Inference in Dirichlet Process Mixtures with Applications to Text Document Clustering

Alberto Bietti
alberto.bietti@gmail.com

Lénaïc Chizat
lenaic.csl@gmail.com

January, 2014

Abstract

Mixture models are widely used to represent data as coming from separate components, but choosing the right number of mixture components can be a hard task. The Dirichlet process (DP) [9] has been introduced by Ferguson (1973) [4] as a convenient prior on probability distributions for nonparametric Bayesian problems, and has become popular in the machine learning community for dealing with Bayesian models with a potentially unbounded number of parameters. A key example is the Dirichlet process mixture model, which extends standard finite mixture models to an infinite number of mixture components, and which has been made practical thanks to the development of suitable Markov chain Monte Carlo (MCMC) and variational inference methods. We provide some background on the Dirichlet process and DP mixtures, and describe algorithms for variational inference and Gibbs sampling, showing the benefits of the variational method for this class of models, based on [2]. We apply the algorithms to text document modeling by deriving versions of the algorithms for DP mixtures with multinomial mixture components, and present results of our experiments.

1 Introduction

The Dirichlet process (DP) was introduced by Ferguson (1973) [4] as a prior on probability distributions (a measure on measures) for dealing with nonparametric problems in the Bayesian setting. The DP is discrete with probability one and can be written as an infinite sum of atomic distributions in what's called the stick-breaking representation [8], which makes it very suitable for defining infinite mixture models. Thanks to this computationally tractable representation, the DP has gained popularity in machine learning for dealing with nonparametric Bayesian models such as DP mixture models [7], and other extensions like the Hierarchical Dirichlet Process [10].

MCMC sampling algorithms have been extensively applied to Bayesian inference problems thanks to their flexibility, and they have been successfully developed for Dirichlet process mixture models, mainly through collapsed Gibbs sampling or blocked Gibbs sampling [5]. Later, Blei and Jordan [2] have developed a variational inference algorithm based on the stick-breaking representation of the DP. Variational inference generally performs slightly worse than sampling methods due to an approximation bias, but runs faster. However, Blei and Jordan show that for DP mixtures the variational algorithm gives almost as good results in the case of DP mixtures, in addition to being faster than sampling methods, and is particularly well suited for large scale problems.

We start by providing some theoretical background on the DP and DP mixture models. In section 4, we derive variational and Gibbs sampling-based inference algorithms for DP mixtures in the case where the distributions on the parameters are in the exponential family. In section 5, we define a DP mixture model for modeling text documents from a corpus as a DP mixture of multinomial distributions. We then derive the appropriate variational inference and Gibbs sampling algorithms and show experimental results on a corpus of news articles from the Associated Press.

2 The Dirichlet Process

Let G_0 be a probability distribution on Θ, and $\alpha > 0$. A random measure G is distributed according to a Dirichlet process [4] with base measure G_0 and concentration parameter α, written $G \sim DP(\alpha, G_0)$, if for every partition $A_1, \ldots, A_k$ of Θ,

$$(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)), \tag{1}$$

Let $G \sim DP(\alpha, G_0)$ be such a random measure and $(A_1, \ldots, A_k)$ a partition of Θ. If $\eta \sim G$, then the vector $(\mathbb{I}_{A_1}(\eta), \ldots, \mathbb{I}_{A_k}(\eta)) = (\delta_\eta(A_1), \ldots, \delta_\eta(A_k))$ (where $\mathbb{I}$ is the indicator function and δ denotes the Dirac delta measure) is distributed according to a multinomial with parameter $(G(A_1), \ldots, G(A_k))$. If $\eta_1, \ldots, \eta_n$ are independent samples from G, because of the conjugacy between the Dirichlet and the multinomial distributions, we have:

$$(G(A_1), \ldots, G(A_k))|\eta_1, \ldots, \eta_n \sim Dir(\alpha G_0(A_1) + \sum_{i=1}^{n} \delta_{\eta_i}(A_1), \ldots, \alpha G_0(A_k) + \sum_{i=1}^{n} \delta_{\eta_i}(A_k)). \tag{2}$$

Since this is true for all partitions, the posterior distribution on G is also a DP, and takes the form:

$$G|\eta_1, \ldots, \eta_n \sim DP(\alpha + n, \frac{1}{\alpha + n}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\eta_i})) \tag{3}$$

Pólya urn scheme and Chinese Restaurant Process

The predictive distribution of a new observation η_{n+1} is obtained by integrating out the random measure G, and is given by the base measure of the posterior:

$$\eta_{n+1}|\eta_1, \ldots, \eta_n \sim \frac{1}{\alpha + n}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\eta_i}). \tag{4}$$

There are two analogies which are useful in describing the generation process of new samples. The *Pólya urn scheme* [1] considers each distinct $\eta_i \in \Theta$ as a different ball color, and puts previously seen balls in an urn. When a new ball is chosen, its color is picked according to the following rule: with probability $\frac{\alpha}{\alpha+n}$, pick a new color according to G_0 (the first ball color is always chosen this way), and with probability $\frac{n}{\alpha+n}$, choose the color of a random ball in the urn. The ball is then painted with the picked color and put in the urn. In the *Chinese Restaurant Process* (CRP) metaphor, we denote by $\{\eta_1^*, \ldots, \eta_K^*\}$ the distinct values of η_i (the tables of customers in a chinese restaurant), and by $\{n_1, \ldots, n_K\}$ their frequencies $n_i = |\{j : \eta_j = \eta_i^*\}|$ (the number of customers at each table). The predictive distribution becomes:

$$\eta_{n+1}|\eta_1, \ldots, \eta_n \sim \frac{1}{\alpha + n}(\alpha G_0 + \sum_{i=1}^{K} n_i \delta_{\eta_i^*}), \tag{5}$$

and the metaphor is the following: when a new customers enters the restaurant, he picks an existing table with probability proportional to the number of customers at that table, or a new table with probability proportional to α.

Stick-breaking representation

As suggested by the previous analogies, the values taken by the η_i will be discrete, and Sethuraman [8] has shown that the DP can be defined with an explicit discrete representation, given by the *stick-breaking construction*, which can be a useful alternative to the characterization given by 1. The construction of $G \sim DP(\alpha, G_0)$ is as follows: let $V_i \sim Beta(1, \alpha)$ and $\eta_i^* \sim G_0$ for $i = 1, 2, \ldots$, G is given by the stick-breaking representation:

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j) \tag{6}$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v})\delta_{\eta_i^*}, \tag{7}$$

where the *stick lengths* $\pi_i(\mathbf{v})$ are given by repeatedly breaking a stick of initial length 1 at points given by the v_i. It is clear from this representation that G can be used to represent infinite mixtures, where the mixture components are the atoms η_i^* of G, and the mixture proportions are given by $\pi_i(\mathbf{v})$.

3 Dirichlet Process Mixtures

A typical finite Bayesian mixture model can be described by the following generative process:

$$\begin{aligned} \pi|\alpha &\sim Dir(\alpha) \\ \eta_k^*|G_0 &\sim G_0, \quad k = 1, \ldots, K \\ Z_n|\pi &\sim Mult(\pi) \\ X_n|z_n &\sim p(\cdot|\eta_{z_n}^*), \end{aligned}$$

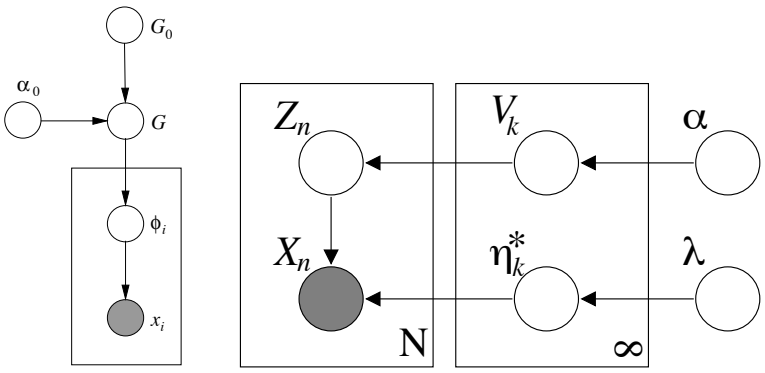
where α is the parameter for a Dirichlet prior on π and G_0 is the prior on η_k. Alternatively, if we write $G = \sum_{k=1}^{K} \pi_k \delta_{\eta_k^*}$, the generation of each data point X_n can be written as:

$$\begin{aligned} \eta_n|G &\sim G \\ X_n|\eta_n &\sim p(\cdot|\eta_n). \end{aligned}$$

This leads to a natural generalization to an infinite mixture model, by changing the prior on G to be a DP:

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \eta_n|G &\sim G \\ X_n|\eta_n &\sim p(\cdot|\eta_n). \end{aligned}$$

This is called a *Dirichlet process mixture model*, and is represented as a graphical model in Figure 1a. Equivalently, the DP mixture model can be modeled using the stick-breaking representation of the DP, leading to the following generative process, illustrated in Figure 1b:



(a) (b)

Figure 1: Two different graphical model representations for the Dirichlet process mixture model. The one on the right is based on the stick-breaking representation. Source: [2].

1. For $i = 1, 2, \ldots$, draw $V_i|\alpha \sim Beta(1, \alpha)$
2. For $i = 1, 2, \ldots$, draw $\eta_i^*|G_0 \sim G_0$
3. For every data point n:
   (a) Choose $Z_n|\mathbf{v} \sim Mult(\pi(\mathbf{v}))$
   (b) Draw $X_n|z_n \sim p(\cdot|\eta_{z_n}^*)$

We will limit ourselves to exponential family distributions for observed data, and the base measure of the DP will be the conjugate prior. Thus, we consider $p(x_n|z_n = i) = h(x_n) \exp(\eta_i^{*\top} x_n - a(\eta_i^*))$, where a is the log-partition function, so that:

$$p(x_n|z_n, \eta_1^*, \eta_2^*, \ldots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp(\eta_i^{*\top} x_n - a(\eta_i^*))\right)^{\mathbb{I}(z_n=i)}. \tag{8}$$

We take G_0 to be in the corresponding conjugate family: $p(\eta^*|\lambda) = h(\eta^*) \exp(\lambda_1^\top \eta^* - \lambda_2 a(\eta^*) - a(\lambda))$, where the sufficient statistics are given by the vector $(\eta^{*\top}, -a(\eta^*))^\top$, and $\lambda = (\lambda_1^\top, \lambda_2)^\top$.

4 Inference in DP mixtures

In order to perform exact inference in the DP mixture, one would need to compute the full posterior of the latent parameters $\mathbf{W} = \{(V_i)_i, (\eta_i^*)_i, (Z_n)_n\}$ given the observations $\mathbf{x} = (x_n)_n$. This typically requires computing the marginal likelihood $p(\mathbf{x}|\alpha, \lambda)$, since

$$p(\mathbf{W}|\mathbf{x}, \alpha, \lambda) = \frac{p(\mathbf{W}, \mathbf{x}|\alpha, \lambda)}{p(\mathbf{x}|\alpha, \lambda)}, \tag{9}$$

and the marginal likelihood is intractable to compute since it requires a complicated integral on the parameters $\mathbf{w}$.

This problem can be circumvented by approximating the posterior, which is usually done in two possible ways. The first is to use MCMC sampling algorithms to construct a Markov chain whose stationary distribution is the posterior, so that samples from this chain will eventually be distributed according to the posterior and will permit to approximated it. The other way is to use variational inference [6, 11], which defines a simpler distribution q on the latent variables $\mathbf{W}$, and optimizes its parameters so as to minimize the KL divergence between q and the targeted posterior. We will look at how to apply these methods to DP mixture models, based on [2].

4.1 Variational Inference

The goal of variational inference [6, 11] is to approximate an intractable target distribution p (the posterior in our case) with a simpler distribution q_ν, called the *variational distribution*, by minimizing the KL divergence $D(q_\nu||p)$ between q_ν and p with respect to the *variational parameters* ν. This turns the problem of approximating the posterior into an optimization problem. The KL divergence we want to minimize takes the following form:

$$D(q_\nu||p(\cdot|\mathbf{x},\alpha,\lambda)) = \mathbb{E}_q[\log q_\nu(\mathbf{W})] - \mathbb{E}_q[\log p(\mathbf{W},\mathbf{x}|\alpha,\lambda)] + \log p(\mathbf{x}|\mathbf{w},\alpha,\lambda). \tag{10}$$

