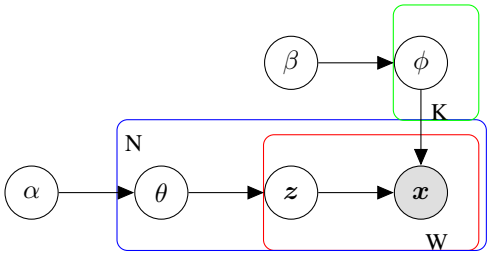

Figure 1: Plate notation for LDA

In this formulation, we have the following relations for the joint distribution of the topic mixtures proportions $\boldsymbol{\theta}$, the set of topic assignments $\boldsymbol{z}$, the words of the corpus $\boldsymbol{x}$, and the topics $\boldsymbol{\phi}$:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) &= p(\boldsymbol{z}, \boldsymbol{\theta} | \alpha) p(\boldsymbol{x}, \boldsymbol{\phi} | \boldsymbol{z}, \beta) \\ &= p(\boldsymbol{z} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \alpha) p(\boldsymbol{x} | \boldsymbol{z}, \boldsymbol{\phi}) p(\boldsymbol{\phi} | \beta) \end{aligned} \tag{1}$$

And we have:

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{\theta}) &= \prod_{j=1}^{D}\prod_{k=1}^{K}\boldsymbol{\theta}_{jk}^{n_{jk\cdot}} \\
p(\boldsymbol{\theta}|\alpha) &= \prod_{j=1}^{D}(\prod_{k=1}^{K}\boldsymbol{\theta}_{jk}^{\alpha-1})\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \\
p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\phi}) &= \prod_{k=1}^{K}\prod_{w=1}^{W}\boldsymbol{\phi}_{kw}^{n_{\cdot kw}} \\
p(\boldsymbol{\phi}|\beta) &= \prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{w=1}^{W}\phi_{kw}^{\beta-1}
\end{aligned}
\tag{2}
$$

And we finally get:

$$
p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\alpha,\beta) = \prod_{j=1}^{D}\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\boldsymbol{\theta}_{jk}^{\alpha-1+n_{jk\cdot}}\prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{w=1}^{W}\phi_{kw}^{\beta-1+n_{\cdot kw}} \tag{3}
$$

with $n_{jkw} = \#\{i|x_{ij} = w, z_{ij} = k\}$, and dot means n_{jkw} is summed on the corresponding index.

Bayesian inference aims at infering the posterior distribution for the latent variables (topic mixtures proportions $\boldsymbol{\theta}$, topic assignments $\boldsymbol{z}$, and topic parameters $\boldsymbol{\phi}$), given the observed words $\boldsymbol{x}$, i.e. at infering: $p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\alpha,\beta)$. Various methods exist for such inference, with their advantages and drawbacks.

1.2 Inference methods

Different inference methods have been proposed for LDA, such as variational Bayesian (VB) inference [2], expectation propagation (EM) [8], collapsed Gibbs sampling [3], or, more recently, stochastic variational inference (created in order to quickly process variational bayesian inference on huge datasets) [9].

While expectation propagation is not computationally efficient, and variational inference suffers from a very large bias, collapsed Gibbs sampling has largely been used, despite some drawbacks (particularly an important computational cost).

1.2.1 Main intuition

The main intuition behind collapsed variational inference presented in [1] was to mix the advantages of variational bayesian inference and of collapsed gibbs sampling. In more details, thanks to collapsed gibbs sampling, one fact appeared: sampling in a collapsed space, with marginalized parameters, works way better than sampling parameters and latent variables simultaneously, and therefore parameters and latent variables appear to be very coupled. Consequently, in joint space of parameters and latent space, fluctuations in parameters can have big impact on latent variables, and a mean field approximation seems very... approximative. On the contrary, when parameters are marginalized, even if some small dependencies are induced between latent variables, the situation is perfect for a mean field approximation. Consequently, one main intuition was used: working in a collapsed space of latent variables should favorize the mean field assumptions, and therefore largely favorize the accuracy of variational bayesian inference.

Before digging this very interesting intuition, let us quickly present variational Bayes inference techniques, which will be partially re-used in collapsed variational bayesian inference.

1.2.2 Variational Bayes

Although, as it can be seen through the cross terms in Eq. 3, latent variables $\boldsymbol{z}$ and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be largely dependent in the true posterior $p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x})$, this dependency is going to be ignored by variational bayes. This hypothesis will have important consequences.

