A COLLAPSED VARIATIONAL BAYESIAN INFERENCE ALGORITHM FOR LATENT DIRICHLET ALLOCATION

Baraille Chloe, Darin Edith, Sebbouh Othmane

Abstract. — We consider the article *A collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation* written by Yee Whye Teh, David Newman and Maw Welling in 2006. The goal of the paper is to propose a collapsed variational inference algorithm for LDA (Latent Dirichlet Allocation), which is to outperform current inference procedures like variational Bayesian inference or collapsed Gibbs sampling. We prove the outlined formulas when this serves the purpose of the explanation. Further, in the light of the article, we discuss CVB0, an amelioration of CVB and stochastic approaches.

Contents

1. Introduction

1.1. Latent Dirichlet Allocation. — Following its publication on 2003, Blei et al.'s *Latent Dirichlet Allocation* has made topic modeling - a subfield of machine learning applied to everything from computational linguistics to informatics and political science - one of the most successful paradigms for both supervised and unsupervised learning.

Latent Dirichlet Allocation is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Formally, we define the following terms:

– A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1,...,V. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, the vth word

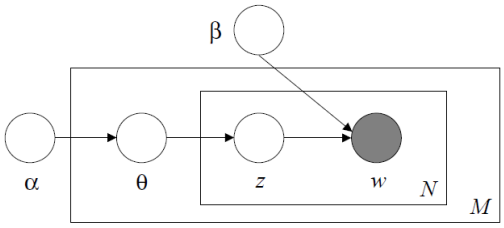


FIGURE 1. The parameters α and β are corpus-level parameters, the variables θ_d are document-level variables, the variables z_{dn} and w_{dn} are word-level variables

in the vocabulary is represented by a V-vector w such that $w_v = 1$ and $w_u = 0$ for $u \neq v$.
- A document is a sequence of N words denoted by $W = (w_1, w_2, ..., w_N)$, where w_n is the nth word in the sequence.
- A corpus is a collection of M documents denoted by $D = W_1, W_2, ..., W_M$.

LDA assumes the following generative process for each document w in a corpus D:
- 1. Choose $N \sim Poisson(\xi)$.
- 2. Choose $\theta \sim Dir(\alpha)$.
- 3. For each of the N words w_n:
  - (a) Choose a topic $z_n \sim Multinomial(\theta)$.
  - (b) Choose a word w_n from $P(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model. First, the dimensionality K of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a kÃ̊UV matrix β where $\beta_{ij} = p(w_j = 1|z_i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and z). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

Given the parameters α and β, the joint distribution of a topic mixture *theta*, a set of N topics z, and a set of N words w is given by:

$$P(\theta, z, w|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(z_n|\theta) P(w_n|z_n, \beta)$$

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation.

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$P(\theta, z|w, \alpha, \beta) = \frac{P(\theta, z, w|\alpha, \beta)}{P(w|\alpha, \beta)}$$

1.2. Inference. — The goal of the paper is to propose a collapsed variational inference algorithm for LDA (Latent Dirichlet Allocation), which is to outperform traditional

inference procedures like variational Bayesian inference or collapsed Gibbs sampling.

The paper leverages on the fact that a Gibbs sampler that operates in a collapsed space, where the parameters are marginalized out, performs better than a Gibbs sampler that samples the parameters and the latent variables simultaneously. This suggests that the parameters and the latent variables are intimately coupled.

Marginalizing out the parameters induces new dependencies between the latent variables, which are conditionally independent given the parameters. These dependencies are spread out over many latent variables, which implies that the dependency between any two latent variables will be small. This is the right setting for a mean field (ie fully factorized variational) approximation: a particular variable interacts with the remaining variables only through summary statistics called the field, and the impact of any single variable on the field is small. This is not true in the joint space of parameters and latent variables because fluctuations in parameters can have a significant impact on latent variables. The paper thus conjectures that the mean field assumptions are much better satisfied in the collapsed space of latent variables than in the joint space of latent variables and parameters.

