# Simulated Annealing for Dirichlet Priors in LDA

Andrey Kuzmenko
for.akuz@gmail.com
http://akuz.me

January 25, 2014

**Abstract**

This note describes a way of applying a special technique for MCMC inference called *simulated annealing* to multinomial random distributions with Dirichlet priors. In particular, we look at speeding up the process of Gibbs sampling inference in Latent Dirichlet Allocation (LDA).

## 1 Dirichlet Priors

Topic models such as LDA [1] use conjugate Dirichlet priors for multinomial latent random variables. In LDA, these multinomial latent variables are word by topic $p(w \mid t)$ and topic by document $p(t \mid d)$ distributions. These *distributions* are assumed to be Dirichlet-distributed *unobserved random variables.*

This note first discusses symmetric Dirichlet priors. Later we extend the discussion to asymmetric Dirichlet priors, in order to be able to specialise the topics by increasing prior probabilities of specific words.

A symmetric Dirichlet prior is defined by a single hyperparameter $\alpha$, which represents a prior mass allocated to each of the dimensions. It is usual to set the the Dirichlet hyperparameter to a small value $\alpha \ll 1$, so that the resulting topics concentrate on a small subset of all words, and documents express a small subset of all topics.

## 2 Burn-in Speed for Small $\alpha$

Let us consider a collapsed Gibbs sampler for LDA [2]. The algorithm iterates in a loop over all word-places in the corpus and samples topic allocation for each of them. The posterior distributions of $p(t \mid d)$ and $p(w \mid t)$ are then constructed using their Dirichlet priors, plus the aggregate statistics of topic allocations of all word-places in the corpus.

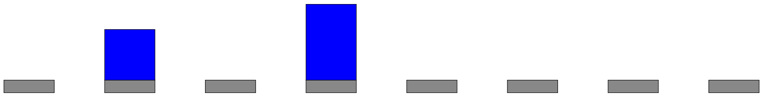


Figure 1: Dirichlet posterior mass with small initial $\alpha$

Let us investigate the dynamics of Dirichlet posteriors during sampling with small $\alpha \ll 1$. Figure 1 shows posterior mass after a few observations have been added to the prior (grey shows prior mass, blue shows added posterior mass).

When a small $\alpha$ is used, the posterior distribution quickly concentrates on a subset of dimensions. After that, however, it will be hard for individual latent variable posteriors to move to other areas of global posterior distribution that have different significant dimensions.

Sampling values from the dimensions that have a very small probability will be unlikely. The posteriors of the multinomial variables will start to move very slowly in the search space. Adding other significant dimensions will be unlikely, because they would need to be added one-by-one.

This leads to a very slow burn-in speed for Gibbs sampling when small values of hyperparameters $\alpha$ are used at the start of the sampling process.

## 3 Simulated Annealing with Dirichlet Priors

The simulated annealing[1] heuristic uses the concept of temperature. In the context of MCMC sampling, the temperature regulates the size of steps that we take when randomly exploring the search space. This allows locating the areas of high probability quickly, and then, by lowering temperature and thus reducing the size of the search steps, exploring them in more details.

Now let us apply this heuristic to sampling a posterior multinomial distribution with a Dirichlet prior.

Basically, we do *not* want our multinomial posteriors to quickly concentrate on a limited subset of dimensions, and then explore the surrounding areas.

Instead, we want out multinomial posteriors to *narrow down* on a set of dimensions, by starting from a larger set of significant dimensions.

## 4 Dynamic Dirichlet $\alpha$

For a posterior multinomial distribution with a Dirichlet prior, the simulated annealing effect can be implemented by dynamically changing the hyperparameter $\alpha$ throughout the burn-in process, starting from some large value, and gradually reducing it to the small target value. For example:

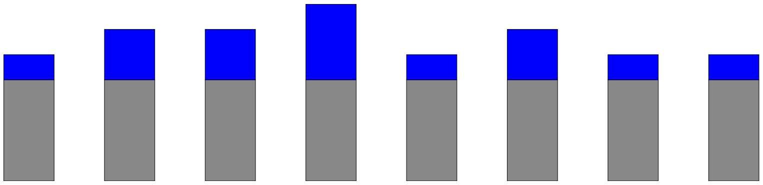
$$\alpha: \; 1.0 \to 0.001$$

[1] http://en.wikipedia.org/wiki/Simulated_annealing

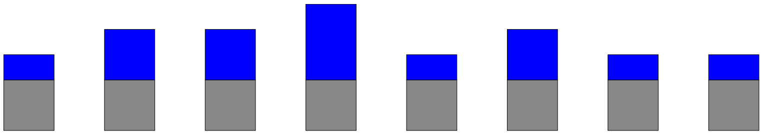
During this process, it will be unlikely for Gibbs sampling to become "trapped" in some small area of the posterior distribution, because the number of significant dimensions of the posterior multinomial will be gradually decreasing from a larger to a smaller number.

This process can schematically be illustrated by the following sequence of posteriors of the multinomial distribution (where temperature $t = C\alpha$, and we discuss the definition of temperature and constant $C$ later).

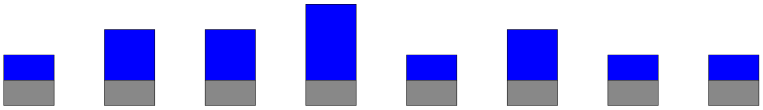
Temperature $t = 2.0$:



Temperature $t = 1.0$:



Temperature $t = 0.5$:



At each temperature, we perform a number of Gibbs sampling iterations to allow the system to burn-in into the current temperature (find a high probability region of posterior for that temperature).

After we have reduced the temperature to the target value, the posterior is likely to be in the area of high probability, and so we can start collecting samples.

By following simulated annealing burn-in process we can significantly reduce the number of steps necessary for Gibbs sampling burn-in, because the high posterior probability area will be identified much faster.

A possible useful computer science analogy would be to consider searching for an element in a sorted list. The usual Gibbs sampling would be equivalent to searching for the element by looking sequentially through the entire list. However, Gibbs sampling with simulated annealing would be equivalent to using a binary search algorithm.

Obviously, a fundamental investigation of convergence properties is necessary. But from the above analogy, one can try to make a conjecture that, if $N$ is the required number of iterations for a regular Gibbs sampling to converge, then Gibbs sampling with simulated annealing might require only $O(log(N))$ iterations.

## 5 Temperature $t$ and Hyperparameter $\alpha$

Let us first consider a *single* multinomial posterior of dimensionality $K$ with a symmetric Dirichlet prior. Let us also assume that we know the number of data samples $S$ that we add to the prior in order to obtain the posterior.

The choice one should make for hyperparameter $\alpha$ is often unclear, so let us try to see if we can define a more transparent approach to setting $\alpha$.

Instead of setting the value $\alpha$ directly, we can define it through a ratio of the total prior mass $V_{prior}$ to total mass of the data samples $V_{data}$.

We define this ratio as the *temperature*:

$$t = \frac{V_{prior}}{V_{data}} = \frac{\alpha K}{S}$$

from where we obtain:

$$\alpha = \frac{t\,S}{K}$$

This gives us the expression for setting $\alpha$ to use during Gibbs sampling, given the dimensionality $K$, number of samples $S$, and a desired temperature $t$.

Now, setting the desired temperature $t$ is easier because we know its interpretation: it regulates how much of the posterior distribution will be affected by the prior versus the data.

For example, for $t = 1.0$ the posterior distribution will be equally affected by the prior and the data. For $t = 0.5$ the posterior distribution will be twice as much affected by the data than by the prior.

As $t \to 0$, the influence of the prior diminishes, and the posterior begins to be defined mostly by the data.

## 6 Temperature in LDA Gibbs Sampling

LDA model assumes the following latent variables:

- $p(t \mid d)$ for each document
- $p(w \mid t)$ for each topic

Given a desired temperature $t$ for sampling from the entire model (joint distribution), we want to set hyperparameter $\alpha$ for every latent variable.

Please note that in LDA literature Dirichlet hyperparameter for $p(t \mid d)$ is denoted as $\alpha$ and Dirichlet hyperparameter for $p(w \mid t)$ is denoted as $\beta$.

Below we will consider setting *asymmetric* Dirichlet priors. The Gibbs sampling formulae for sampling with asymmetric priors are described in [4].

## 7 Temperature $t$ and LDA Hyperparameter $\beta$

Let us first consider setting $\beta$ on Dirichlet priors of $p(w \mid t)$.