Since $D(q_\nu||p(\cdot|\mathbf{x},\alpha,\lambda))$ is non-negative, an alternative formulation is that of maximizing the right hand side of the following equation, which is a lower bound on the marginal log likelihood (sometimes called *evidence lower bound*):

$$\log p(\mathbf{x}|\alpha,\lambda) \geq \mathbb{E}_q[\log p(\mathbf{W},\mathbf{x}|\alpha,\lambda)] - \mathbb{E}_q[\log q_\nu(\mathbf{W})]. \tag{11}$$

For our DP mixture model, this bound becomes:

$$\begin{aligned}\log p(\mathbf{x}|\alpha,\lambda) \geq\ & \mathbb{E}_q[\log p(\mathbf{V}|\alpha)] + \mathbb{E}_q[\log p(\boldsymbol{\eta}^*|\lambda)] \\ & + \sum_{n=1}^{N}(\mathbb{E}_q[\log p(Z_n|\mathbf{V})] + \mathbb{E}_q[\log p(x_n|Z_n)]) \\ & - \mathbb{E}_q[\log q_\nu(\mathbf{V},\boldsymbol{\eta}^*,\mathbf{Z})]\end{aligned} \tag{12}$$

In the mean-field variational inference setting, the variational distribution is taken to be fully factorized, as a product of distributions on each parameter. We consider these distributions to be in the exponential family, and of the same nature as the corresponding conditional distribution in the full model. To allow this representation, we limit ourselves to a truncated stick-breaking representation by fixing T and constraining $q(v_T = 1) = 1$, so that the mixture proportions $\{\pi_t(\mathbf{v})\}_{t>T}$ are equal to zero with probability one. The variational distribution takes the form:

$$q(\mathbf{v},\boldsymbol{\eta}^*,\mathbf{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n), \tag{13}$$

where $q_{\gamma_t}(v_t)$ are beta distributions, $q_{\tau_t}(\eta_t^*)$ are exponential family distributions with natural parameter τ_t and $q_{\phi_n}(z_n)$ are multinomials. Note that even though we consider a truncated variational distribution, there is no such constraint on the distribution of the full model, hence the algorithm will still try to approximate the full, infinite stick-breaking representation.

We now derive expressions for the terms in the lower bound on the log marginal likelihood from Equation 12:

$$
\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{V}|\alpha)] &= \textstyle\sum_{t=1}^{T-1}(\alpha-1)\mathbb{E}_q[\log(1-V_t)] - (T-1)(\log\Gamma(\alpha) - \log\Gamma(1+\alpha)) \\
\mathbb{E}_q[\log p(\boldsymbol{\eta}^*|\lambda)] &= \textstyle\sum_{t=1}^{T}(\log h(\eta_t^*) + \lambda_1^\top \mathbb{E}_q[\eta_t^*] - \lambda_2 \mathbb{E}_q[a(\eta_t^*)]) - Ta(\lambda) \\
\mathbb{E}_q[\log p(Z_n|\mathbf{V})] &= \mathbb{E}_q\left[\log\left(\prod_{i=1}^{\infty} V_i^{\mathbb{I}(Z_n=i)}(1-V_i)^{\mathbb{I}(Z_n>1)}\right)\right] \\
&= \textstyle\sum_{i=1}^{T} q(z_n=t)\mathbb{E}_q[\log V_t] + q(z_n>t)\mathbb{E}_q[\log(1-V_t)] \\
\mathbb{E}_q[\log p(x_n|Z_n)] &= \textstyle\sum_{t=1}^{T} q(z_n=t)\left(\log h(x_n) + \mathbb{E}_q[\eta_t^*]^\top x_n - a(\eta_t^*)\right) \\
\mathbb{E}_q[\log q_{\gamma_t}(V_t)] &= (\gamma_{t,1}-1)\mathbb{E}_q[\log V_t] + (\gamma_{t,2}-1)\mathbb{E}_q[\log(1-V_t)] - (\log\Gamma(\gamma_{t,1}) + \log\Gamma(\gamma_{t,2}) - \log\Gamma(\gamma_{t,1}+\gamma_{t,2})) \\
\mathbb{E}_q[\log q_{\tau_t}(\eta_t^*)] &= \log h(\eta_t^*) + \tau_{t,1}^\top \mathbb{E}_q[\eta_t^*] - \tau_{t,2}\mathbb{E}_q[a(\eta_t^*)] - a(\tau_t) \\
\mathbb{E}_q[\log q_{\phi_n}(Z_n)] &= \phi_n^\top \log \phi_n.
\end{aligned}
$$