2. Approximate inference in LDA

We now consider the model defined by the article of Teh, Neman and Welling. The authors assume there are :

- K latent topics. Each being a multinomial distribution over a vocabulary of size W, which we note $z_n \sim Multinomial(W)$
- D documents. For each one, we draw a mixing proportion $\theta_j = \{\theta_{jk}\}$ over K from a $Dir(\alpha)$
  - for the i-th word of the document, a topic z_{ij} is drawn with probability θ_{jk}
  - then, a word x_{ij} is drawn from the z_{ij} topic, with x_{ij} taking the value w with probability ϕ_{kw}, where $\phi_k = \{\phi_{kw}\}$ is drawn from a $Dir(\beta)$

Then, the joint distribution of the model is given by:

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

where $n_{jkw} = \#\{i : x_{ij} = w, z_{ij} = k\}$, i.e. the number of times the word w, drawn from topic k, appears in the document j.

Proof. —

$$p(\mathbf{x}, \mathbf{z}, \theta, \phi) = p(\theta, \phi | \alpha, \beta) p(\mathbf{x}, \mathbf{z} | \theta, \phi, \alpha, \beta)$$

Since θ and ϕ are independent, their joint density is the product of their densities. Then:

$$p(\mathbf{x}, \mathbf{z}, \theta, \phi) = p(\theta | \alpha) p(\phi | \beta) p(\mathbf{x}, \mathbf{z} | \theta, \phi, \alpha, \beta) \tag{1}$$

For all $j = 1...d$, θ_j is drawn from a $Dir(\alpha)$, hence, since all thetas are i.i.d.:

$$p(\theta | \alpha) = \prod_{j=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{jk}^{\alpha - 1} \tag{2}$$

The same goes for ϕ:

$$p(\phi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \phi_{kw}^{\beta-1} \tag{3}$$

We also have:

$$\begin{aligned}
p(\mathbf{x}, \mathbf{z}|\theta, \phi, \alpha, \beta) &= \prod_{w=1}^{W} \prod_{j=1}^{D} p(x_{ij}, z_{ij}|\theta, \phi, \alpha, \beta) \\
&= \prod_{w=1}^{W} \prod_{j=1}^{D} p(z_{ij}|\theta, \alpha, \beta) p(x_{ij}|z_{ij}, \phi, \alpha, \beta) \\
&= \prod_{w=1}^{W} \prod_{j=1}^{D} \prod_{k=1}^{K} \theta_{jk}^{n_{jkw}} \times \prod_{w=1}^{W} \prod_{j=1}^{D} \prod_{k=1}^{K} \phi_{kw}^{n_{jkw}} \\
&= \prod_{j=1}^{D} \prod_{k=1}^{K} \prod_{w=1}^{W} \theta_{jk}^{n_{jkw}} \times \prod_{w=1}^{W} \prod_{k=1}^{K} \prod_{j=1}^{D} \phi_{kw}^{n_{jkw}} \\
&= \prod_{j=1}^{D} \prod_{k=1}^{K} \theta_{jk}^{\sum_{w=1}^{W} n_{jkw}} \times \prod_{w=1}^{W} \prod_{k=1}^{K} \phi_{kw}^{\sum_{j=1}^{D} n_{jkw}} \\
&= \prod_{j=1}^{D} \prod_{k=1}^{K} \theta_{jk}^{n_{jk.}} \times \prod_{w=1}^{W} \prod_{k=1}^{K} \phi_{kw}^{n_{.kw}}
\end{aligned} \tag{4}$$

Where we used in (4):

- In the third equality, the fact that z_{ij} is drawn with topic k with probability θ_{jk} and that knowingg z_{ij}, x_{ij} takes the value w with probability ϕ_{kw}.
- In the last equality, the definitions: $n_{jk.} = \sum_{w=1}^{W} n_{jkw}$ and $n_{.kw} = \sum_{j=1}^{D} n_{jkw}$

Plugging the result of the equations (2), (3) and (4) in (1):

$$\begin{aligned}
p(\mathbf{x}, \mathbf{z}, \theta, \phi) &= \prod_{j=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{jk}^{\alpha-1} \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \phi_{kw}^{\beta-1} \prod_{j=1}^{D} \prod_{k=1}^{K} \theta_{jk}^{n_{jk.}} \times \prod_{w=1}^{W} \prod_{k=1}^{K} \phi_{kw}^{n_{.kw}} \\
&= \prod_{j=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{jk}^{\alpha-1} \theta_{jk}^{n_{jk.}} \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \phi_{kw}^{\beta-1} \phi_{kw}^{n_{.kw}} \\
&= \prod_{j=1}^{D} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{jk}^{\alpha-1+n_{jk.}} \times \prod_{k=1}^{K} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \theta_{kw}^{\beta-1+n_{.kw}}
\end{aligned} \tag{5}$$

□

The task we have to perform in Bayesian inference is to find the posterior distributions of the topics $\mathbf{z}$, the proportions θ and the topic parameters ϕ. This amounts to finding:

$$p(\theta, \phi, z|x, \alpha, \beta) = \frac{p(\theta, \phi, z, x|\alpha, \beta)}{p(x|\alpha, \beta)}$$

Unfortunately, this distribution is intractable to compute. The normalization factor in particular, $p(x|\theta,\beta)$, cannot be computed exactly. To solve this issue, a number of approximate inference techniques are available. The authors examine two approaches: variational Bayes (VB) and collapsed Gibbs sampling, which we are going to explain in detail in the next two sections.

2.1. Variational Bayesisan. — Variational Bayes inference leads to upper bounding the negative log marginal likelihood $-\log p(x|\alpha,\beta)$ using the variational free energy:

$$-\log p(\mathbf{x}|\alpha,\beta) \leq \tilde{\mathcal{F}}(\tilde{q}(\mathbf{z},\theta,\phi)) = E_{\tilde{q}}[-\log p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta)] - \mathcal{H}(\tilde{q}(\mathbf{z},\theta,\phi))$$

where we call the term: $\tilde{\mathcal{F}}(\tilde{q}(\mathbf{z},\theta,\phi))$ the variational free energy.

Proof. — The variational Bayes inference tries to resolve the intractability of the joint distribution in (1). To do so, it aims to approximate this joint distribution by another one, which is easier to calculate. The approximation is considered in the sense of the Kulback-Leibler divergence, i.e. we have to find the distribution q that minimizes the KL divergence. The choice of the KL divergence is motivated by the fact that we need an asymmetric criterion for the approximation of the joint distribution. Indeed, we want q to be accurate when we predict it to be high, even if we are wrong when p is high. This is achieved with the KL divergence.

In the following, we drop α and β for clarity, and we note $Y = (\mathbf{z},\theta,\phi)$:

$$KL(q||p_x) = \int q(y) \log \frac{q(y)}{p(y|x)} dy$$

Since $p(y|x) = \frac{p(y,x)}{p(x)}$:

$$\begin{aligned} KL(q||p_x) &= \int q(y)(\log \frac{q(y)}{p(y,x)} + \log p(x))dy \\ &= \int q(y) \log \frac{q(y)}{p(y,x)} dy + \int q(y) \log p(x) dy \\ &= \int q(y) \log \frac{q(y)}{p(y,x)} dy + \log p(x) \end{aligned}$$

We define $L(q) = \int q(y) \log \frac{q(y)}{p(y,x)} dy$. Then:

$$-\log p(x) = L(q) - KL(q||p_x)$$

Since $KL(q||p_x) \geq 0$, $-\log p(x)$ is upper bounded by $L(q)$. And since the residual term doesn't depend on q, minimizing the KL divergence amounts to minimizing the upper bound $L(q)$ in q.

We can rewrite $L(q)$ as :

$$L(q) = -\int q(y)p(y,x)dy + \int q(y) \log q(y) dy \quad = E_q[-\log p(x,Y)] - \mathcal{H}(q(y))$$

where : $\mathcal{H}(Y) = E_q[-\log q(Y)] = \int -q(y) \log q(y) dy$

The upper bounding by $L(q)$ rewrites, using our notation $Y = (\mathbf{z}, \theta, \phi)$ and using α and β again:

$$-\log p(\mathbf{x}|\alpha, \beta) \leq E_{\tilde{q}}[-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)] - \mathcal{H}(\tilde{q}(\mathbf{z}, \theta, \phi))$$

□

We have stated that Variational Bayes inference consists in finding the closest distribution possible to the posterior distribution: $p(\theta, \phi, z|x, \alpha, \beta)$ in the sense of the KL divergence.
First, we make a hypothesis over this distribution, which is a consequence of the structure of the data. Since the latent variables (z, θ, ϕ are very numerous, there is a good chance that the dependency between any two of them is very small. We make the hypothesis that the distribution we are looking for to be fully factorized.

