
Probabilistic Graphical Models IFT6269 Project

Latent Dirichlet Allocation

Patrice B  chard ^{* 1} Th  ophile Gervet ^{* 2}

Abstract

TODO

1. Introduction and Related Works

1.1. Latent Dirichlet Allocation

1.1.1. GENERATIVE PROCESS

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus of D documents $\{\mathbf{w}_1, \dots, \mathbf{w}_D\}$. Each document $\mathbf{w}_d = (w_{d,1}, \dots, w_{d,N_d})$ is a sequence of N_d words from a vocabulary indexed by $\{1, \dots, V\}$.

The basic idea behind LDA is that each document is represented as a random mixture over K latent topics, where each topic is characterized by a multinomial distribution over words, parametrized by a vector β_k , $k \in 1, \dots, K$.

Formally, LDA assumes the following generative process for each document \mathbf{w}_d :

1. Sample a distribution over topics $\theta_d \sim \text{Dir}(\alpha)$
2. For each of the N_d words $w_{d,n}$ independently:
 - (a) Draw a topic from the distribution over topics $z_{d,n} \sim \text{Mult}(\theta_d)$
 - (b) Draw a word from this topic $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

This generative process is illustrated in the directed graphical model of figure 1.

The joint distribution associated with this model is:

$$\prod_{k=1}^K p(\beta_k | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta_{1:K}) \quad (1)$$

^{*}Equal contribution ¹Universit   de Montr  al, Montr  al, Canada ²McGill University, Montr  al, Canada. Correspondence to: Patrice B  chard <patrice.bechard@umontreal.ca>, Th  ophile Gervet <theophile.gervet@umontreal.ca>.

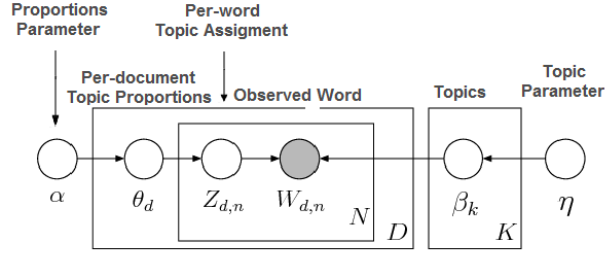


Figure 1. Graphical model corresponding to the generative process for LDA.

where

$$\begin{aligned} z_{d,n} &\sim \text{Mult}(\theta_d), \text{ i.e } p(z_{d,n} | \theta_d) = \theta_{d,z_{d,n}} \\ w_{d,n} &\sim \text{Mult}(\beta_{z_{d,n}}), \text{ i.e } p(w_{d,n} | z_{d,n}, \beta_{1:K}) = \beta_{z_{d,n}, w_{d,n}} \\ \beta_k &\sim \text{Dir}(\eta) \\ \theta_d &\sim \text{Dir}(\alpha) \end{aligned}$$

This generative process is just a model of the structure we assume to be in our data. In reality, we only observe documents $\mathbf{w}_{1:D}$ and our goal is to infer the underlying topic structure: the topics $\beta_{1:K}$, the per document distributions over topics θ_d , and per document per word topic assignments $z_{d,n}$.

Since the generative process assumes independence between documents, the probability of the whole corpus decomposes as a product of terms for individual documents. In order to clarify the notation, from now on let us consider equations for a single document \mathbf{w} with N words.

Assuming the Dirichlet hyper-parameters α and η are fixed, we want to infer the posterior distribution:

$$p(\beta, \theta, \mathbf{z} | \mathbf{w}, \alpha, \eta) = \frac{p(\beta, \theta, \mathbf{z}, \mathbf{w} | \alpha, \eta)}{p(\mathbf{w} | \alpha, \eta)} \quad (2)$$

Unfortunately, this distribution is intractable to compute because the normalization factor $p(\mathbf{w} | \alpha, \eta)$ cannot be computed exactly. We must use approximate inference to estimate the posterior over latent variables β , θ and \mathbf{z} . We will explore both variational inference and Gibbs sampling

to solve this problem.