The summations in $\mathbb{E}_q[\log p(Z_n|\mathbf{V})]$ and $\mathbb{E}_q[\log p(x_n|Z_n)])$ can be truncated at $t = T$ since $q(z_n = t) = q(z_n > t) = 0$ for $t > T$. We have

$$
\begin{aligned}
q(z_n = t) &= \phi_{n,t} \\
q(z_n > t) &= \textstyle\sum_{i=t+i}^{T}\phi_{n,i} \\
\mathbb{E}_q[\log V_t] &= \Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1}+\gamma t, 2) \\
\mathbb{E}_q[\log(1-V_t)] &= \Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1}+\gamma t, 2).
\end{aligned}
$$

The last two equalities come from the well-known fact that the expectation of the sufficient statistics in the exponential family is equal to the gradient of the log partition function. Ψ is the digamma function, which appears when taking the derivative of the log normalizer in the beta distribution.

Coordinate ascent algorithm

We can now maximize the lower bound from Equation 12 with a coordinate ascent algorithm by computing the gradient with respect to each parameter and setting it to zero. The derivation is straightforward and leads to the following update rules:

$$
\begin{aligned}
\gamma_{t,1} &= 1 + \textstyle\sum_n \phi_{n,t} \\
\gamma_{t,2} &= \alpha + \textstyle\sum_n \sum_{i=t+1}^{T}\phi_{n,i} \\
\tau_{t,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,t} x_n \\
\tau_{t,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,t} \\
\phi_{n,t} &\propto \exp(S_{n,t}),
\end{aligned}
$$

where the $\phi_{n,t}$ are normalized to have $\sum_t \phi_{n,t} = 1$, and where

$$
S_{n,t} = \mathbb{E}_q[\log V_t] + \sum_{i=1}^{T-1}\mathbb{E}_q[\log(1-V_i)] + \mathbb{E}_q[\eta_t^*]^\top x_n - \mathbb{E}_q[a(\eta_t^*)].
$$

Once the parameters have been learnt, we can use them to approximate the posterior. For example, the predictive distribution can be approximated in the following way:

$$p(x_{N+1}|\mathbf{x},\alpha,\lambda) = \int \left(\sum_{t=1}^{\infty} \pi_t(\mathbf{v}) p(x_{N+1}|\eta_t^*) dP(\mathbf{v},\boldsymbol{\eta}^*|\mathbf{x},\lambda,\alpha) \right)$$
$$\approx \sum_{t=1}^{T} \mathbb{E}_q[\pi_t(\mathbf{v})]\mathbb{E}_q[p(x_{N+1}|\eta_t^*)], \tag{14}$$

where the sum becomes truncated and we have $\mathbb{E}_q[\pi_t(\mathbf{v})p(x_{N+1}|\eta_t^*)] = \mathbb{E}_q[\pi_t(\mathbf{v})]\mathbb{E}_q[p(x_{N+1}|\eta_t^*)]$ thanks to the factorized form of q.

4.2 Gibbs Sampling

The two main MCMC sampling algorithms for DP mixtures are collapsed Gibbs sampling and blocked Gibbs sampling. The former use the Pólya urn scheme to iteratively sample cluster assignements of each observation C_n from the following conditional distribution:

$$p(c_n = k|\mathbf{x}, c_{-n}, \lambda, \alpha) \propto p(x_n|x_{-n}, c_{-n}, c_n = k, \lambda) p(c_n = k|c_{-n}, \alpha). \tag{15}$$

In contrast, blocked Gibbs sampling uses the stick-breaking representation and samples in turn each block of parameters Z, V and η conditionned on the other two blocks. The resulting algorithm looks quite similar to variational inference, and we describe it for multinomial components in section 5. With similar parameterization, the predictive distribution is obtained by computing Monte Carlo expectations using a set of B samples of the parameters from the chain after convergence:

$$p(x_{N+1}|\mathbf{x},\alpha,\lambda) \approx \sum_{t=1}^{T} \mathbb{E}[\pi_k(V)|\gamma_1,\ldots,\gamma_T]\mathbb{E}[p(x_{N+1}|\tau_t)], \tag{16}$$

where $p(x_{N+1}|\tau_t)$ is the marginal likelihood of a Dirichlet-Multinomial model, given by $B(\tau_t + x_{N+1})/B(\tau_t)$, where B is the multinomial beta function, the normalization constant of a Dirichlet distribution.