Then, to find this distribution, we restrain our research to a particular family of distributions. We consider the family of joint distributions of a different yet close model, where we include three new variables model $\tilde{\alpha}, \tilde{\beta}$, and γ (called variational variables) such that:

- $z_{ij}|\gamma_{ij}$ is a $Multinomial(\gamma_{ij})$
- $\theta_j|\tilde{\alpha_j}$ is a $Dir(\tilde{\alpha_j})$
- $\phi_k|\tilde{\beta_j}$ is a $Dir(\tilde{\beta_j})$

These variables are the ones over which we want to minimize the KL divergence:

$$(\gamma^*, \tilde{\theta^*}, \tilde{\beta^*}) = \underset{(\tilde{\gamma}, \tilde{\theta}, \tilde{\beta})}{\operatorname{argmin}} KL(\tilde{q}(\mathbf{z}, \theta, \phi)||p(\theta, \phi, z, x|\alpha, \beta))$$

Which amounts, as explained in the proof, to minimize the variational free energy $E_{\tilde{q}}[-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)] - \mathcal{H}(\tilde{q}(\mathbf{z}, \theta, \phi))$.

In this project, we will not detail this optimization problem. But we note that we can find solutions numerically.

The problem of the Variational Bayes inference is that it assumes all latent variables and the Dirichlet parameters to be independent, which enabled us above to write the posterior distribution in a fully factorized form. In reality, these variables can be very dependent. This can lead to inaccurate results.

2.2. Collapsed Gibbs Sampling. — Gibbs sampling provides an alternative solution for approximating the posterior distribution $p(\theta, \phi, z|x, \alpha, \beta)$.

We first recall the standard algorithm of Gibbs Sampling. Gibbs Sampling is a member of a family of algorithms from the Markov Chain Monte Carlo (MCMC) framework which aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to be close to sampling.from the desired posterior. Gibbs Sampling is based on sampling from conditional distributions of the variables of the posterior. For example, to sample x from the joint distribution $p(x) = p(x_1, ..., x_m)$, where there is no closed form solution for $p(x)$, but a representation for the conditional distributions is available, using Gibbs Sampling one would perform the following:

- Randomly initialize each x_i

- For $t = 1...T$:
  - $x_1^{t+1} \sim p(x_2^t, x_3^t, ..., x_N^t)$
  - $x_2^{t+1} \sim p(x_1^t, x_3^t, ..., x_N^t)$
  - ...
  - $x_N^{t+1} \sim p(x_1^t, x_2^t, ..., x_{N-1}^t)$

In the framework on LDA, Gibbs sampling can be used to approximate the posterior distribution $p(\theta, \phi, z|x, \alpha, \beta)$. The words x are observed, while the latent variable z, corresponding to the topics of words in documents, on the one hand, and the parameters θ and ϕ, which refer respectively to the document-topic and topic-word distributions, on the other hand, are unobserved. Standard Gibbs sampling, which iteratively samples latent variables z and parameters θ, ϕ can potentially have slow convergence due to strong dependencies between the parameters and latent variables. (The convergence is theoretically guaranteed with Gibbs Sampling, but there is no way of knowing how many iterations are required to reach the stationary distribution.) However, one can note that both θ (latent document-topic proportions) and ϕ (latent topic-word distributions) can be calculated using z (topic index assignments): z is a sufficient statistic for both these distributions. Indeed:

$$\theta_{d,z} = \frac{n(d,z) + \alpha}{\sum_{|Z|} n(d,z) + \alpha}, \phi_{z,x} = \frac{n(z,x) + \beta}{\sum_{|W|} n(z,x) + \beta}$$

Therefore, a simpler algorithm can be used if we integrate out the multinomial parameters θ and ϕ and simply sample z. This is called a collapsed Gibbs sampler.

In the framework of collapsed Gibbs sampling, we are interested in the following posterior:

$$\begin{aligned} p(z_i|z_{-i}, \alpha, \beta, x) &= \frac{p(z_i, z_{-i}, x|\alpha, \beta)}{p(z_{-i}, x|\alpha, \beta)} \\ &\propto p(z_i, z_{-i}, x|\alpha, \beta) = p(z, x|\alpha, \beta) \\ &= \int\int p(z, x, \theta, \phi|\alpha, \beta) d\theta d\phi \\ &= \int\int p(\phi|\beta) p(\theta|\alpha) p(z|\theta) p(x|\phi_z) d\theta d\phi \\ &= \int p(z|\theta) p(\theta|\alpha) d\theta \int p(x|\phi_z) p(\phi|\beta) d\phi \end{aligned}$$

where z_{-i} refers to all topic allocations z_k except z_i.

