PGM Intermediary Report
Latent Dirichlet Allocation

Quentin LEROY, Vincent MATTHYS, Bastien PONCHON
{qleroy,vmatthys,bponchon}@ens-paris-saclay.fr

The paper we are working on is **Latent Dirichlet Allocation** *by Bley, Ng and Jordan.*

1 The model

In this paper, the authors present a generative probabilistic model of a corpus of text documents or of other kinds of collections of discrete data. This model is parametrized by some hyper-parameters $\xi \in \mathbb{R}$, and by a number k of possible topics and a number V of possible words that can be found in the corpus of M documents. Each corpus D is then parametrized by $\alpha \in \mathbb{R}_{*+}^k$, and $\beta \in [0,1]^{k \times V}$, and the model assumes that each document $\boldsymbol{w_d}$ of the corpus has been generated as follows:

1. The size of the document is $N \sim \text{Poisson}(\xi)$
2. The topic mixture parameter $\theta \in [0,1]^k, \sum_{i=1}^k \theta_i = 1$ is drawn from $\text{Dir}(\alpha)$
3. For each $n \in [1, N]$:
   (a) The topic $z_n \in [0,1]^k$ of the n^{th} word is drawn from $\text{Multinomial}(\theta)$
   (b) The word w_n is drawn from $p(w_n|z_n, \beta)$ where $p(w_n^j = 1|z_n^i = 1, \beta) = p(w_k^j = 1|z_k^i = 1, \beta) = \beta_{ij} \quad \forall k \in [1, N]$

The probabilistic graphical model of the representation is shown in Figure 1. The authors then discuss the assumptions on which this model is built (especially the assumption of exchangeability), compare it with other latent variables models (such as mixture of unigrams), and provide a graphical interpretation.

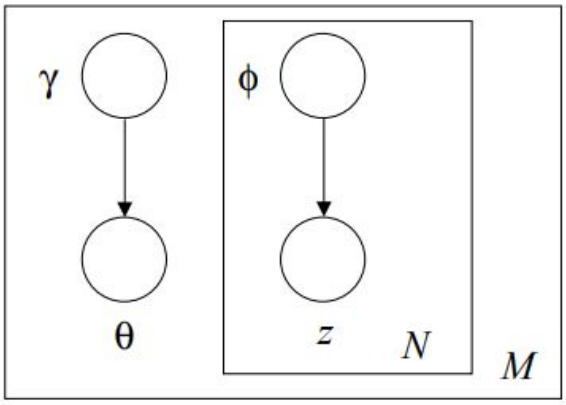

Figure 2: Graphical model representation of the variational distribution used to approximate the posterior in LDA

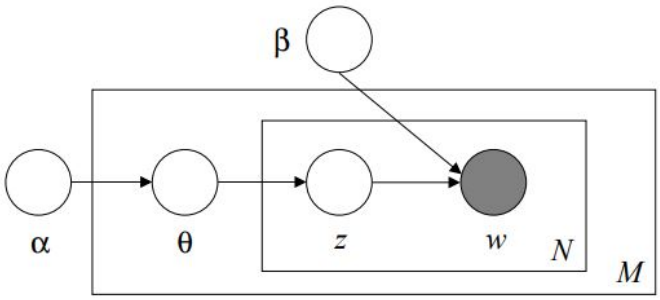

Figure 1: Probabilistic graphical model representation of the Latent Dirichlet Allocation model

2 Inference

Once we learned the parameters α and β of the document corpus, we can use the generative model to estimate the distributions of the hidden variables given a new document: $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$. However this distribution is intractable, the authors hence have to resort to variational inference to solve it. The idea behind variational inference is to find a lower bound of the log likelihood thanks to a more tractable distribution, shown in figure 2. The authors show that maximizing this lower bound with regard to the variational parameters γ and ϕ is equivalent to minimizing the Kullback-Leibler divergence between this approximate distribution and the actual distribution of the LDA. We then have a convex optimization problem in the parameters γ and ϕ, leading to an iterative algorithm to compute their approximated value.

3 Parameters estimation

Before being able to do inference on new documents, we have to train the model on our corpus of documents, by finding the parameters α and β that maximize the marginal log likelihood of the data $\sum_{d=1}^{M} log(p(\boldsymbol{w}_d|\alpha, \beta)$. But as stated above we have a lower bound on this likelihood, depending on α and β and on (approximate) variational distributions of parameters γ_d and ϕ_d (for each document $\boldsymbol{w}_d$). We can therefore use a variational EM algorithm to estimate α and β.

1. In the E-step we maximize our lower bound with regard to γ_d and ϕ_d for each $d \in [1, D]$ as presented in the *Inference* section.
2. In the M-step, we maximize the resulting lower bound with respect to α and β, for which we have a closed form depending on the γ_d and ϕ_d.

4 Details of computations

$$
\begin{aligned}
\ell(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \log \int_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \log \int_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} b(\boldsymbol{\theta}, \boldsymbol{z}) \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{b(\boldsymbol{\theta}, \boldsymbol{z})} d\boldsymbol{\theta} \\
&\geq \int_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} b(\boldsymbol{\theta}, \boldsymbol{z}) \log \left(\frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{b(\boldsymbol{\theta}, \boldsymbol{z})} \right) d\boldsymbol{\theta} \quad \text{thanks to Jensen's inequality} \\
&= \mathbb{E}_b \left[\ell_c(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right] - H(b(\boldsymbol{\theta}, \boldsymbol{z})) \\
&= \mathbb{E}_b \left[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \right] - \mathbb{E}_b \left[\log b(\boldsymbol{\theta}, \boldsymbol{z}) \right] \\
&= \mathcal{L}(b, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
b^{(t+1)} &= \arg\max_{b} \mathcal{L}(b, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}) = p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) && \text{(E-step)} \\
(\boldsymbol{\alpha}, \boldsymbol{\beta})^{(t+1)} &= \arg\max_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} \mathcal{L}(b^{(t+1)}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg\max_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} \mathbb{E}_b(\ell_c(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) && \text{(M-step)}
\end{aligned}
\tag{2}
$$

Which is equivalent to the following system:

$$
\begin{aligned}
b^{(t+1)} &= \arg\min_{b} \mathcal{D}(b(\boldsymbol{\theta}, \boldsymbol{z}) \mid\mid p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})) && \text{(E-step)} \\
(\boldsymbol{\alpha}, \boldsymbol{\beta})^{(t+1)} &= \arg\min_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} \mathcal{D}(b^{(t)}(\boldsymbol{\theta}, \boldsymbol{z}) \mid\mid p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) && \text{(M-step)}
\end{aligned}
\tag{3}
$$

Nevertheless the exact inference is not possible due to the untractable posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$. A relaxation of the problem is proposed by introduction of the varational distribution $q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})$:

$$
\begin{aligned}
(\boldsymbol{\gamma}, \boldsymbol{\phi})^{(t+1)} &= \arg\max_{(\boldsymbol{\gamma},\boldsymbol{\phi})} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \arg\max_{(\boldsymbol{\gamma},\boldsymbol{\phi})} \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}, \boldsymbol{z})\right] && \text{(E-step)} \\
(\boldsymbol{\alpha}, \boldsymbol{\beta})^{(t+1)} &= \arg\min_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \mathcal{D}(b^{(t)}(\boldsymbol{\theta}, \boldsymbol{z}) \mid\mid p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) && \text{(M-step)}
\end{aligned}
\tag{4}
$$

which leads to the following updates: