GIBBS SAMPLING

GIBBS SAMPLING

My Project consists of two parts. In the first part, I used Gibbs Sampling to get samples from a Normal Distribution. In the second part of the project I explored the applications of Gibbs Sampling in Topic Modelling specifically Latent Dirichilet Allocation (LDA) where we used Gibbs Sampling for generating samples from complex distributions. I explored a computionally optimized way of Gibbs Sampling technique also known as Collapsed Gibbs Sampler.

GIBBS SAMPLING

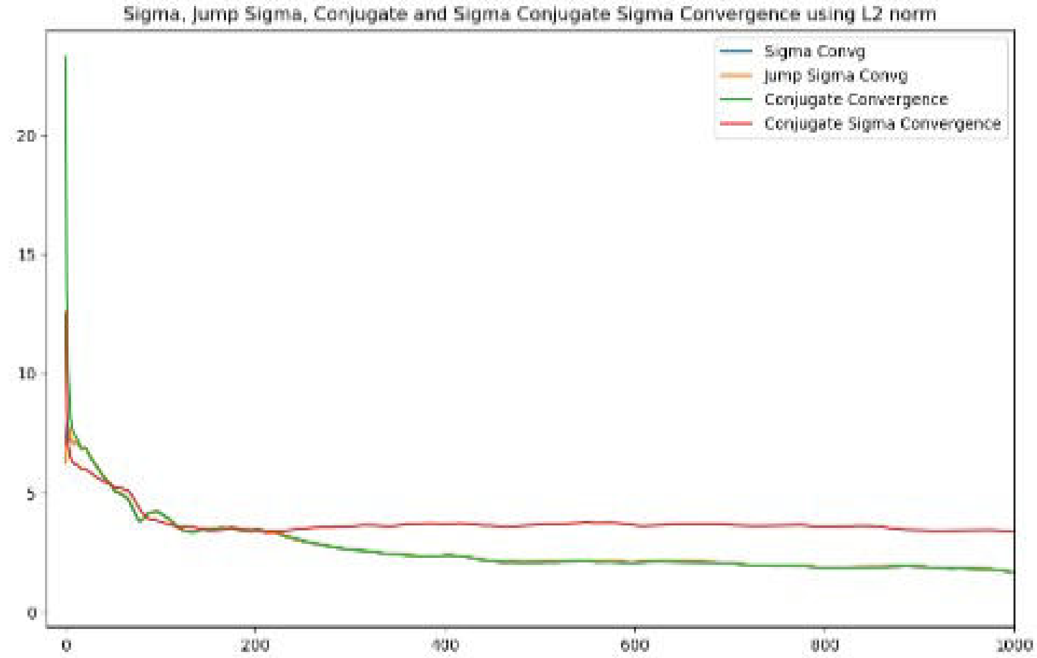
OVERVIEW

- GIBBS SAMPLING
- USING GIBBS SAMPLING TO FIND SOLUTION TO LINEAR SYSTEM OF EQUATIONS (GIBBS SAMPLING FOR GUASSIAN DISTRIBUTION)
- USING GIBBS SAMPLING IN TOPIC MODELLING AND LATENT DIRICHLET ALLOCATION (GIBBS SAMPLING FOR DIRICHLET AND OTHER DISTRIBUTIONS)
- DERIVING GIBBS SAMPLER FOR LDA
- COLLAPSED GIBBS SAMPLER