5 Document clustering with DP mixtures

We consider the problem of clustering text documents. We use a bag of words representation for each document and assume it is sampled from a DP mixture model, where the mixture components are multinomial distributions with parameters θ_t on the size of the vocabulary M, and the base measure is a Dirichlet with parameter λ on the M-dimensional simplex, so that it is conjugate to the multinomial.

In the algorithms, we consider the natural parameters of the multinomials, $\eta_t = \log \theta_t$, and define the corresponding variational distributions $q_{\tau_t}(\eta_t)$ to be Dirichlet with parameter τ_t. The coordinate ascent algorithm for variational inference becomes:

$$\begin{aligned}
\gamma_{t,1} &= 1 + \textstyle\sum_n \phi_{n,t} \\
\gamma_{t,2} &= \alpha + \textstyle\sum_n \sum_{i=t+1}^{T} \phi_{n,i} \\
\tau_t &= \lambda + \textstyle\sum_n \phi_{n,t} x_n \\
\phi_{n,t} &\propto \exp(S_{n,t}),
\end{aligned}$$

with

$$S_{n,t} = \mathbb{E}_q[\log V_t] + \sum_{i=1}^{T-1} \mathbb{E}_q[\log(1 - V_i)] + \mathbb{E}_q[\eta_t]^\top x_n,$$

where we have $\mathbb{E}_q[\eta_{t,m}] = \mathbb{E}_q[\log \theta_{t,m}] = \Psi(\tau_{t,m}) - \Psi(\sum_m \tau_{t,m})$ for every word m in the vocabulary. The multinomial parameters $\theta_{t,m}$ can be recovered from $\exp(\eta_{t,m})$ by normalizing with the constraint $\sum_m \theta_{t,m} = 1$.

A similar parameterization can be obtained for blocked Gibbs sampling, giving the following sampling scheme:

1. For $t \in \{1, \ldots, T\}$, independently sample $\eta_t \sim Dir(\tau_t)$, where

$$\tau_t = \lambda + \textstyle\sum_n \mathbb{I}(z_n = k) x_n$$

2. For $n \in \{1, \ldots, N\}$, independently sample z_n from

$$p(z_n = t | \mathbf{v}, \boldsymbol{\eta}, \mathbf{x}) \propto \pi_t(\mathbf{v}) p(x_n | \eta_t) \propto \pi_t(\mathbf{v}) \textstyle\prod_m \eta_{t,m}^{x_{n,m}}$$

3. For $t \in \{1, \ldots, T\}$, independantly sample $v_t \sim p(v_t | \mathbf{z}) = Beta(\gamma_{t,1}, \gamma_{t,2})$, where

$$\gamma_{t,1} = 1 + \textstyle\sum_{n=1}^N \mathbb{I}(z_n = t)$$
$$\gamma_{t,2} = \alpha + \textstyle\sum_{i=t+1}^T \sum_{n=1}^N \mathbb{I}(z_n = i)$$

6 Experiments

We ran our algorithms on a corpus of news articles from the Associated Press[1]. The dataset contains 2246 documents with a vocabulary of 10473 words. The variational inference algorithm converged in about 1 second after 10 iterations, where convergence was assessed by looking at the lower bound on the log marginal likelihood. We set the scaling parameter α to one, the parameter of the Dirichlet base measure to $(1, \ldots, 1)^\top$ and the truncation level T to 100. Figure 2 shows some examples of the obtained multinomial clusters after 50 iterations, where the words on the top have higher probability.

We compared variational inference and blocked Gibbs sampling by using 200 documents for inference and computing the mean held out log-probability on 100 different documents (mean over the held out set of the logarithm of the predictive probability, given by Equations 14 and 16). We focused on blocked Gibbs sampling since it has computational benefits compared to the collapsed Gibbs sampler, where variables are updated one at a time. Table 1 shows results of running time (on a Macbook Pro) and held out log probability (averaged over a few rounds) for $T = 100$ and the same parameters α and λ. Part of the reason for the faster iteration time is that variational inference can fully exploit vectorized computations, while this cannot be done in MATLAB for Gibbs sampling, but most of the additional cost in Gibbs sampling is due to the sampling itself. We ran both algorithms for 15 iterations, which is enough for variational inference to reach convergence, and in the case of Gibbs sampling we didn't observe any improvements in held out probability by letting the chain run longer: the clusters assignments remain practically the same after the first few iterations. More accurate diagnostics can be used to assess statistical convergence of the chain (see [2]).

