

# Implementing Latent Dirichlet Allocation with Stochastic Variational Inference

Kevin Song

December 5, 2016

## 1 Introduction

When analyzing a large body of text, we are very rarely interested in what happens at the level of individual words or sentences. Instead, we want to learn some higher-level information about the body of documents. For example, we may want to know what the prevailing opinion is on a certain topic (e.g. analyzing legal documents), we may want to extract some data held in the documents (e.g. analyzing medical literature for treatment outcomes), or, in some cases, when we are handed a stack of documents with no further information, we might just want to know what the documents are talking about.

Topic models, introduced around the turn of the millennium, are a mechanism for solving the last problem: without any additional information about the corpus, aside from maybe the number of topics expected, it can determine which words belong to which topic, and what topic (or set of topics) each document exhibits. Examples of topic models include probabilistic latent semantic indexing (pLSI) and latent dirichlet allocation (LDA).

In this project, I write a small-scale LDA inference program using stochastic variational inference (SVI). This program is applied to a set of test documents which are generated by the rules of the LDA model (explained below). Unfortunately, due to time constraints, I was not able to optimize my code well enough to attempt to feed all of Wikipedia as input, as I had originally planned. Therefore, to provide a reference point for how well this algorithm can potentially work, I provide results from an online Github repository that implements a very limited variant of SVI designed specifically for LDA.

## 2 Methods Overview

A note on notation: the papers I'm working out of (cited at the end of this document) have horrible notational issues.

### 2.1 Topic Models and LDA

Topic models tend to assume a generative process, meaning the words in a document are drawn at random according to certain rules. For example, we may assume that there are two topics, that each document exhibits exactly one topic, and the words are drawn from that topic distribution. Solving the topic model involves figuring out what the parameters of the model are, in this case, what the word distribution for each topic is, and which (single) topic each document has.

The challenge is to create a model which is strong enough to make interesting predictions about a document corpus, but can still be solved. To illustrate this, I briefly explore unigram models, the pLSI model, and the LDA model.

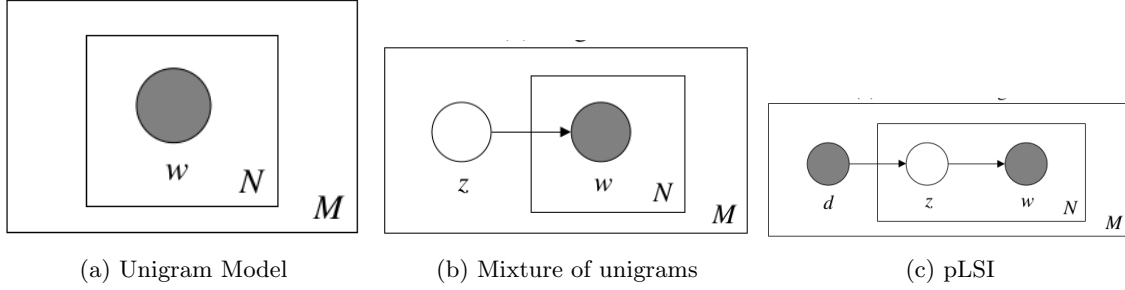


Figure 1: Graphical representations for topic models

### 2.1.1 Unigram Model

In the unigram model, there is one global distribution of words, and all words in every document are assumed to be sampled from this distribution. This is not a particularly interesting model, since it essentially means that every document should look the same.

Mathematically, this manifests as the following conditional for the words:

$$p(w) = \prod_{n=1}^N p(w_n)$$

We can graphically represent this as simply observing words coming from a distribution, as seen in Figure 1(a).

### 2.1.2 Mixture of Unigrams

In the mixture of unigrams, each document has its own multinomial distribution. These distributions do not represent topics—they are just a document-specific distribution of words. The generative process draws a parameter  $z$  for each document, and then words are drawn from a multinomial conditioned on  $z$ . Mathematically, this is represented by

$$p(w) = \sum_z p(z) \prod_{n=1}^N p(w_n|z)$$

Graphically, this is represented by the process shown in Figure 1(b)

### 2.1.3 Probabilistic Latent Semantic Indexing

pLSI extends the mixture of unigrams model further, by postulating that a document  $d$ 's words are independent of the document index given an unobserved topic,  $z$ . It is represented graphically by Figure 1(c).