Both terms are multinomial with Dirichlet priors, and Dirichlet distributions are conjugate with multinomial distributions. Therefore:

$$
\begin{aligned}
\int p(z|\theta)p(\theta|\alpha)d\theta &= \int \prod_i \theta_{j,z_i} \frac{1}{B(\alpha)} \prod_k \theta_{j,z}^{\alpha^k} d\theta_j \\
&= \frac{1}{B(\alpha)} \int \prod_k \theta_{j,k}^{n_{j,k}+\alpha_k} d\theta_j \\
&= \frac{B(n_{j..}+\alpha)}{B(\alpha)}
\end{aligned} \tag{6}
$$

where $n_{j,k,w}$ indicates the number of times word w in document j is assigned to topic k. A . indicates a sum over the index. $B(\alpha)$ is the multinomial beta function defined by:

$$
B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)}
$$

$$
B(n_{j..}+\alpha) = \frac{\prod_k \Gamma(n_{jk}+\alpha_k)}{\Gamma(\sum_k n_{jk.}\alpha_k)} = \frac{\prod_k \Gamma(n_{jk.}+\alpha)}{\Gamma(K\alpha+n_{j..})}
$$

Similarly:

$$
\begin{aligned}
\int p(x|\phi_z)p(\phi|\beta)d\phi &= \int \prod_j \prod_i \phi_{z_{j,i},x_{j,i}} \prod_k \frac{1}{B(\beta)} \prod_x \phi_{k,x}^{\beta_x} d\phi_k \\
&= \prod_k \frac{1}{B(\beta)} \int \prod_x \phi_{k,x}^{\beta_x,+n_{k,x}} d\phi_k \\
&= \prod_k \frac{B(n_{.k.}+\beta)}{B(\beta)}
\end{aligned} \tag{7}
$$

Combining (7) and (8), we obtain:

$$
p(z,x|\alpha,\beta) = \prod_j \frac{B(n_{j..}+\alpha)}{B(\alpha)} \prod_k \frac{B(n_{.k.}+\beta)}{B(\beta)} \tag{8}
$$

We finally derive the Gibbs Sampling equation, using (8):

$$
\begin{aligned}
p(z_i|z^{-i},x,\alpha,\beta) &= \frac{p(x,z|\alpha,\beta)}{p(x,z^{-i}|\alpha,\beta)} = \frac{p(z|\alpha,\beta)}{p(z^{-i}|\alpha,\beta)} . \frac{p(x|z,\alpha,\beta)}{p(x^{-i}|z^{-i}\alpha,\beta)p(x_i|\alpha,\beta)} \\
&\propto \prod_j \frac{B(n_{j..}+\alpha)}{B(n_{j..}^{-i}+\alpha)} \prod_k \frac{B(n_{.k.}+\beta)}{B(n_{.k.}^{-i}+\beta)} \\
&\propto \frac{\Gamma(n_{jk.}+\alpha)\Gamma(n_{j..}^{-i}+K\alpha)}{\Gamma(n_{jk.}^{-i}+\alpha)\Gamma(n_{j..}+K\alpha)} . \frac{\Gamma(n_{.kw}+\beta)\Gamma(n_{.k.}^{-i}+W\beta)}{\Gamma(n_{.kw}^{-i}+\beta)\Gamma(n_{.k.}+W\beta)} \\
&\propto (n_{jk.}^{-i}+\alpha) . \frac{n_{.kw}^{-i}+\beta}{n_{.k.}^{-i}+\beta}
\end{aligned} \tag{9}
$$

Collapsed Gibbs sampling has been observed to converge quickly. One can notice from (9) that z_i depends on z_{-i} only through $njk.^{-i}, n.k.^{-i}, n.kw^{-i}$. In particular, the dependence of z_{ij} on any particular other variable $z_{i'j'}$ is very weak, especially for large

datasets. As a result the convergence of collapsed Gibbs sampling is expected to be fast. However, as with other MCMC samplers, and unlike variational inference, it is often hard to diagnose convergence, and a sufficiently large number of samples may be required to reduce sampling noise.

Last thing noteworthy is that, in theory, the CVB algorithm requires the calculation of very expensive averages. However, the averages only depend on sums of independent Bernoulli variables, and thus are very closely approximated with Gaussian distributions (even for relatively small sums).