[1] Available at `http://www.cs.princeton.edu/~blei/lda-c/ap.tgz`

	Time	Mean held out log probability
Variational inference	0.6 seconds (0.04s/iteration)	−1661.04
Blocked Gibbs sampling	60 seconds (4.1s/iteration)	−1617.27

Table 1: Comparison of running time and average held out log probability for variational inference and blocked Gibbs sampling.

7 Discussion and Conclusion

We have seen how the DP makes it possible to define mixture models with an unbounded number of mixture components that increases with the number of observed samples. We saw how Bayesian posterior inference can be done with variational inference and Gibbs sampling algorithms, and explored the advantages of variational inference in a large-scale, high-dimensional setting with text document modeling. However, one can show that the number of clusters in a DP typically increases as $O(\alpha \log n)$, where n is the number of observations (see [9]), hence this might not be the perfect way to find the right number of clusters in the data, and when one has some prior domain knowledge on the data, then using a finite mixture model with an explicitely chosen number of components might give better results.

In the context of text document modeling, we saw that when using DP mixtures with multinomial components, there can be words shared across multiple components, and different components sharing a common set of words which belong together in a given topic. This happens because each document is modeled as a sample from a single "topic". If the goal is to find a set of topics appearing in a corpus of documents, rather than clustering documents, one can model each document as being sampled itself from a mixture of topics, where each topic is a multinomial. This problem has been studied extensively with a class of models called topic models. The most popular of these models is Latent Dirichlet Allocation (LDA) [3], which considers a fixed number of topics, and nonparametric extensions of LDA base on the DP have been developed, e.g. with the Hiererarchical Dirichlet Process (HDP) [10].

References

[1] Blackwell, D., and MacQueen, J. B. Ferguson distributions via pólya urn schemes. *The annals of statistics* (1973), 353–355.

[2] Blei, D., and Jordan, M. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis* (2006).

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[4] Ferguson, T. S. A bayesian analysis of some nonparametric problems. *The annals of statistics* (1973), 209–230.

[5] Ishwaran, H., and James, L. F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 453 (2001).

government	percent	i	police	court	soviet	iraq
police	million	dukakis	i	police	party	iraqi
people	year	bush	people	drug	i	saudi
president	billion	new	two	years	government	united
two	market	campaign	officials	south	gorbachev	kuwait
last	stock	people	water	two	two	states
united	new	president	hospital	i	people	state
new	index	jackson	state	government	first	military
states	prices	two	air	judge	political	two
i	rose	state	three	prison	new	i
(318)	(201)	(170)	(165)	(135)	(131)	(111)

i	dollar	northern	israel	mecham	testimony	demjanjuk
new	late	inches	soviet	senate	bishops	death
dukakis	new	rain	peace	state	sexuality	sheftel
rep	yen	snow	israeli	office	vatican	nazi
president	bid	central	talks	impeachmen	hunthausen	court
house	london	southern	occupied	trial	archbishop	holocaust
people	gold	texas	palestinian	governor	satellite	trial
bush	ounce	temperature	territories	threat	catholic	crimes
republican	york	heart	minister	republican	grand	sentence
year	troy	new	jewish	attorney	congressiona	camp
(73)	(64)	(12)	(16)	(6)	(5)	(2)

Figure 2: Examples of topics (multinomial clusters) obtained on all 2246 documents of the Associated Press dataset. For each topic, we show the 10 words with highest probability, and the number of documents in the cluster is shown in brackets.

[6] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning 37*, 2 (1999), 183–233.

[7] Ranganathan, A. The dirichlet process mixture (dpm) model. *Journal of Machine Learning Research* (2006).

[8] Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica* (1994).

[9] Teh, Y. W. Dirichlet process. *Encyclopedia of Machine Learning* (2007).

[10] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the american statistical association 101*, 476 (2006).

[11] Wainwright, M. J., and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*, 1-2 (2008), 1–305.