Classicaly, the negative log marginal likelihood $-\log p(\boldsymbol{x}|\alpha,\beta)$ is upper bounded by variational free energy, as follows:

$$-\log p(\boldsymbol{x}|\alpha,\beta) \leq \tilde{\mathcal{F}}(\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})) = \mathbb{E}_{\tilde{q}}[-\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta)] - \mathcal{H}(\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})) \tag{4}$$

where $\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$ is an approximate posterior, and $\mathcal{H}(\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}))$ the variational entropy. $\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$ is assumed to be factorized:

$$\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}) = \prod_{ij} \tilde{q}(z_{ij}|\tilde{\gamma}_{ij}) \prod_{j} \tilde{q}(\theta_i|\tilde{\alpha}_j) \prod_{k} \tilde{q}(\phi_k|\tilde{\beta}_k) \tag{5}$$

with $\tilde{q}(z_{ij}|\tilde{\gamma}_{ij}) \sim$ Multinomial$(\tilde{\gamma}_{ij})$, $\tilde{q}(\theta_j|\tilde{\alpha}_j) \sim$ Dirichlet$(\tilde{\alpha}_j)$ and $\tilde{q}(\phi_k|\tilde{\beta}_k) \sim$ Dirichlet$(\tilde{\beta}_k)$.

After an optimization of $\tilde{\mathcal{F}}(\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}))$ with respect to variational parameters, we get a set of updates for $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{kw}$ and $\tilde{\gamma}_{ijk}$, which will converge to a local minimum. As given in [1]:

$$\begin{cases} \tilde{\alpha}_{jk} & = \alpha + \sum_i \tilde{\gamma}_{ijk} \\ \tilde{\beta}_{kw} & = \beta + \sum_{ij} \mathbb{1}(x_{ij} = w)\tilde{\gamma}_{ijk} \\ \tilde{\gamma}_{ijk} & \propto \exp \Psi(\tilde{\alpha}_{jk}) + \Psi(\tilde{\beta}_{kx_{ij}}) - \Psi(\sum_w \tilde{\beta}_{kw}) \end{cases} \tag{6}$$

with the digamma function $\Psi(y) = \frac{\delta \log \Gamma(y)}{\delta y}$.

It is an efficient algorithm (time complexity: $O(MK)$, with M number of unique document/word pairs), but with a very loose upper bound condition, because of the dependency condition between $\boldsymbol{z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Putting ourselves in a collapse environment will give way more accurate estimates of the posterior.

2 Collapsed Variational Bayesian Inference for LDA

This section introduces a modified version of Variational Bayesian (VB) inference from [1], for the special case of LDA. Instead of assuming full independence between the latent variables $\boldsymbol{z}$ and the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ conditionally to the observed data $\boldsymbol{x}$, it actually models perfectly the dependence of the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ on the latent variables $\boldsymbol{z}$. Nevertheless, the mutual independence between the latent variables $\boldsymbol{z}$ is still assumed (as in VB inference). Obviously, this way of proceeding allows for a search of an approximation for the true posterior $p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\alpha,\beta)$ in a much bigger space than the space of fully factorized distributions of the form $\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$ (see 1.2.2). Thus, if the approximation is tractable, it would certainly give a better approximation of the true posterior than VB inference.

2.1 A more accurate method than Variational Bayesian Inference

In details, as in VB inference, the goal is to approximate the true posterior $p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\alpha,\beta)$ by a distribution $\hat{q}$ which is decomposed - without loss of generality - as $\hat{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}) = \hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})\ \hat{q}(\boldsymbol{z})$. The mutual independence of the latent variables $\boldsymbol{z}$ gives:

$$\hat{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}) = \hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z}) \prod_{i,j} \hat{q}(z_{ij}|\hat{\gamma}_{ij}) \tag{7}$$

where the posterior $\hat{q}(z_{ij}|\hat{\gamma}_{ij})$ is assumed to be multinomial (with new parameters $\hat{\gamma}_{i,j}$) as in VB inference. This expression leads to a new version of variational free energy $\hat{\mathcal{F}}[\hat{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})]$ (see 1.2.2 for the original definition). Indeed, since for any suitable function $f(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$, we have that $\mathbb{E}_{\hat{q}(\boldsymbol{z})\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})}[f(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})] = \mathbb{E}_{\hat{q}(\boldsymbol{z})}[\mathbb{E}_{\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})}[f(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})]]$ (simply write the definitions and fix a value for $\boldsymbol{z}$ in $\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})$), we obtain the following expression by using some basic property of the logarithm function:

$$\hat{\mathcal{F}}[\hat{q}(\boldsymbol{z})\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})] = \mathbb{E}_{\hat{q}(\boldsymbol{z})}\Big[\mathbb{E}_{\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})}[-\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\alpha,\beta)] - \mathcal{H}(\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z}))\Big] - \mathcal{H}(\hat{q}(\boldsymbol{z})) \tag{8}$$

Recall that the goal of variational inference is to minimize the variational free energy $\hat{\mathcal{F}}[\hat{q}(\boldsymbol{z})\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})]$ with respect to the approximate posterior $\hat{q}$. The only parametrization that has been made concerns $\hat{q}(\boldsymbol{z})$ and the optimization deals with the new parameters $\hat{\gamma}_{ij}$. Indeed, the optimal choice for $\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})$ is obviously the real corresponding posterior $p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta)$. Thus, the term in squared brackets of Eq. 8 becomes:

$$\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta)}\Big[-\log p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\alpha,\beta)-\mathcal{H}(p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta))\Big] \tag{9}$$

$$=\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta)}\Big[-\log p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta)-\log p(\boldsymbol{x},\boldsymbol{z}|\alpha,\beta)-\mathcal{H}(p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta))\Big] \tag{10}$$

$$=\mathcal{H}(p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta))-\log p(\boldsymbol{x},\boldsymbol{z}|\alpha,\beta)-\mathcal{H}(p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{x},\boldsymbol{z},\alpha,\beta)) \tag{11}$$

$$=-\log p(\boldsymbol{x},\boldsymbol{z}|\alpha,\beta) \tag{12}$$

The new variational free energy $\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))=\min_{\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})}\hat{\mathcal{F}}[\hat{q}(\boldsymbol{z})\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})]$ to minimize is then:

$$\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))=\mathbb{E}_{\hat{q}(\boldsymbol{z})}[-\log p(\boldsymbol{x},\boldsymbol{z}|\alpha,\beta)]-\mathcal{H}(\hat{q}(\boldsymbol{z})) \tag{13}$$

One can observe that the parameters $\boldsymbol{\theta},\boldsymbol{\phi}$ have been marginalized out: this is the reason for which this method is called Collapsed Variational Bayesian (CVB) inference. Recall that for vanilla VB inference, we were looking for distributions of the form $\tilde{q}(\boldsymbol{z})\tilde{q}(\boldsymbol{\theta})\tilde{q}(\boldsymbol{\phi})$ to approximate the true posterior. The corresponding variational free energy $\tilde{\mathcal{F}}[\tilde{q}(\boldsymbol{z})\tilde{q}(\boldsymbol{\theta})\tilde{q}(\boldsymbol{\phi})]$ can also be minimized as what has been done for CVB until now. This gives a new variational free energy $\tilde{\mathcal{F}}[\tilde{q}(\boldsymbol{z})]=\min_{\tilde{q}(\boldsymbol{\theta})\tilde{q}(\boldsymbol{\phi})}\tilde{\mathcal{F}}[\tilde{q}(\boldsymbol{z})\tilde{q}(\boldsymbol{\theta})\tilde{q}(\boldsymbol{\phi})]$. While only one assumption has been done on $\hat{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})=\hat{q}(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})\,\hat{q}(\boldsymbol{z})$ – which is the family of distribution of $\hat{q}(\boldsymbol{z})$ – the assumptions on $\tilde{q}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})=\tilde{q}(\boldsymbol{z})\tilde{q}(\boldsymbol{\theta})\tilde{q}(\boldsymbol{\phi})$ are way stronger (independence between the latent variables and the parameters), so that the space of distributions over which the variational free energy is minimized is much bigger in the case of CVB inference than for VB inference. As a consequence, we have that:

$$\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))\leq\tilde{\mathcal{F}}(\hat{q}(\boldsymbol{z})) \tag{14}$$

and thus the CVB approximation is better than the VB one. The quantity $\tilde{\mathcal{F}}(\hat{q}(\boldsymbol{z}))-\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))$ then characterizes the supplementary bias that VB inference has compared to CVB inference.

2.2 Iterative solution

Now, in the same way than for VB inference, it remains to minimize the new variational free energy $\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))$ with respect to $\hat{q}(\boldsymbol{z})$, that is with respect to the parameters $\hat{\gamma}_{ij}$ because of the parametrization that has been made above. The fact that there is only 'one' parameter to optimize is a significant difference with the VB inference. Recall that $\hat{q}(z_{ij}|\hat{\gamma}_{ij})$ is multinomial with parameters $\hat{\gamma}_{ij}=(\hat{\gamma}_{ijk})_{1\leq k\leq K}$, that is $\hat{q}(z_{ij}=k|\hat{\gamma}_{ij})=\hat{\gamma}_{ijk}$. The result of the minimization of $\hat{\mathcal{F}}(\hat{q}(\boldsymbol{z}))$ with respect to $\hat{\gamma}_{ijk}$ is given in [1] in the form of updates since the closed-form solution involves the parameter itself. At step $t+1$ of the updates of parameters $(\hat{\gamma}_{ijk})_{1\leq k\leq K}$, the new value of $\hat{\gamma}_{ijk}^{(t+1)}$ depends on the previous values $(\hat{\gamma}_{ijk}^{(t)})_{1\leq k\leq K}$ as follows:

$$\hat{\gamma}_{ijk}^{(t+1)}=\frac{e^{\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma}^{(t)})}\Big[p(\boldsymbol{x},\boldsymbol{z}^{-ij},z_{ij}=k|\alpha,\beta)\Big]}}{\sum_{k'=1}^{K}e^{\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma}^{t)})}\Big[p(\boldsymbol{x},\boldsymbol{z}^{-ij},z_{ij}=k'|\alpha,\beta)\Big]}} \tag{15}$$

where the superscript $-ij$ put on a variable $\boldsymbol{z}$ (for instance) means that the associated variable includes all its components except the ij-th, that is all z_{kl} except z_{ij} in this example.

The problem of that result is that it involves the marginal distribution over $\boldsymbol{x}$ and $\boldsymbol{z}$. The derivation of this quantity is done in [10, 1] and gives the result:

$$p(\boldsymbol{x},\boldsymbol{z}|\alpha,\beta)=\prod_{j=1}^{D}\frac{\Gamma(K\alpha)}{\Gamma(K\alpha+n_{j\cdot\cdot})}\prod_{k=1}^{K}\frac{\Gamma(\alpha+n_{jk\cdot})}{\Gamma(\alpha)}\prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(W\beta+n_{\cdot k\cdot})}\prod_{w=1}^{W}\frac{\Gamma(\beta+n_{\cdot kw})}{\Gamma(\beta)} \tag{16}$$

Incorporating this expression in Eq. 15 and using an expansion of the logarithm of the Gamma function gives another version of the updates:

$$\hat{\gamma}_{ijk}^{(t+1)} = \frac{e^{\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma}^{(t)})}\left[\log(\alpha + n_{jk\cdot}^{-ij}) + \log(\beta + n_{\cdot kx_{ij}}^{-ij}) - \log(W\beta + n_{\cdot k\cdot}^{-ij})\right]}}{\sum_{k'=1}^{K} e^{\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma}^{(t)})}\left[\log(\alpha + n_{jk'\cdot}^{-ij}) + \log(\beta + n_{\cdot k'x_{ij}}^{-ij}) - \log(W\beta + n_{\cdot k'\cdot}^{-ij})\right]}} \tag{17}$$

2.3 An approximation necessity

The pain point is in calculating the expectations in the above formula. While the authors of [1] give an exact evaluation, it is computationally infeasible and thus they appeal to a "Gaussian approximation" for the expectations.