1.2. Related Works

Most of the work done in this article is based on the original Latent Dirichlet Allocation paper (Blei et al., 2003), which covers the general theory of the model as well as the variational inference approach to estimate the posterior. Other works present the Gibbs sampling approach for the same problem and provide useful implementation details (Darling, 2011; Griffiths, 2002; Heinrich, 2005).

Latent Dirichlet Allocation is an extension of the probabilistic latent semantic analysis (pLSA) model (Hofmann, 1999), which postulates that a document \mathbf{w} and a word w_n are conditionally independent given a latent topic z :

$$p(\mathbf{w}, w_n) = p(\mathbf{w}) \sum_z p(w_n|z)p(z|\mathbf{w})$$

Some problems arise with this model. First, while it does make it possible for a document to contain multiple topics, the model only learns the topic mixtures $p(z|\mathbf{w})$ for the documents it has seen during the training, making it difficult to use it to assign probability to previously unseen documents. Furthermore, the number of parameters which the model tries to estimate grows linearly with the number of documents in the training set, which makes the model prone to overfitting. LDA overcome both of the main problems with pLSA by adding a Dirichlet prior over the documents, defining a complete generative model.

However, LDA is by no means perfect. There have been many new ways to enhance the model to make it more realistic by relaxing some assumptions made by the model. The bag-of-words assumption is fine for our use of the LDA, which is to uncover the topics present in a document and to assign a certain word to a topic. For more complicated goals, such as generating text, the assumption becomes inappropriate. There have been a number of extensions of the model that relaxes the bag-of-words assumption, for example, by combining the LDA with a Hidden Markov Model (HMM) to take into account the ordering of the words (Griffiths et al., 2005). It is also possible to relax the assumption that the ordering of the documents is irrelevant if we wish to find the evolution of certain topics over time, for example. A dynamic topic model would then be more appropriated than the normal LDA (Blei & Lafferty, 2006).

2. Methods

2.1. Variational EM

To make our life easier, we derive a variational EM algorithm for a slightly simpler graphical model for LDA (the

one presented in the original paper). We remove the Dirichlet prior over topics, this corresponds to removing the dependence of the β plate on the η random variable (see Figure 2, left). By doing so, we lose our ability to smooth topics through the η parameter.

Assuming the α parameter and topics β are fixed, the posterior for a single document \mathbf{w} now takes the form:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \quad (3)$$

This expression is still intractable because of the denominator $p(\mathbf{w}|\alpha, \beta)$.

Let us motivate a variational EM algorithm by looking at how this intractable posterior comes back to bite us when we try to maximize the likelihood of the observed data. Since the probability of the whole corpus decomposes as a product of terms for individual documents, the log likelihood of the whole corpus decomposes as a sum of terms for individual documents. The log likelihood of a single document \mathbf{w} takes the form:

$$\log p(\mathbf{w}|\alpha, \beta) = \log \int \sum_z p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) d\theta \quad (4)$$

Because of the integral over θ and the sum over \mathbf{z} inside the log, we cannot maximize this objective directly. One common way to solve this problem is the Expectation Maximization (EM) algorithm. We make the following observation:

$$\begin{aligned} \log p(\mathbf{w}|\alpha, \beta) &= \log \int \sum_z \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\ &= \mathbb{E}_q \left[\log \int \sum_z \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z})} d\theta \right] \\ &\geq \mathbb{E}_q [\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] + \mathbb{E}_q [\log q(\theta, \mathbf{z})] \\ &= \mathcal{L}(q, \alpha, \beta) \end{aligned}$$

where we have introduced an arbitrary distribution $q(\theta, \mathbf{z})$ over the problematic latent variables θ and \mathbf{z} , and the last step follows from Jensen's inequality.

The EM algorithm consists in alternatively maximizing the lower bound \mathcal{L} on the log likelihood with respect to the distribution q (the E-step), and with respect to parameters α and β (the M-step).

One can easily show that maximizing \mathcal{L} with respect to q is equivalent to minimizing KL ($q(\theta, \mathbf{z}) || p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$). A simple solution is thus to set $q(\theta, \mathbf{z})$ to be $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$, the true posterior over latent variables under the model.

