**Latent Dirichlet Allocation**

[**http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf\\**](http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf\\)

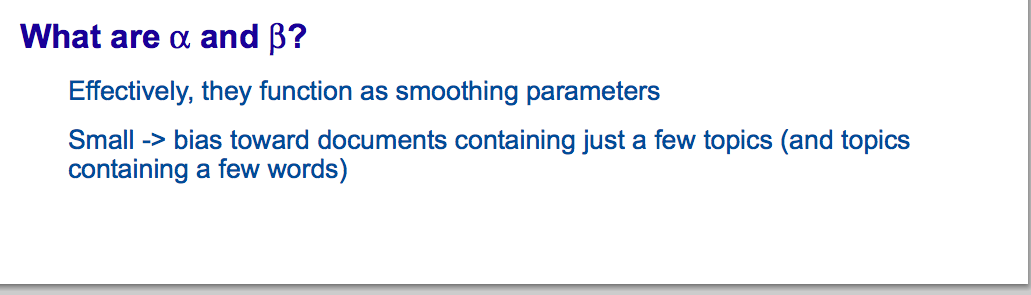
**Results of perplexity vs number of topics as in this paper**[**http://ai.stanford.edu/~ang/papers/nips01-lda.pdf**](http://ai.stanford.edu/~ang/papers/nips01-lda.pdf)

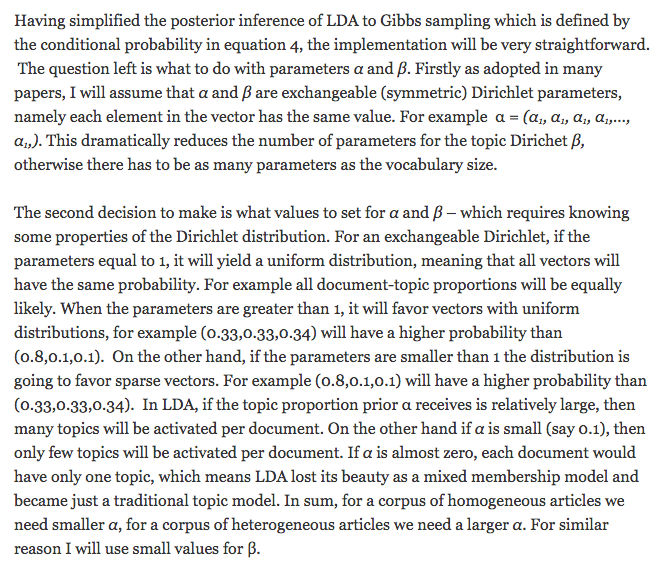
**Explains almost everything for LDA**[**http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf**](http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf)

**Difference between variational inference and gibbs sampling applied to LDA**

[**http://stats.stackexchange.com/questions/76276/what-are-the-main-differences-between-classical-and-gibbs-sampling-latent-dirich?rq=1**](http://stats.stackexchange.com/questions/76276/what-are-the-main-differences-between-classical-and-gibbs-sampling-latent-dirich?rq=1)

[**https://liucanblog.wordpress.com/2014/01/12/15/**](https://liucanblog.wordpress.com/2014/01/12/15/)

****

****

**Questions in LDA and Convergence Issues (IMPORTANT)**[**https://lists.cs.princeton.edu/pipermail/topic-models/2008-August/000319.html**](https://lists.cs.princeton.edu/pipermail/topic-models/2008-August/000319.html)

Model documents as arising from multiple topics

K topics are associated with a collection of documents

Each document exhibits these topics in different proportions

Hidden variable model – posit a hidden structure in the observed data and then learn that structure using posterior probabilistic inference

LDA – hidden variables represent the latent topic structure, ie, the topics themselves and how each document exhibits them

Given a collection, the posterior distribution of the hidden variables given the observed documents determines a hidden topical decomposition of the collection