In details, each expectation term involved in Eq. 17 is approximated. The method is only explained for the term $\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma})} \log(\alpha + n_{jk\cdot}^{-ij})$ but it is directly applicable to other terms. Recall that $n_{jk\cdot}^{-ij} = \#\{i'|z_{i'j} = k, i' \neq i\} = \sum_{\substack{i'=1 \\ i'\neq i}}^{n_{j\cdot\cdot}} \mathbb{1}_{\{z_{i'j}=k\}}$. It is the sum of $n_{j\cdot\cdot}$ independent Bernoulli variables. Now assume that the number $n_{j\cdot\cdot}$ of words in document j is large (which is often the case in practice). Then, the variable $n_{jk\cdot}^{-ij}$ is close to a Gaussian distribution. This is the main approximation made in this part. Now, each of the Bernoulli variables $\mathbb{1}_{\{z_{i'j}=k\}}$ has a clear expectation according to distribution $\hat{q}$ since $\mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})} \mathbb{1}_{\{z_{i'j}=k\}} = \hat{q}(z_{i'j} = k|\hat{\gamma}_{i'j}) = \hat{\gamma}_{i'jk}$ and have thus variance $\hat{\gamma}_{i'jk}(1 - \hat{\gamma}_{i'jk})$. Those Bernoulli variables being independent, we have that:

$$\mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}] = \sum_{\substack{i'=1 \\ i'\neq i}}^{n_{j\cdot\cdot}} \hat{\gamma}_{i'jk} \qquad \text{Var}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}] = \sum_{\substack{i'=1 \\ i'\neq i}}^{n_{j\cdot\cdot}} \hat{\gamma}_{i'jk}(1 - \hat{\gamma}_{i'jk}) \tag{18}$$

The authors of [1] approximate the expectations of logarithms $\mathbb{E}_{\hat{q}(\boldsymbol{z}^{-ij}|\hat{\gamma})} \log(\alpha + n_{jk\cdot}^{-ij})$ with a second order Taylor series which involves $\mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}]$, $\mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}]^2$, $\text{Var}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}]$ and evaluate those terms under the Gaussian approximation written in Eq. 18. Their justification for which that second order approximation is "good enough" is that $\mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma})}[n_{jk\cdot}^{-ij}]$ is large.

Putting the results of the expansion of each expectation of logarithm present in Eq. 17, not yet evaluated with Gaussian approximations, yields the approximate update of the CVB algorithm:

$$\begin{aligned}\hat{\gamma}_{ijk}^{(t+1)} \propto\ & \frac{\left(\alpha + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{jk\cdot}^{-ij}]\right)\left(\beta + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot kx_{ij}}^{-ij}]\right)}{W\beta + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot k\cdot}^{-ij}]} \\ & e^{-\frac{\text{Var}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{jk\cdot}^{-ij}]}{2\left(\alpha + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{jk\cdot}^{-ij}]\right)^2} - \frac{\text{Var}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot kx_{ij}}^{-ij}]}{2\left(\beta + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot kx_{ij}}^{-ij}]\right)^2}} \\ & e^{\frac{\text{Var}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot k\cdot}^{-ij}]}{2\left(W\beta + \mathbb{E}_{\hat{q}(\boldsymbol{z}|\hat{\gamma}^{(t)})}[n_{\cdot k\cdot}^{-ij}]\right)^2}}\end{aligned} \tag{19}$$

For a final result, put in the above equation the results of Eq. 18. This update can be seen as the mean-field version of the collapsed Gibbs sampling updates for LDA [2].

3 Experiments

The authors of [1] did not provide any implementation of their algorithm, that they have tested on various datasets. Studying the variational bounds of the log marginal probabilities for their training sets (as presented in Eq. 4 for VB, and Eq. 14 for CVB), as functions of numbers of iterations, they pointed out that collapsed VB converged to a better variational bound than standard VB, but to a worse one than collapsed Gibbs. Moreover, as expected, they showed that collapsed VB (complexity in $O(MK)$, with M number of unique document/word pairs) converged quicker than collapsed Gibbs (complexity in $O(NK)$, with N total number of words in the corpus), but approximately as fast as standard variational bayes (complexity in $O(MK)$).

In our case, we tried to re-implement such results on new datasets. To do so, we used David Andrzejewski CVB-LDA implementation [11], and created an easy-to-use script to test it on any corpus of documents. All detailed informations on our work can be found on the following git: [12]. Because of a lack of time, we decided just to test the CVB-LDA algorithm on a toy example: the study of the n words the most representative of each one of the K topics explaining the corpus (n and K being chosen by the user). In fact, it simply consists in pointing out the first n indices of vector ϕ_k, for each $k \in K$. We then qualitatively compared these results to the one obtained using the Sklearn LDA implementation [13], which is based on stochastic variational bayes.

The results of our experiments, on the french Sequoia dataset [14] without parsing information (just sentences), can be seen in Fig. 2.