In our case, things are not so simple: as we have seen, the posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ is intractable. This is where vari-

ational inference comes into play. The basic idea is to introduce a parametrized family of distributions over latent variables and phrase inference as an optimization problem.

We parametrize q as follows, according to the mean field independence assumption:

$$q(\boldsymbol{\theta}, \mathbf{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta} | \boldsymbol{\gamma}) \prod_{n=1}^N q(z_n | \phi_n) \quad (5)$$

where the dirichlet parameters $\boldsymbol{\gamma}$ and the multinomial parameters ϕ_n , $n \in \{1, \dots, N\}$ are free variational parameters. The graphical model corresponding to this parameterization is illustrated in Figure 2, right.

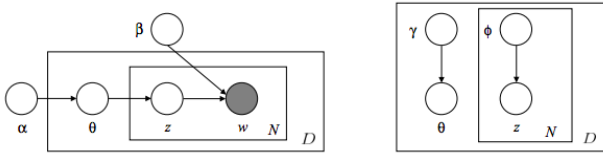


Figure 2. Graphical model corresponding to the generative process for LDA.

Now variational inference consists in solving the following optimization problem:

$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg \min_{\boldsymbol{\gamma}, \boldsymbol{\phi}} \text{KL}(q(\boldsymbol{\theta}, \mathbf{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) || p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \quad (6)$$

Note that $q(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \boldsymbol{\gamma}^*, \boldsymbol{\phi}^*)$ is actually a conditional distribution varying as a function of \mathbf{w} because the optimization objective depends on \mathbf{w} . It can thus be seen as an approximation to the posterior $p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

With this tool at our disposal, we can come back and complete the EM algorithm described in section 1.1.2. We can expand our variational lower bound on the per-document log likelihood as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \mathbb{E}_q[\log p(\boldsymbol{\theta} | \boldsymbol{\alpha})] + \mathbb{E}_q[\log p(\mathbf{z} | \boldsymbol{\theta})] + \\ &\quad \mathbb{E}_q[\log p(\mathbf{w} | \mathbf{z}, \boldsymbol{\beta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta} | \boldsymbol{\gamma})] - \mathbb{E}_q[\log q(\mathbf{z} | \boldsymbol{\phi})] \end{aligned} \quad (7)$$

where the first three terms correspond to the expected complete log likelihood under q and the last two terms to the entropy of q .

2.1.1. E-STEP

In the E-step, we maximize \mathcal{L} with respect to the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. This corresponds to applying the following update rules iteratively until convergence for each document:

$$\phi_{n,k} \propto \beta_{k,w_n} \exp\{\Psi(\gamma_k)\} \quad (8)$$

$$\gamma_k = \alpha_k + \sum_{n=1}^N \phi_{n,k} \quad (9)$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function.

Theses updates make sense intuitively. Equation 8 tells us that $\phi_{n,k}$, the probability that word w_n is generated by latent topic k , is proportional to the probability of drawing w_n from topic k times a quantity proportional to how much the document likes topic k (the dirichlet parameter γ_k). Equation 9 tells us that the dirichlet parameter γ_k is equal to the pseudo-count α_k given by the dirichlet prior plus the expected count of words picking topic k .

2.1.2. M-STEP

For the M-step, we maximize the overall variational lower bound with respect to parameter $\boldsymbol{\beta}$. We could also maximize with respect to $\boldsymbol{\alpha}$ but we chose to keep it fixed to make things simpler. As said before, the overall log likelihood of the corpus is the sum of the log likelihoods for the individual documents. Likewise, the overall variational lower bound is the sum of the individual variational bounds.

Maximizing with respect to $\boldsymbol{\beta}$ gives the following update:

$$\beta_{k,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{d,n,k} w_{d,n}^{(j)} \quad (10)$$

where we abuse notation and index word $w_{d,n}$ which was a scalar until here as a one hot vector.

Equation 10 also has an intuitive meaning: the probability of word j under topic k is proportional to the expected count under the variational distribution of the number of times word j is assigned topic k across the whole corpus.