3. Collapsed Variational Bayesian

Recall the disadvantages of Variational Bayes and Collapsed Gibbs sampling:

- VB assumes that the latent variables and the parameters are independent, which allows for a fully factorized form of the posterior density. This assumption is often wrong as the dependency between the latent variable and the parameters can be high, which can lead to inaccurate results.
- Collapsed Gibbs sampling is too time consuming

Here, we keep one assumption from collapsed Gibbs sampling: we still assume that the latent variables $\mathbf{z}$ are mutually independent. The posterior distribution of the latent $(\mathbf{z}, \theta, \phi)$ variables writes:

$$\hat{q}(\mathbf{z}, \theta, \phi) = \hat{q}(\theta, \phi|\mathbf{z})\hat{q}(\mathbf{z})$$

Since the latent variables $\mathbf{z}$ are independent:

$$\hat{q}(\mathbf{z}, \theta, \phi) = \hat{q}(\theta, \phi|\mathbf{z}) \prod_{ij} \hat{q}(\mathbf{z}_{ij}, \gamma_{ij})$$

with $\hat{q}(\mathbf{z}_{ij}|\gamma_{ij})$ the density of a $Multinomial(\gamma_{ij})$, as in the VB inference method presented above.

We still aim to minimize the variational free energy with respect to the joint posterior distribution $\hat{\mathcal{F}}(\hat{q}(\mathbf{z}, \theta, \phi))$.

$$\begin{aligned}\hat{\mathcal{F}}(\hat{q}(\mathbf{z}, \theta, \phi)) &= \hat{\mathcal{F}}(\hat{q}(\theta, \phi|\mathbf{z})\hat{q}(\mathbf{z})) \\ &= E_{\hat{q}(\mathbf{z})\hat{q}(\theta,\phi|\mathbf{z})}[-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)] - \mathcal{H}(\hat{q}(\theta, \phi|\mathbf{z})\hat{q}(\mathbf{z}))\end{aligned}$$

The first term rewrites:

$$\begin{aligned}E_{\hat{q}(\mathbf{z})\hat{q}(\theta,\phi|\mathbf{z})}[-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)] &= \int -\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)\hat{q}(\theta, \phi|\mathbf{z})\hat{q}(\mathbf{z})dzd(\theta, \phi) \\ &= \int (-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)\hat{q}(\theta, \phi|\mathbf{z})d(\theta, \phi))\, \hat{q}(\mathbf{z})dz \\ &= \int E_{\hat{q}(\theta,\phi|\mathbf{z})}[-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta)]\hat{q}(\mathbf{z}) \\ &= E_{\hat{q}(\mathbf{z})}\left[E_{\hat{q}(\theta,\phi|\mathbf{z})}(-\log p(\mathbf{x}, \mathbf{z}, \phi, \theta|\alpha, \beta))\right]\end{aligned}$$

The second term rewrites, $\mathcal{H}$ being the entropy:

$$\begin{aligned}\mathcal{H}(\hat{q}(\theta,\phi|\mathbf{z})\hat{q}(\mathbf{z})) &= E_{\hat{q}(\theta,\phi|\mathbf{z})\hat{q}(\mathbf{z})}[-log(\hat{q}(\theta,\phi|\mathbf{z})\hat{q}(\mathbf{z}))] \\ &= E_{\hat{q}(\theta,\phi|\mathbf{z})\hat{q}(\mathbf{z})}[-log(\hat{q}(\theta,\phi|\mathbf{z}))] + E_{\hat{q}(\theta,\phi|\mathbf{z})\hat{q}(\mathbf{z})}[-log(\hat{q}(\mathbf{z}))] \\ &= E_{\hat{q}(\theta,\phi|\mathbf{z})}[-log(\hat{q}(\theta,\phi|\mathbf{z}))] + E_{\hat{q}(\mathbf{z})}[-log(\hat{q}(\mathbf{z}))] \\ &= E_{\hat{q}(\mathbf{z})}[\mathcal{H}(\hat{q}(\theta,\phi|\mathbf{z}))] + \mathcal{H}(\hat{q}(\mathbf{z})))\end{aligned}$$

Summing the first term and the second leads to the following formulation of the leads to the following expression of the variational free energy:

$$\hat{\mathcal{F}}(\hat{q}(\mathbf{z},\theta,\phi)) = E_{\hat{q}(\mathbf{z})}\left[E_{\hat{q}(\theta,\phi|\mathbf{z})}(-\log p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta)) - \mathcal{H}(\hat{q}(\theta,\phi|\mathbf{z}))\right] - \mathcal{H}(\hat{q}(\mathbf{z}))$$