### 2.1.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation extends all of the previous models by assuming that there is some fixed global parameter which controls the topic distributions of the documents. There is then a per-topic probability of a word showing up in a document, which is modified by another global parameter. Graphically, it is represented by Figure 2.

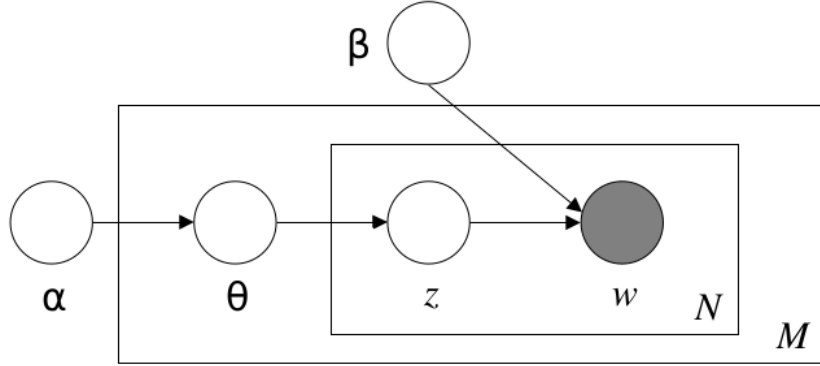


Figure 2: Graphical Model for LDA

Algorithmically, the generative process is assumed to draw topic distributions from a Dirichlet distribution. Then, for each individual document in the corpus, the following happens:

1. Choose the document length  $N$ , either as a fixed parameter or from a Poisson distribution.
2. Choose a document topic composition  $\theta \sim \text{Dir}(\alpha)$
3. For each of the  $N$  words:
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta, 1)$
  - (b) Choose a word  $w_n$  from  $p(w_n | z_n, \beta_{z_n})$

As can be seen from Figure 2, this model is significantly more complicated than any of its predecessors. What it loses in simplicity, however, it more than makes up for in power. On top of being able to account for a probabilistic distribution of words in a topic, LDA is additionally posits that a document can contain multiple topics, and that there is some global distribution of topics (controlled by  $\alpha$ ) which the topic mixtures come from.

The question that arises now is this: can we efficiently solve this model? For a long time, the only available solution method was batch variational inference, which is very slow when the number of documents and topics is large. Instead, a more recent invention, stochastic variational inference, can be used to quickly calculate the posterior distribution.

## 2.2 Stochastic Variational Inference

Variational inference is one technique that can be used when a posterior distribution is intractable to compute explicitly. By introducing some parametrized distribution over the hidden variables and minimizing the difference between the variational distribution and the posterior, we can approximate the posterior distribution.

Variational inference minimizes the evidence lower bound (ELBO), which is equal to the negative Kullback-Leibler divergence plus an additive constant.

To derive the ELBO, we assume we have a probability distribution over all variable in the model (in this example case,  $x, z, \beta$ ), and introduce a distribution over some hidden variables,  $q(z, \beta)$ .

$$\begin{aligned}
\log p(x) &= \log \int p(x, z, \beta) dz d\beta \\
&= \log \int p(x, z, \beta) \frac{q(z, \beta)}{q(z, \beta)} dz d\beta \\
&= \log \left( \mathbb{E}_q \left[ \frac{p(x, z, \beta)}{q(z, \beta)} \right] \right) \\
&\geq \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \\
&\triangleq L(q)
\end{aligned}$$

By making appropriate choices for the forms of  $q$  and modifying the gradient of the PDF by the Fisher matrix (which results in the so-called natural gradient), we can arrive at the general algorithm for stochastic variational inference in Appendix A.