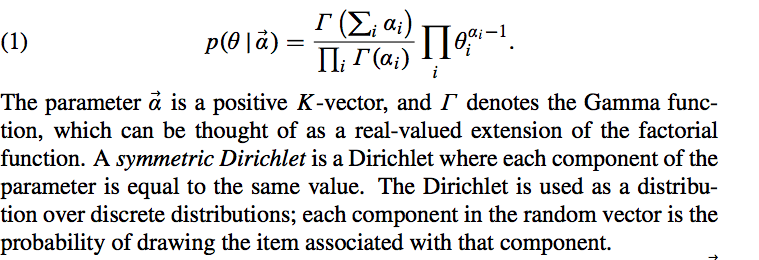
Mixture models – since documents can exhibit multiple topics

BETA\_ (1:k) – topics

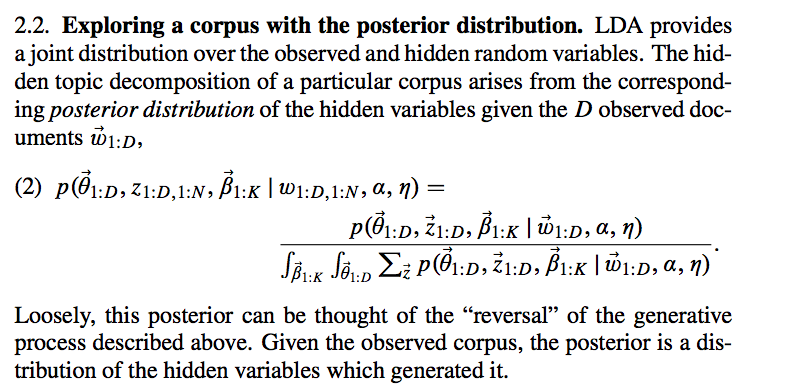
THETA: per document topic proportions

Z – per word topic assignment

Below is the probability density for the Dirichlet

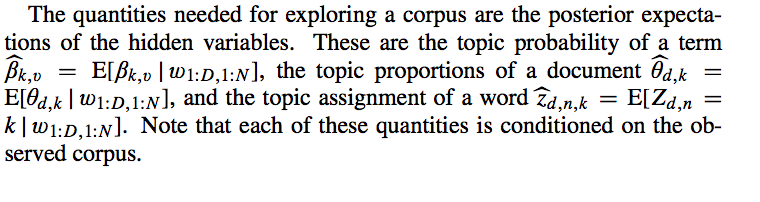


LDA contains two Dirichlet random variables: topic proportions THETA and the topics BETA which are the distributions over the vocabulary



The above distribution is intractable to compute because of the above denominator

The quantities needed for exploring the corpus are the posterior expectations of the hidden variables. These are the topic probability of a term, the topic proportion of a document and the topic assignment of a word.



Dirichlet distributions were used as conjugate priors for multinomials in Bayesian modeling

- it is also preferable to think of the Dirichlet in the model as a component of the likelihood.

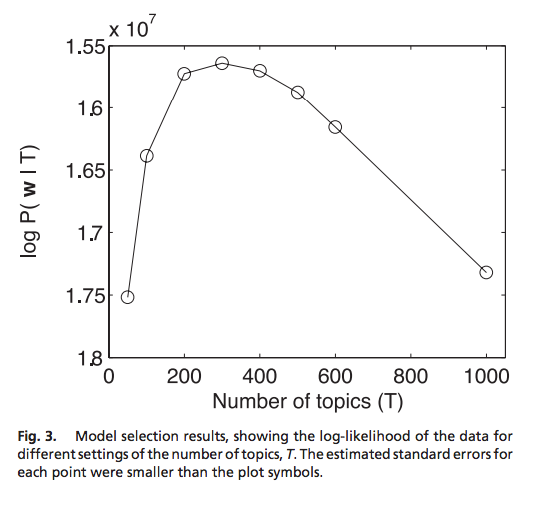
**Difference between LDA and simple Dirichlet multinomial clustering model**

In the clustering model, the Dirichlet is sampled once for the corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. Such a model restricts a document to being associated with a single topic.