2.1.3. OVERALL ALGORITHM

We implemented a variational "EM flavored" algorithm which we described in detail in algorithm 1. Note that to be strictly an EM algorithm, we should maximize \mathcal{L} with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ all the way at each E-step. Instead, we alternate a single step of improvement with respect to the variational parameters with one step of improvement with respect to $\boldsymbol{\beta}$. Our algorithm has the same guarantee to increase the variational lower bound at each iteration, it is just more convenient to check for convergence only for the overall algorithm instead of for each E-step.

2.2. Gibbs Sampling

Once again, we consider a single document \mathbf{w} with N words. Recall that we are interested in the latent topics $\beta_{1:K}$, the distribution over topics $\boldsymbol{\theta}$, and the per word topic assignments z_n , $n \in \{1, \dots, N\}$.

Algorithm 1 Variational EM**Input:** documents $\mathbf{w}_1, \dots, \mathbf{w}_D$ Set $\alpha_k = 1/K$ for all k Randomly initialize variational parameters γ, ϕ **repeat** **for** $d = 1$ **to** D **do** **for** $k = 1$ **to** K **do** **for** $n = 1$ **to** N_d **do** $\phi_{d,n,k} = \beta_{k,w_n} \exp\{\Psi(\gamma_{d,k})\}$ **end for** normalize $\phi_{d,n}$ to sum to 1 $\gamma_{d,k} = \alpha_k + \sum_{n=1}^{N_d} \phi_{d,n,k}$ **end for** **end for** **for** $k = 1$ **to** K **do** **for** $j = 1$ **to** V **do** $\beta_{k,j} = \sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{d,n,k} w_{d,n}^{(j)}$ **end for** normalize β_k to sum to 1 **end for****until** objective \mathcal{L} doesn't improve anymore

We could derive conditional distributions for each of these latent variables, and therefore an LDA Gibbs sampling algorithm, but we note that both θ_d and β_k can be computed using the topic assignments $z_{d,n}$ which are sufficient statistics. This allows us to integrate out the multinomial parameters and simply sample $z_{d,n}$. This strategy is referred to as *collapsed* or *Rao-Blackwellized* Gibbs sampling (Heinrich, 2005).

The conditional we want to sample the topic assignment z_n from is given by :

$$p(z_n | \mathbf{z}_{-n}, \alpha, \eta, \mathbf{w})$$

where \mathbf{z}_{-n} denotes all topic allocations except for z_n and where we have reintroduced the hyperparameter α controlling the topics-per-document proportions as well as the hyperparameter η controlling the probability of a word being assigned to a topic. We can develop this expression using Bayes' rule:

$$p(z_n | \mathbf{z}_{-n}, \mathbf{w}, \alpha, \eta) \propto p(w_n | z_n, \mathbf{z}_{-n}, \mathbf{w}_{-n}, \eta) p(z_n | \mathbf{z}_{-n}, \alpha) \quad (11)$$

where we have used the fact that $\mathbf{w} \perp \alpha | \mathbf{z}$ and $\mathbf{z} \perp \eta$. Now, let's develop each term of equation 11 for a given topic assignment $z_n = k$:

$$\begin{aligned} p(w_n | z_n = k, \mathbf{z}_{-n}, \mathbf{w}_{-n}) &= \int p(w_n | z_n = k, \beta_k) p(\beta_k | \mathbf{z}_{-n}, \mathbf{w}_{-n}) d\beta_k \\ &= \frac{n_{-n,k}^{(w_n)} + \eta_n}{\sum_{n'=1}^V (n_{-n,k}^{(w_{n'})} + \eta_{n'})} \end{aligned} \quad (12)$$

$$\begin{aligned} p(z_n = k | \mathbf{z}_{-n}) &= \int p(z_n = k | \theta_{\mathbf{w}_n}) p(\theta_{\mathbf{w}_n} | \mathbf{z}_{-n}) d\theta_{\mathbf{w}_n} \\ &= \frac{n_{-n,k}^{(\mathbf{w}_n)} + \alpha_k}{\sum_{k'=1}^K (n_{-n,k'}^{(\mathbf{w}_n)} + \alpha_{k'})} \end{aligned} \quad (13)$$

The results obtained follow a Dirichlet $(n_{-n,k}^{(w_n)} + \eta_n)$ and a Dirichlet $(n_{-n,k}^{(\mathbf{w}_n)} + \alpha_k)$, respectively. These distributions were obtained by the integration of the product between a Multinomial distribution and its Dirichlet conjugate prior. A detailed derivation of these equations is presented in (Heinrich, 2005).