In the process of collapsed Gibbs sampling, at each moment in time, every word-place is allocated to some topic, and the aggregate statistics are updated (used in posterior).

Therefore, the total mass of the data being added to obtain posteriors of $p(w \mid t)$ for all topics combined is equal to $M$, the total number of words-places in the corpus.

To achieve the desired temperature of sampling $t$, the total prior mass we want to allocate to all topics must be $t\,M$ (from the definition of temperature).

However, we do *not* know yet which proportion of total prior mass $t\,M$ to allocate to individual topics. We only know that the total prior mass allocated to all topics must be $t\,M$.

It is convenient at this point to introduce additional parameters $\gamma_k > 0$:

$$\sum_{k=1}^{K} \gamma_k = 1$$

where $K$ is the number of topics.

We can now use parameters $\gamma_k$ to distribute total prior mass $t\,M$ at a given temperature $t$ among the individual topics by setting $\beta_k$ for each topic $k$ as:

$$\beta_k = \gamma_k \frac{t\,M}{W}$$

where $W$ is the number of *unique* words in the corpus.

The newly introduced parameters $\gamma_k$ can be called *topic proportions*.

If we set all $\gamma_k$ to the same value $1\big/K$, we indicate that at a given temperature all topics should be influenced by the data in equal proportions.

If we set some of the $\gamma_k$ to values larger than the others, then we indicate that these topics should be influenced by the data less, compared to the topics with lower values of $\gamma_k$.

Setting some of the $\gamma_k$ to larger values can help to differentiate vague topics distributed over many words, from the topics distributed over more compact sets of significant words.

In the example program for running LDA Gibbs sampling on a set of text files[2], a large value of $\gamma_0 = 0.5$ is set in order to detect a topic of most common words (that are not stop words). The rest of the prior mass is distributed among all the other topics equally.

## 8 Temperature $t$ and LDA Hyperparameter $\alpha$

Let us now consider setting $\alpha$ on Dirichlet priors of $p(t \mid d)$.

[2] http://akuz.me/software

Again, in the process of collapsed Gibbs sampling, at each moment in time, every word-place is allocated to some topic.

Therefore, for a given document, the total mass of the data being added to the prior to obtain posterior of $p(t \mid d)$ will be equal to $L_d$, the length of the document.

To achieve the desired temperature of sampling $t$, the total prior mass we want to allocate to all topics in a specific document must be $t\,L_d$ (from the definition of temperature).

Now we need to decide how to distribute the total prior mass $t\,L_d$ among the topics. In order to do this, we can make the following additional assumption.

Remember, we defined $\gamma_k$ topic proportions in order to differentiate topics with many significant words from topics with fewer significant words. We now assume that the topics *also* occur in documents with frequencies proportional to $\gamma_k$.

This is reasonable, because we can expect that words from vague topics (higher $\gamma_k$) will in general occur in text more frequently than words from more specific topics (lower $\gamma_k$).

Therefore, for a given document $d$ and temperature $t$, we can distribute the total prior mass $t\,L_d$ for $p(t \mid d)$ among the topics proportionally to $\gamma_k$:

$$\alpha_{d,k} = \gamma_k\, t\, L_d$$

Note that by setting the $\alpha_{d,k}$ in this way, we take the length of the document into account.

## 9 Conclusion

Using simulated annealing heuristic allows Gibbs sampling to converge much faster on the area of high posterior probability.

This article describes how simulated annealing can be applied to Gibbs sampling for inferring topics in the LDA model.

Finally, alternative parameterisation is introduced for the priors on $p(w \mid t)$ and $p(t \mid d)$ in order to unequally spread the prior among different words and obtain topics centered around specific words of interest.

The simulated annealing algorithm, including the reparametrisation, is implemented in the LDAGibbs (and related classes) in the NLP library available at http://akuz.me/software.

## References

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

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[3] Gregor Heinrich. Parameter estimation for text analysis. *Technical Note, vsonix GmbH + University of Leipzig*, Germany, version 2.4: August 2008.

[4] Andrey Kuzmenko. Gibbs Sampling for LDA with Asymmetric Dirichlet Priors. *Research Note, 1 July 2011, http://akuz.me/notes*