**Posterior Inference for LDA**

**Gibbs sampling for intractable posterior**[**http://psiexp.ss.uci.edu/research/papers/sciencetopics.pdf**](http://psiexp.ss.uci.edu/research/papers/sciencetopics.pdf)

**Extended Results for the assignment**

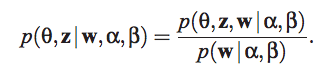
**1. **

**Question I**\*topic posteriors as function of Gibbs sweeps

\*Perplexity for docs after Gibbs

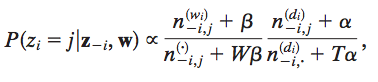
In LDA, we consider the documents to be represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

The key problem in LDA is that of computing the posterior distribution of the hidden variables given a document:



The above distribution is intractable to compute in general, and hence to normalize the distribution we marginalize over the hidden variables.

We therefore estimate the posterior of the word topic assignments given the observed words directly, by marginalizing out theta and beta.



here the first ratio expresses the probability of w\_i under topic j and the second ration expresses the probability of topic j in document d.

These counts are the only information necessary for computing the full conditional distribution allowing the algorithm to be implemented efficiently by caching the relatively small set of nonzero counts.

Below, we plot how the topic posteriors depend on the number of Gibbs sweeps for each document out of the 2000 documents in training set A.

\*\* Plotting topic posteriors vs number of Gibbs sweeps shows how the posterior distribution of the topics for each document evolves with number of iterations. The plot can be interpreted as: how does the mixture of our belief about proportion of different topics within a document evolves with number of iterations.

***Relation of Theta with Number of Gibbs sweeps***

As more sweeps – better convergence – higher proportions assigned to multiple topics within each document..

***Relation with Number of Topic Categories***

***Relation of Beta with Gibbs sweeps:***It’s how the proportion of topics for a given word

***Perplexity and Number of Gibbs sweeps:***10 – 1.897e+03

50 – 1.6525e+03

**\*For the topic proportion of a word – compare BMM and LDA**

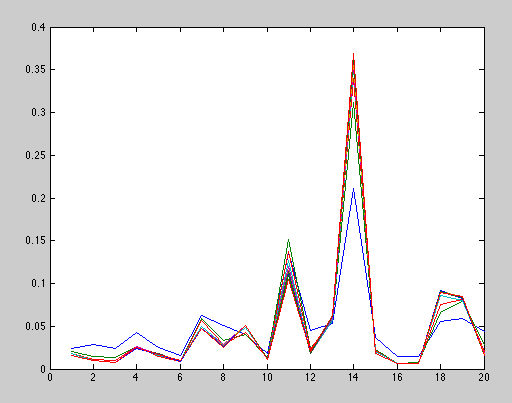
**Relation of Perplexity vs Number of Gibbs iterations**

**Question G**Perplexity of a held-out test set to evaluate the models. Perplexity is monotonically decreasing the likelihood of the test data and is equal to the inverse of the geometric mean per-word likelihood.

Mixture proportions – posterior probabilities of the mixture components.

\*\* Across number of Gibbs iterations – becoming more certain that any document may contain topic 14

\*\* For topic 14 below – posterior probability for that particular topic shows that as you have more sweeps, the uncertainty that this topic may be present across any document over the training set decreases.  
\*\* Comparing topic 14 with topic 2 – after 20 gibbs sweeps, you are more certain that topic 14 is present across any document in the training set, compared to the certainty about topic 2



**Perplexity and evaluating models:**Test set is a collection of unseen documents -   
  
In LDA – theta represents the topic distributions for the documents in the training set and so ignored to compute likelihood of unseen documents.   
  
Lower the perplexity – higher the likelihood – better the model.

**For LDA – better model structure. LDA is a generative process of creating documents given that you know the topic structure.**

**Better informative you are about the document structure – better the document generative process.**

Document clustering is an unsupervised learning task that partitions a set of documents to produce disjoint subsets of clusters.

**Question H**

Convergence of BMM   
<https://faculty.cs.byu.edu/~ringger/CS679/papers/WalkerRingger-Gibbs-kdd2008.pdf>

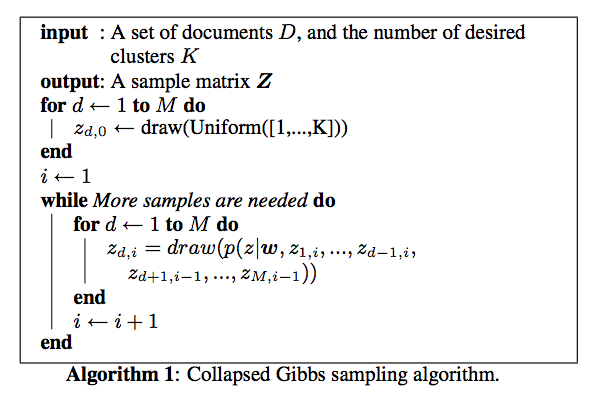
Collapsed Gibbs sampling on a mixture of multionomials model  
Evidence that the sampler converges quickly within a relatively small number of samples. – Collapsed sampler clustering algorithm

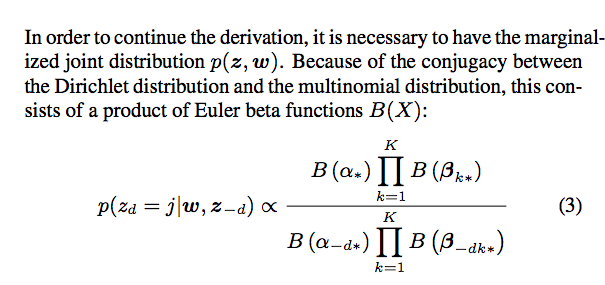
Papers on automatic selection of the mixture of components (K)  
[14] M. Meila and D. Heckerman. An experimental comparison ˘ of model-based clustering methods. Machine Learning, 42(1–2):9–29, Jan. 2001. [15] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265, June 2000.

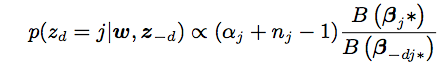
In the case of clustering, the only specifically relevant variables are the hidden document cluster labels, z. There may be some reason to sample the values for the θ matrix as well, but this is not strictly part of the clustering task. In addition to the storage benefit of not sampling from uninteresting variables, collapsed samplers have also been shown to converge relatively quickly because they contain fewer dependencies between sampled parameters and treat the marginalized parameters exactly [4, 9, 22].

Collapsed samplers cannot be used in all cases but have the limiting requirement that the variables to be marginalized out must be marginalizable in closed form. Fortunately, conjugacy between the Dirichlet and Multinomial distributions makes this possible for the model shown in Figure 1.

Collapsed Gibbs Sampling Algorithm







Using the Gibbs sampling algorithm above, we get a matrix Z of samples for the documents in D, such that z\_\_d,i is the ith label sampled for the dth document in D.

Discussions – relating to using Z to choose the clustering for D:  
1. How long does it take to converge

2. How many samples should be taken?  
3. How should the collected samples be summarized?  
  
  
**MCMC sampling techniques are guaranteed to converge in the limit to the target distribution. However, because consecutive draws in the chain can be highly correlated, the samples from the beginning of the chain can be highly influenced by the random initialization state. MCMC algorithms often include a parameter called “burn-in” which specifies the number of initial samples that should be discarded from the beginning of the chain to reduce of random initialization of the samples used for parameter estimation and inference.**

**Label switching is a problem when clustering using MCMC techniques on mixture models. It is possible for label switching to occur mid-chain. Therefore, averaging across multiple samples can be worse than taking on any individual samples as the chain summary, because the meaning of each label can change across multiple samples.**

We wanted to determine the extent to which label switching occurs using the collapsed sampler with our model.