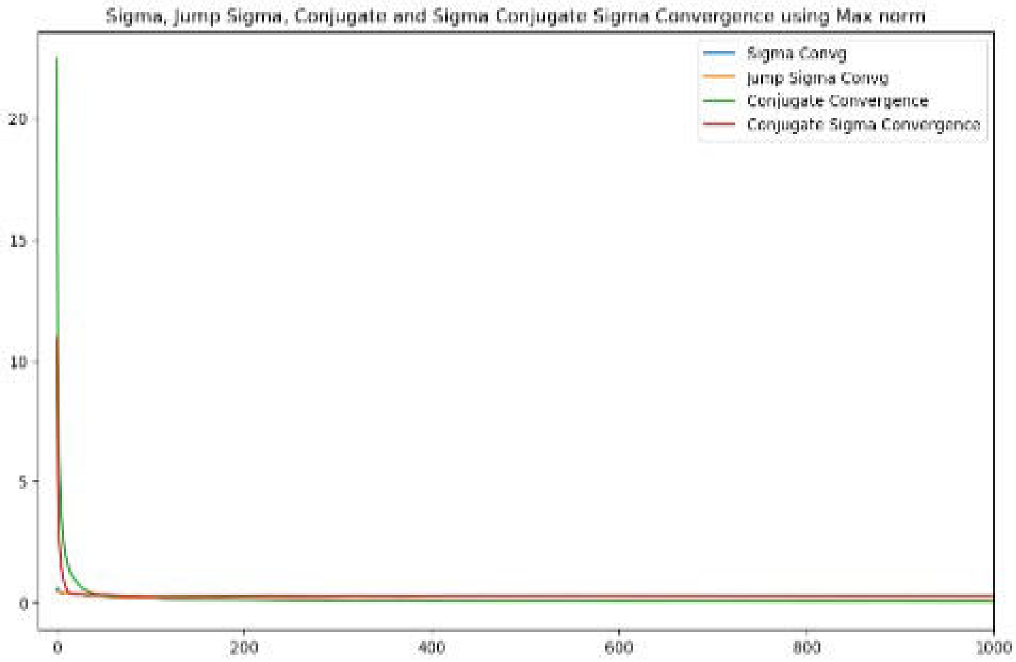
GIBBS SAMPLING

- We used Gibbs Sampling to sample from a Multivariate Normal distribution iteratively and checked whether the samples are converging to the expected distribution and the samples were converging to expected distribution.
- The following are the plots of the convergence metrics for a 1225 x 1225 sparse matrix with both L2 Norm and Max Norm. We can clearly observe that the norm difference is very less.

GIBBS SAMPLING

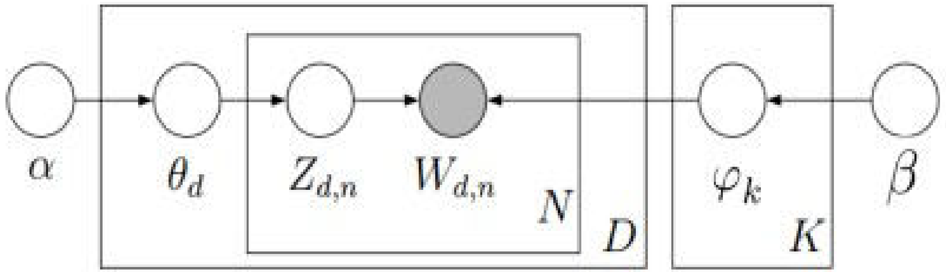


GIBBS SAMPLING



TOPIC MODELLING

- **Latent Dirichlet Allocation is one of Topic Modelling techniques where we classify a document into a particular topic.**
- **Graphical Representation of LDA Model:**



GIBBS SAMPLING

Notations:

D : Number of Documents

K : Number of topics

V : Number of words in the vocabulary

N_d : Number of words in document d or Length of document d

θ_d : $< \theta_{d,t}$ for $t \in \{1, 2, ...K\} >$ Topic distribution for document d

ψ_k :$< \psi_{k,v}$ for $k \in \{1, 2, ...V\} >$Word distribution over Vocabulary V for topic k

$z_{d,n}$: Latent topic assignment to the n^{th} word in document d

$w_{d,n}$: n^{th} word in document d

Z : $\{z_{d,n}\}$, W : $\{w_{d,n}\}$, θ : $\{\theta_d\}$, ϕ : $\{\psi_k\}$

$n_{d,k,v}$: Number of {i: $w_{d,i} = v, Z_{d,i,k} = 1$}

$n_{d,k,:}$: $[n_{d,k,1}, ...n_{d,k,V}]$

$n_{d,k,.}$: $\sum_v n_{d,k,v}$ = Number of times Document d has topic k in it

$n_{.,k,v}$: $\sum_d n_{d,k,v}$= Number of times word v has topic k in all documents

GIBBS SAMPLING

LDA Model can be described in the following way where *Dir* is nothing but dirichilet distribution.

Model

$$\psi_k \sim Dir(\beta) \text{ for } i \in \{1, 2 ... K\}$$
$$\theta_d \sim Dir(\alpha) \text{ for } d \in \{1, 2 ... D\}$$
$$z_{d,n} \mid \theta_d \sim Discrete(\theta_d)$$
$$w_{d,n} \mid z_{d,n}, \phi_{z_{d,n}} \sim Discrete(\phi_{z_{d,n}})$$

TOPIC MODELLING

Our Primary Task in LDA is to find the following posterior

$$P(Z, \theta, \phi \mid W) = \frac{P(Z,\theta,\phi,W)}{P(W)}$$

From the above model detals joint can be further simplified as the following

TOPIC MODELLING

Joint Probability:

$$P(Z,\theta,\phi,W)=\prod_{d=1}^{D}p(\theta_d \mid \alpha_d)\prod_{d=1}^{D}\prod_{i=1}^{N_d}p(z_{d,i} \mid \theta_d)\prod_{d=1}^{D}p(\psi_k \mid \beta_k)\prod_{d=1}^{D}\prod_{i=1}^{N_d}p(w_{d,i} \mid z_{d,i},\phi)$$

$$P(Z,\theta,\phi,W)=\prod_{d=1}^{D}D(\theta_d;\alpha_d)\prod_{d=1}^{D}\prod_{i=1}^{N_d}(\Pi_{k=1}^{K}\theta_{d,k}^{Z_{d,i,k}})\prod_{d=1}^{D}D(\psi_k;\beta_k)\prod_{d=1}^{D}\prod_{i=1}^{N_d}(\Pi_{k=1}^{K}\psi_{k,w_{d,i}}^{Z_{d,i,k}}) \tag{1}$$

$$P(Z,\theta,\phi,W)=\prod_{d=1}^{D}\frac{\Pi_K\,\tau(\alpha)}{\tau(\sum_K\alpha)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha+n_{d,k,\cdot}-1}\prod_{k=1}^{K}\frac{\Pi_V\,\tau(\beta)}{\tau(\sum_V\beta)}\prod_{v=1}^{V}\psi_{d,k}^{\beta+n_{\cdot,k,v}-1} \tag{2}$$

TOPIC MODELLING

Marginalizing above intergral with respect to Z is difficult. So we calculate the posterior by using Gibbs Sampling. Where we conditionally update each variable iteratively.

$$P(\Theta) \sim P(\Theta \mid Z, W, \phi)$$

$$P(\phi) \sim P(\phi \mid Z, W, \theta)$$

$$P(Z) \sim P(Z \mid \Theta, W, \phi)$$

TOPIC MODELLING

- **We can integrate out the remaining parameters and update the hidden variable Z.This is called Collapsed Gibbs Sampling**

$$\psi_{k,t} = \frac{n_{.,k,v} + \beta}{\sum_{v=1}^{V}(n_{.,k,v} + \beta)}$$

$$\theta_{d,k} = \frac{n_{d,k,.} + \alpha}{\sum_{k=1}^{K}(n_{d,k,.} + \alpha)}$$

TOPIC MODELLING

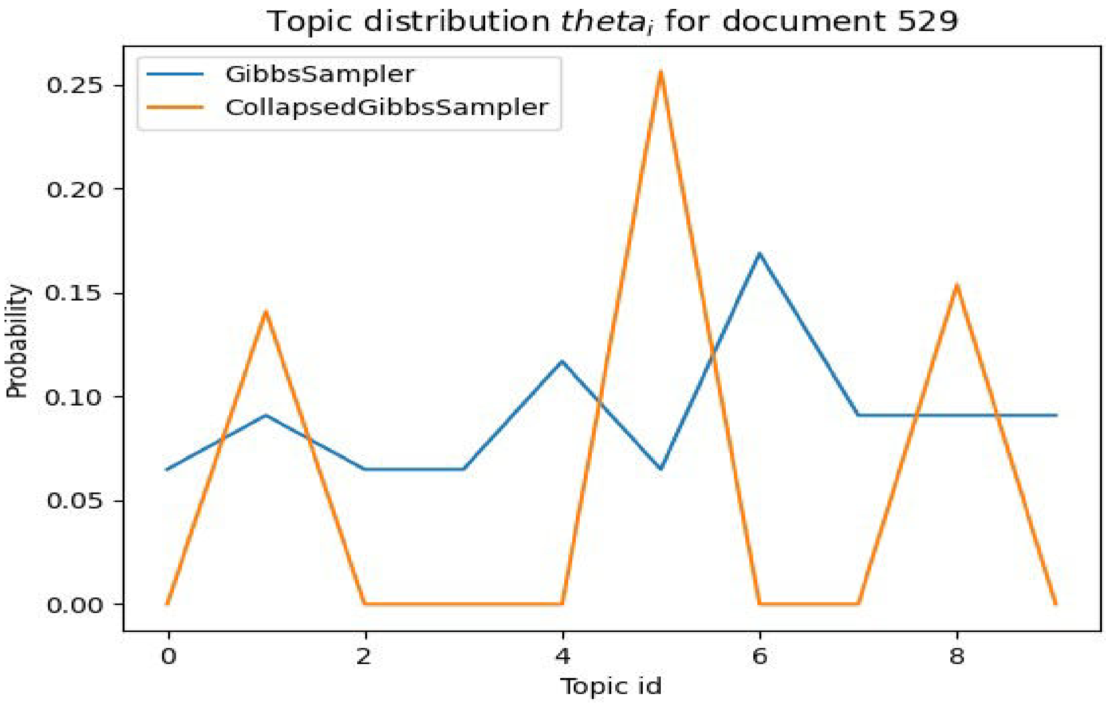
- **Collapsed Gibbs Sampler is faster than Normal Gibbs Sampler as we need not take all parameters into consideration**

Sampler	Number of iterations	Training Time(in sec)
Gibbs Sampler	500	1137
Collapsed Gibbs Sampler	500	830