The two main topics of the sequoia corpus are 'detected': medical articles and political / judicial ones. Qualitatively, CVB LDA and Sklearn LDA give approximately the same results. Interestingly (and strangely), CVB LDA runs quite faster than Sklearn LDA. It could be explained by the stochasticity of the Sklearn LDA inference algorithm: it needs more iterations than VB to reach convergence on very small datasets, as Sequoia. Indeed, stochastic variational inference is thought to be used on massive datasets, where VB (and GS) cannot handle.

4 Conclusion

[1] is based on a smart intuition, derived from collapsed gibbs sampling, which gives birth to an elegant new inference method, computationally efficient and accurate. Tested on a small french dataset, as a toy example, it gave results comparable to the Sklearn LDA implementation (which uses stochastic variational inference). But tests on bigger datasets would have been necessary to draw true conclusions on the performances of CVB compared to stochastic variational inference and other inference methods, such as Gibbs sampling.

In 2012, a new approximate algorithm was introduced: CVB0 [15], which processes the update $\hat{\gamma}_{ijk}^{(t+1)}$ with zeroth-order information only, and not with a second-order Taylor expansion as in Eq. 19. This new inference algorithm is presented to be very fast (the fastest among VB, CVB and GS), while accurate. In [16], the CVB0 algorithm has been studied in the light of the α-divergence projection, showing in particular that CVB0 is not affected by the zero-forcing effect in LDA.

```
######## DATA LOADING AND PROCESSING ########

VOCAB SIZE: 8251 WORDS

######## COLLAPSED VARIATIONAL INFERENCE LDA ########

Collapsed variational inference LDA exec time: 0.381174087524s

Topics found via collapsed variational inference LDA:

Topic 1: 10 most important words, with p(w|z):
affaire, 0.8127481%
paris, 0.4366119%
president, 0.4365667%
deux, 0.3874088%
etait, 0.3427752%
commission, 0.3388877%
ans, 0.3265192%
dont, 0.2801829%
juge, 0.2785372%
ancien, 0.2730116%

Topic 2: 10 most important words, with p(w|z):
patients, 1.5233556%
aclasta, 1.3945385%
angiox, 0.71278%
bivalirudine, 0.6824946%
perfusion, 0.6370641%
traitement, 0.637059%
mg, 0.5916037%
doit, 0.4835385%
effets, 0.4552107%
voir, 0.4395397%

######## SKLEARN LDA COMPARISON ########

Sklearn LDA exec time: 6.565928936s

Topics found via Sklearn LDA:

Topic 1: 10 most important words, with p(w|z):
rrb, 1.9544326%
lrb, 1.9532899%
patients, 0.9570906%
aclasta, 0.8762633%
angiox, 0.4479074%
bivalirudine, 0.4297532%
traitement, 0.4013508%
perfusion, 0.4008852%
mg, 0.3725444%
doit, 0.3497594%

Topic 2: 10 most important words, with p(w|z):
affaire, 0.6752488%
commission, 0.4087772%
president, 0.3725456%
paris, 0.3635402%
jean, 0.3146196%
conseil, 0.2580898%
2006, 0.2577587%
solution, 0.2411031%
etait, 0.2382889%
juge, 0.2326179%
```

Figure 2: CVB-LDA script example on Sequoia dataset for $n = 10$ and $K = 2$, compared with Sklearn-LDA results on the same dataset

References

[1] Yee Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. volume 19, pages 1353–1360, 01 2006.

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

[3] Thomas Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35, 04 2004.

[4] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents, 2012.

[5] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, page 524–531, USA, 2005. IEEE Computer Society.

[6] Xiaogang Wang and Eric Grimson. Spatial latent dirichlet allocation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1577–1584. Curran Associates, Inc., 2008.

[7] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, page 433–434, New York, NY, USA, 2003. Association for Computing Machinery.

[8] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 352–359, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[9] Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference, 2012.

[10] `https://lingpipe.files.wordpress.com/2010/07/lda3.pdf`.

[11] `https://github.com/davidandrzej/cvbLDA`.

[12] `https://github.com/tristandot/collapsed_variational_inference_LDA_script`.

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#examples-using-sklearn-decomposition-latentdirichletallocation`.

[14] `https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia`.

[15] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models, 2012.

[16] Issei Sato and Hiroshi Nakagawa. Rethinking collapsed variational bayes inference for lda. *arXiv preprint arXiv:1206.6435*, 2012.