**SEE THIS DOCUMENT FURTHER**[**http://metaoptimize.com/qa/questions/13729/gibbs-sampling-in-document-clustering**](http://metaoptimize.com/qa/questions/13729/gibbs-sampling-in-document-clustering)

> *(2) What are the guarantees for Gibbs sampling for convergence to a* > *local optimum? A global optimum? An optimum of what precisely?*  In the LDA model, Gibbs produces samples from the posterior distribution of topics given words. The term “converged” is perhaps misleading because Gibbs does not explicitly seek optima in the posterior distribution. Instead, as you probably know, we say that the sampler has “converged in distribution”, meaning that the samples produced occur in proportion to their posterior density. A “converged” Gibbs sampler will sample more frequently around the modes of the posterior distribution, which, by definition are maxima. However, sampling is typically a bad way to optimize; optima finding is an optimization task, which Gibbs is not meant to do. However, Gibbs IS (when properly converged) good at exploring, in the sense that it frequents the (possibly many) regions of high probability. Unfortunately, when variables are highly correlated, Gibbs can "get stuck" and not actually sample the full distribution. In the paper by Walker and Ringger (KDD 2008), they touched on this for Gibbs sampling applied to a mixture of multinomials model. Although Gibbs outperforms garden variety EM (which only finds local maxima) on their task, Gibbs on their model is likely NOT exploring the full space (i.e. confined to a local region that includes multiple maxima), and hence, is likely NOT sampling around the global optima at all (when it should be there most often). The same could be true for LDA, though I’m not aware of anyone who has analyzed this. However, it is unclear how much this “stuckedness” hurts Gibbs.

Why convergence of LDA is more difficult than BMM?

**Question J**

**Word Entropy**

This computes the entropy over the word distribution for a topic. Is all the mass centered on a few words, or is the mass evenly spread out across a lot of words? Word entropy is to type count as document entropy is to token count. They are the same metrics, just looking at words instead of documents.

### Document Entropy

This metric computes the entropy of the distribution over documents for a given topic. I.e., given that a token is labeled with this topic, what is the probability distribution over documents that it probably came from? We get that distribution from the Mallet state file. This tells you how broadly a particular topic was used in the corpus. We have found that it is correlated closely with the log of the token count of the topic.

For each topic – word entropy means – a lower entropy corresponds to more certainty (less uncertainty) about which bag of words is representing that topic. In other words – you are more certain about the bag of words that correspond to the topic.  
A category or topic with higher entropy means – you are not sure which bag or sets of words are responsible for that topic.

**EXTENSION:  
Variational inference and other approximate inference methods in LDA instead of Gibbs sampling**[**http://papers.nips.cc/paper/3113-a-collapsed-variational-bayesian-inference-algorithm-for-latent-dirichlet-allocation.pdf**](http://papers.nips.cc/paper/3113-a-collapsed-variational-bayesian-inference-algorithm-for-latent-dirichlet-allocation.pdf)

* **three current approachers:  
  a. variational Bayes  
  b. expectation propagation**
* **c. collapsed gibbs sampling**

**Paper: Model-based document clustering with a collapsed gibbs sampler. I**

[**https://faculty.cs.byu.edu/~ringger/CS679/papers/WalkerRingger-Gibbs-kdd2008.pdf**](https://faculty.cs.byu.edu/~ringger/CS679/papers/WalkerRingger-Gibbs-kdd2008.pdf)

[**http://www.anthology.aclweb.org/W/W08/W08-2106.pdf**](http://www.anthology.aclweb.org/W/W08/W08-2106.pdf)

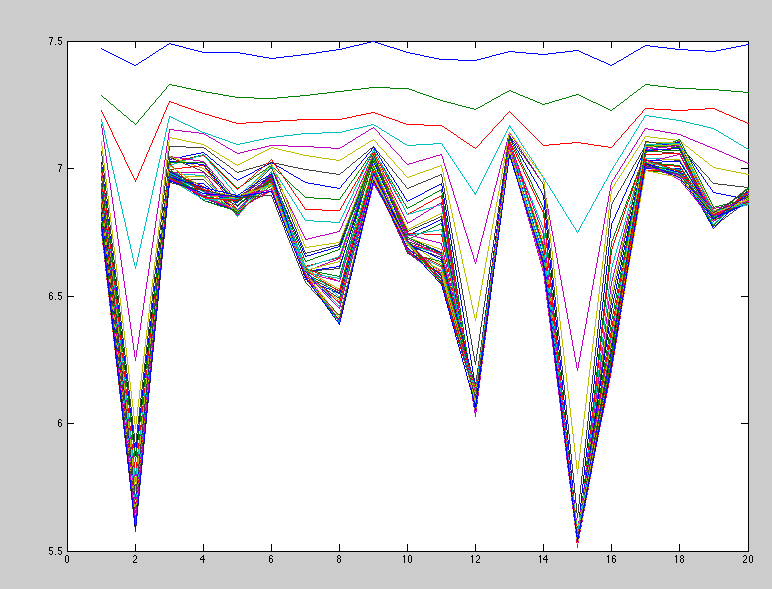
**\*\*\*\*\* - includes discussion of speed of convergence of gibbs sampler**