Table 1: Traing times

TOPIC MODELLING

- **Gibbs Sampler Vs Collapsed Gibbs Sampler**



TOPIC MODELLING

- **Topics classified by Gibbs Sampler**

```
Topic #0: com space new available data 20 information use 10 aids
Topic #1: just know like people don year years car use time
Topic #2: server support edu supported os joseph david file readme vga
Topic #3: edu graphics pub mail ray send ftp objects server files
Topic #4: section firearm license military shall weapon person following means u
se
Topic #5: like know think problem windows use don just good does
Topic #6: like just key don good government people think encryption use
Topic #7: people god just don think like time good know israel
Topic #8: edu navy vote votes health mil misc car hp thomas
Topic #9: goal game finnish shot puck roger peter sweden good slave
```

TOPIC MODELLING

- **Topics classified by Collapsed Gibbs Sampler**

```
Topic #0: god people good brothers work did like jews just 12
Topic #1: just like don use people time good does want right
Topic #2: people just like don know use think time edu good
Topic #3: like people time just know good space use think don
Topic #4: bibles zyxel enjoy engine engineer engineered engineering engineers en
gines england
Topic #5: edu just graphics like people pub know new don mail
Topic #6: space like people don good does use just chip live
Topic #7: people seek order religion philosophy mail users like said bit
Topic #8: edu graphics like good think don just know time mail
Topic #9: zyxel eng engine engineer engineered engineering engineers engines eng
land english
```

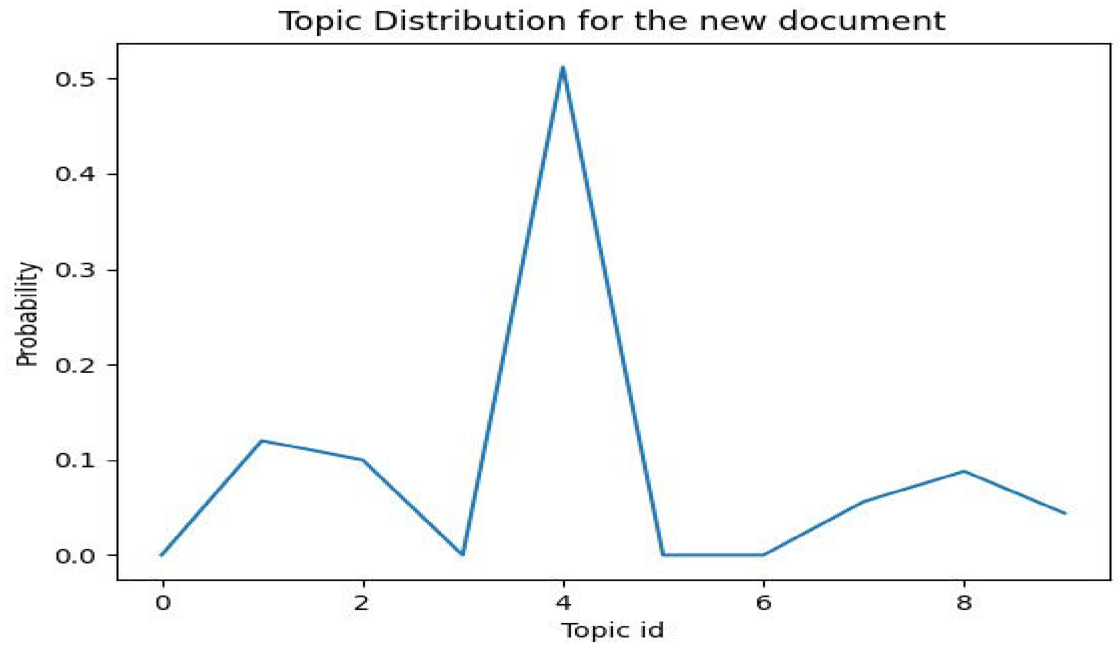
TOPIC MODELLING

- <u>**Querying:**</u>

  Estimating topic proportions of unseen or a new document. We do this by updating Z using Gibbs Sampling and then update the topic proportions of the new document according to our new sampled Z vector.

TOPIC MODELLING

- **Topic proportions of a new set of words classified by Collapsed Gibbs Sampler.**



FURTHER SCOPE

- **Implementing Topic Modelling using mixture models and doing inference using Gibbs Sampling. Comparing the inference results over Mixture Models and LDA.**

- **Analysis of our topic model for similarity checking.**

WHAT DID I LEARN

- Methods like Steepest Descent, Conjugate Gradient
- Multivariate Normal Distributions, Sampling stratagies from a Multivariate Normal.
- Gibbs Sampling
- MLE, MAP and Bayesian Estimators
- Conjugate Priors and Exponential Family Distributions mainly Dirichilet distribution
- Topic Modelling specifically LDA

THANK YOU