The specific variational inference algorithm for LDA is provided in Listing 1. For ease of understanding, some of the types and sizes of parameters have been added to the algorithm. The numbers used for dimensions are as follows:  $K$  is the number of topics,  $N$  is the number of distinct words in the document corpus,  $D$  is the number of documents, and  $M$  is the number of words in a given document. Superscripts in parenthesis are used to refer to the time series, while superscripts without parenthesis are used to refer to a document ID, e.g.  $w_n^d$  is the  $n$ th word in the  $d$ th document. This differs slightly from the original notation because, quite frankly, I wouldn't wish the original notation on my enemies.

In addition, words are represented indicator vectors, that is  $w_n^d$  is a vector with  $N$  entries, all but one of which are zero. When used as a subscript, it refers to the nonzero index in the indicator vector, instead of the full vector.

If a matrix-valued variable is subscripted with a single subscript, it refers to that column of the matrix, e.g. if  $X$  is a matrix,  $X_i$  is the  $i$ th column of  $X$ .

Alpha and eta ( $\alpha, \eta$ ) are hyperparameters provided by the user. A sane value is somewhere between 0.1 and 1.

Finally, the variable  $\Psi$  is used to refer to the digamma function, which is equal to the derivative of the log of the gamma function:

$$\Psi(x) = \frac{d}{dx} \log(\Gamma(x))$$

```

1 Declare  $\lambda$ : a matrix of size  $k \times n$ 
2 Set  $\rho^{(0)}$  in  $[0.5, 1)$ 
3 while  $\lambda$  not converged:
4   while  $\phi$  and  $\gamma$  not converged:
5     Sample a document indexed by  $d$  randomly from the corpus
6     Declare  $\gamma^d$ : a vector with  $K$  elements
7     Declare  $\phi^d$ : a matrix of size  $k \times M$ 
8
9     Initialize  $\gamma_k^d = 1$  for  $k \in \{1, \dots, K\}$ 
10    For  $m \in \{1, \dots, M\}$ : (words in document)
11      For  $k \in \{1, \dots, K\}$ : (topics)
12        Set  $\phi_{m,k}^d \propto \left( \Psi(\gamma_k^d) - \sum_{j=1}^K \Psi(\gamma_j^d) \right) - \left( \Psi(\lambda_{k,w_n^d}) - \sum_{j=1}^M \Psi(\lambda_{k,j}) \right)$ 
13
14    Set  $\gamma^d = \alpha + \sum_m \phi_m^d$ 
15  endwhile
16
17  For  $k \in \{1, \dots, K\}$ :
18    Set intermediate topics  $\hat{\lambda}_k = \eta + D \sum_{m=1}^M \phi_{mk}^d w_{dm}$ 
19  endfor

```

```

21 |   Update  $\lambda^{(t)} = (1 - \rho^{(t)})\lambda^{(t-1)} + \rho^{(t)}\hat{\lambda}$ 
    |   Update  $\rho^{(t)}$  by some rule, e.g. Robbins Munro or Adagrad
23 | endwhile

```

Listing 1: The SVI-LDA algorithm

The per-document parameters  $\gamma$  and  $\phi$  can either be cached during computation and reused, or the global parameter  $\lambda$  can be used to reconstruct them once it is found, using a procedure similar to that between lines 5 and 15 in Listing 1.

### 3 Results

## 4 Appendix A: General Stochastic Variational Inference

This algorithm is taken near-verbatim out of Hoffman et. al., and so uses some of their curious notational conventions.

```
1 Initialize  $\lambda^{(0)}$  randomly
   Set  $\rho_t$ 
3 repeat until converged:
   Sample a data point  $x_i$  from the data set
5   Compute its local variational parameter:
        $\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})]$ 
7
   Computer intermediate global parameters:
        $\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})]$ 
9
11  Update estimates of lambda:
        $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}$ 
13
15  Update  $\rho_t$ , e.g. using a Robbins Munro schedule
end
```