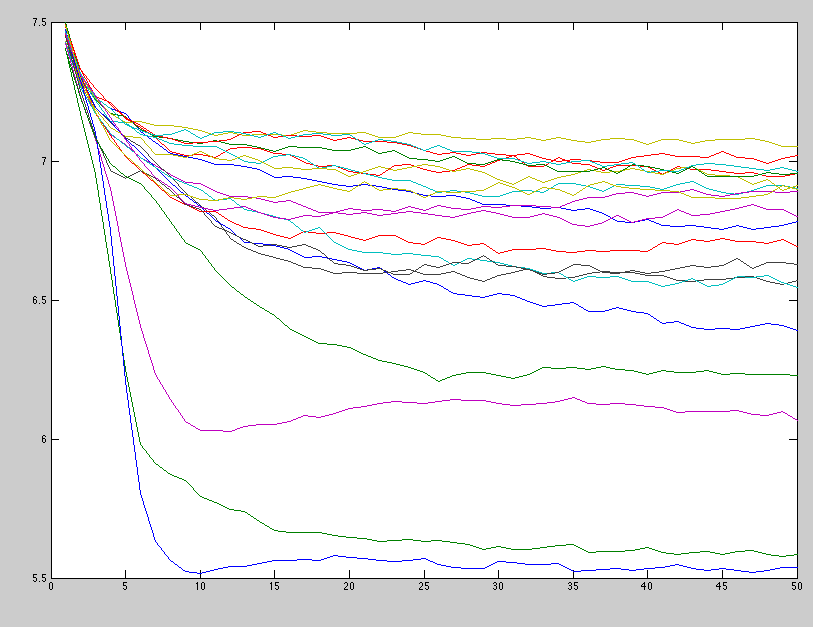
**\*\*\* Label switching – how often does the k switch along document – for a fixed document, observe the k along iterations**

**Plots for word entropy in LDA – plotting for word 10**

[**http://qpleple.com/word-relevance/**](http://qpleple.com/word-relevance/)

**Instead of the global word frequency, we consider the frequency of words within a topic. The entropy is the distribution of topics given a word – capturing how much the word w is shared across several topics. x**

****

****

Entropy, in this context, expresses the disorder of words within each topic (cluster) where for each cluster, the category distribution of the words is calculated first, and the probability that a member of topic j belongs to category i. We can then compute the word entropy for each topic.

From the figures above, the entropy decreases as we have more Gibbs iterations for the LDA model. For the LDA model, as we have more iterations, our informative measure of how the topics generated the document model increases. The topics, being consisting of a measure of words representing each topic becomes more aligned to represent the documents – which is to say that the word entropy (or disorder of orders for each topic) decreases. This further means that, our uncertainty about which words are responsible for representing the particular topics that can generate the documents decreases (more informative) which is why the word entropy decreases. The word entropy measures the expected value of the information contained in each topic. With higher Gibbs iterations, the LDA model reaches close to convergence approximating the true posterior distribution of topic proportions – and so the words become more informative about the topic it represents.   
Word entropy decreases because the words becomes more certain about the topics at convergence.

The above figure shows the word entropy for each topic across the number of Gibbs iterations. Out of 20, two word entropy for the topics are lower than the others. This suggests that the words representing topic 15 and topic 2 are more informative (less disordered) about that topic.

<http://www.ijarcce.com/upload/may/Validation%20of%20Document%20Clustering%20based%20on%20Purity%20and%20Entropy%20measures.pdf>