Putting together equations 12 and 13, we obtain the posterior distribution from which we sample a new topic assignment z_n :

$$\begin{aligned} p(z_n = k | \mathbf{z}_{-n}, \mathbf{w}) &\propto \frac{n_{-n,k}^{(w_n)} + \eta_n}{\sum_{n'=1}^V (n_{-n,k}^{(w_{n'})} + \eta_{n'})} \frac{n_{-n,k}^{(\mathbf{w}_n)} + \alpha_k}{\sum_{k'=1}^K (n_{-n,k'}^{(\mathbf{w}_n)} + \alpha_{k'})} \end{aligned} \quad (14)$$

Finally, after the *burn-in* period of the Gibbs sampling, we can find the expressions for the parameters we are looking for, namely the term-by-topic distribution β and the topic-by-document distribution θ as such:

$$\beta_{k,n} = \frac{n_k^{(w_n)} + \eta_t}{\sum_{t=1}^V n_k^{(t)} + \eta_t} \quad (15)$$

$$\theta_{d,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^K n_m^{(k)} + \alpha_k} \quad (16)$$

Algorithm 2 presents the various steps of the implementation of LDA using collapsed Gibbs sampling to estimate the posterior distribution.

Algorithm 2 Gibbs Sampling

Input: documents $\mathbf{w}_1, \dots, \mathbf{w}_D$
 Initialize hyperparameters α and η
 Randomly initialize topic assignments $z_{d,n}$ for all d, n
 Increment counters $n_{d,k}$ and $n_{k,n}$
repeat
 for $d = 1$ **to** D **do**
 for $n = 1$ **to** N_d **do**
 $n_{d,z_{d,n}} -= 1$; $n_{z_{d,n},n} -= 1$
 Sample topic assignment $z_{d,n}$ from equation 14
 $n_{d,z_{d,n}} += 1$; $n_{z_{d,n},n} += 1$
 end for
 end for
until convergence
 compute $\beta_{k,n}$ from equation 15 for all d, k
 compute $\theta_{d,k}$ from equation 16 for all k, n

2.2.1. CHOICE OF HYPERPARAMETERS

Both α and η are hyperparameters for our model. Typically, the a priori assumption of LDA is that all words are equally likely to be assigned to a topic, and all topics have equal chance of being assigned to a document. In this case, values of $50/K$ for each component of α and 0.01 for each component of η have been reported to yield good results (Griffiths & Steyvers, 2004).

3. Experiments

The experiments conducted with both implementations of the latent Dirichlet allocation, one using variational inference and the other using Gibbs sampling to estimate the posterior distribution of our model (equation 2), used data taken from the 20 Newsgroups dataset (Rennie, 2008).

3.1. Querying

TODO

3.2. Perplexity

TODO

4. Results

TODO

References

- Blei, David M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- Blei, David M and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd international conference*

on Machine learning, pp. 113–120. ACM, 2006.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Darling, William M. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011.

Griffiths, Thomas L and Steyvers, Mark. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Griffiths, Thomas L, Steyvers, Mark, Blei, David M, and Tenenbaum, Joshua B. Integrating topics and syntax. In *Advances in neural information processing systems*, pp. 537–544, 2005.

Griffiths, Tom. Gibbs sampling in the generative model of latent dirichlet allocation. Technical report, Department of Psychology, Stanford University, Stanford, CA, 2002.

Heinrich, Gregor. Parameter estimation for text analysis. Technical report, vsonix GmbH + University of Leipzig, Leipzig, Germany, 2005.

Hofmann, Thomas. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

Rennie, Jason. 20 newsgroups, 2008. URL <http://qwone.com/~jason/20Newsgroups/>.