To minimize the variational free energy, we first minimize what is inside the expectation with respect to $\hat{q}(\theta,\phi|\mathbf{z})$. Recall, in the framework of VB inference, the fact that $\hat{q}(\theta,\phi|\mathbf{z})$ is an upper bound of the joint distribution $p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)$. Hence the minimum of $\hat{q}(\theta,\phi|\mathbf{z})$ is achieved at the true joint distribution. Hence, replacing $\hat{q}(\theta,\phi|\mathbf{z})$ by the true distribution in the formula to minimize, we have:

$$\begin{aligned}&E_{\hat{q}(\theta,\phi|\mathbf{z})}(-\log p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta)) - \mathcal{H}(\hat{q}(\theta,\phi|\mathbf{z})) \\ &= E_{p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)}(-\log p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta)) - \mathcal{H}(p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)) \\ &= E_{p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)}[-\log p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta) - \log(p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta))] \\ &= -\log\left(p(\mathbf{x},\mathbf{z},\phi,\theta|\alpha,\beta)p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)\right) \\ &= -\log p(\mathbf{x},\mathbf{z}|\alpha,\beta)\end{aligned}$$

Now that we minimized what was inside the expectation taken with respect to $q(\hat{z})$, the problem becomes finding the minimum, with respect to $q(\hat{z})$, of:

$$\hat{\mathcal{F}}(\hat{q}(z)) = E_{\hat{q}(z)}[-log(p(\mathbf{x},\mathbf{z}|\alpha,\beta))] - \mathcal{H}(\hat{q}(\mathbf{z})) \tag{10}$$

Note that since with CVB, we are making weaker assumptions than with VB, the universe on which we minimize the Variational free energy is larger in CVB than in VB. Hence, if we find a minimum to the CVB problem, it is lower than the one found with VB.

Minimizing (10) with respect to $\gamma_{\hat{ijk}}$, we get:

$$\hat{\gamma}_{ijk} = \hat{q}(z_{ij}=k) = \frac{\exp(E_{\hat{q}(z^{-ij})}[p(\mathbf{x},\ \mathbf{z}^{-ij}, z_{ij}=k|\alpha,\beta)])}{\sum_{k'=1}^{K}\exp(E_{\hat{q}(z^{-ij})}[p(\mathbf{x},\ \mathbf{z}^{-ij}, z_{ij}=k'|\alpha,\beta)])} \tag{11}$$

Plugging in (8) and expanding $\log\frac{\Gamma(\eta+n)}{\Gamma(\eta)}$ for positive reals η and positive integers n, and canceling terms appearing both in numerator and denominator, we get:

(12)

$$\hat{\gamma}_{ijk} = \frac{\exp\left(E_{\hat{q}(z^{-ij})}[\log(\alpha + njk.^{-ij}) + \log(\beta + n^{-ij}_{.kx_{ij}}) - \log(W\beta + n_{.k.^{-ij}})]\right)}{\sum_{k'=1}^{K}\exp\left(E_{\hat{q}(z^{-ij})}[\log(\alpha + njk'.^{-ij}) + \log(\beta + n^{-ij}_{.k'x_{ij}}) - \log(W\beta + n_{.k'.^{-ij}})]\right)}$$

4. Discussion

Traditional inference techniques such as Gibbs sampling and variational inference do not readily scale to corpora containing millions of documents or more. In such cases it

is very time-consuming to run even a single iteration of the standard collapsed Gibbs sampling or variational Bayesian inference algorithms.

The collapsed representation, where parameters are marginalized out, leaving only latent variables, contributed to improve inference in LDA. In this framework, it is possible to perform inference in the collapsed space and recover estimates of the parameters afterwards. Algorithms such as Gibbs Sampling and Variational Bayes that operate in a collapsed space have proven to be more efficient, because the per-token updates propagate updated information sooner, the update equations of these algorithms are simpler, there are fewer parameters to update, these algorithms make no expensive calls to the digamma function...

I the paper we discussed above, Teh et al. introduced the collapsed variational inference (CVB). For variational inference, an important advantage of the collapsed representation is that the variational bound is strictly better than that for the uncollapsed representation, leading to the potential for collapsed variational algorithms to learn more accurate topic models than uncollapsed variational algorithms.

Teh et al. showed that an algorithm using approximate updates works well in practice, outperforming the classical VB algorithm in terms of prediction performance. Asuncion et al. later showed that a simpler version of this method called CVB0, based on additional approximations, is much faster while still maintaining the accuracy of CVB. The CVB0 algorithm iteratively updates each γ_{ijk} via :

$$\gamma_{ijk} \propto \frac{N_{x_{ij}k}^{\phi-ij} + \eta_{x_{ij}}}{N_k^{Z-ij} + \sum_x \eta_x}(N_{jk}^{\theta-ij} + \alpha)$$

where the N^Z, N^θ and N^ϕ variables are variational expected counts corresponding to their indices. Specifically, N^Z is the vector of expected number of words assigned to each topic, N^θ is the equivalent vector for document j only, and each entry w, k of matrix N^ϕ is the expected number of times word w is assigned to topic k across the corpus.

A disadvantage of CVB0 is that the memory requirements are large as it needs to store a variational distribution γ for every token in the corpus

A significant advance was made by Hoffman et al., who proposed a stochastic variational inference algorithm for LDA topic models. Because the algorithm does not need to see all of the documents before updating the topics, this method can often learn good topics before a single iteration of the traditional batch inference algorithms would be completed. The algorithm processes documents in an online fashion, so it can be applied to corpora of any size, or even to never-ending streams of documents.

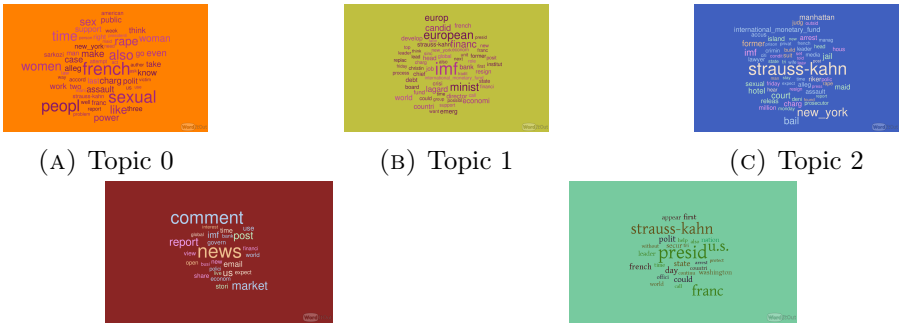
Leveraging on both the collapsed variational inference and on the stochastic variational inference, Foulds et al. propose a stochastic algorithm for collapsed variational Bayesian inference for LDA. In experiments on large-scale text corpora, the algorithm was found to converge faster and often to a better solution than previous methods. It has also been demonstrated that the method can learn coherent topics in seconds on small corpora, facilitating the use of topic models in interactive document analysis software.

5. Implementation

In order to illustrate the Latent Dirichlet Allocation modeling and its implementation, we've decided to work on a corpus that gathers English news articles on the scandal of Dominique Strauss-Kahn in New-York. The application's idea is to understand which frames have been used by the media and the journalists to report on this issue. Indeed we can think of an article on a given matter as a choice over a set of framing and then a choice over the words in that framing set. This fits particularly well in the LDA modeling and its bayesian approach.

The full implementation is to be read in a separate notebook. We can already state that to our great discontent we have finally chosen to implement along the collapsed gibbs sampler strategy since it was easier to understand in a given set of time due to its closeness to other algorithm that we've seen in the semester. Furthermore since it was stated that this methodology was performing quite well but just running slowly, we solve the dilemma by drastically reducing the corpus sample size.

We can already spoil one practical result by showing the following wordclouds that depict the topic set generated from the corpus. Let's imagine what have been the underlying frames !



(A) Topic 0 (B) Topic 1 (C) Topic 2

(D) Topic 3 (E) Topic 4

FIGURE 2. Wordclouds for each topic estimated with LDA

Documentation and sources

[1] TEH, NEWMAN, WELLING — *A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation*, NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing Systems, 2006.

[2] BLEI, NG AND JORDAN — *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 2003.

[3] WILLIAM M.DARLING — *A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling*, 2011.

[4] FOX AND ROBERTS — *A Turorial on Variational Bayesian inference*, 2011.

[5] FOULDS, BOYLES, DUBOIS, SMYTH, WELLING — *Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation*, 2013.

[6] ASUNCION, SMYTH, WELLING — *On smoothing and inference for topic models*, Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence, 2009.

[7] Hoffman, Blei, Bach — *Online Learning for Latent Dirichlet Allocation*, Advances in Neural Information Processing Systems, 2010.

January 30th 2018

Baraille Chloe, Darin Edith, Sebbouh Othmane