Machine Learning: Data to Models
Assignment 3b

Qun Gao, JHED ID qgao6:
Li-Yi Lin, JHED ID: llin34

190

4 Blocked Gibbs Sampler

4.1 Analysis Questions

From the Appendix, we know that the collapsed joint distribution is

$$P(\mathbf{z}, \mathbf{x}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda) = P(\mathbf{w}|\mathbf{z}, \mathbf{x}, \mathbf{c}, \beta)P(\mathbf{z}|\alpha)P(\mathbf{x}|\lambda)$$

$$= \prod_k (\frac{\Gamma(\sum_w \beta)}{\Gamma(\sum_w n_w^k + \beta)} \prod_w \frac{\Gamma(n_w^k + \beta)}{\Gamma(\beta)}) \times \prod_c \prod_k (\frac{\Gamma(\sum_w \beta)}{\Gamma(\sum_w n_w^{c,k} + \beta)} \prod_w \frac{\Gamma(n_w^{c,k} + \beta)}{\Gamma(\beta)})$$

$$\times \prod_d (\frac{\Gamma(\sum_k \alpha)}{\Gamma(\sum_k n_k^d + \alpha)} \prod_k \frac{\Gamma(n_k^d + \alpha)}{\Gamma(\alpha)}) \times \prod_{(d,i)} P(x_{d,i}|\lambda)$$

To sample $z_{d,i}, x_{d,i}$ at one time, we derive

$$P(z_{d,i}, x_{d,i}|\mathbf{z} - z_{d,i}, \mathbf{x} - x_{d,i}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda) = \frac{P(\mathbf{z}, \mathbf{x}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda)}{P(\mathbf{z} - z_{d,i}, \mathbf{x} - x_{d,i}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda)}$$

According to similar derivation as in the appendix, we have

$$P(z_{d,i} = k, x_{d,i} = 0|\mathbf{z} - z_{d,i}, \mathbf{x} - x_{d,i}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda) = \frac{\frac{\Gamma(n^k_{w_{d,i}}+1+\beta)}{\Gamma(1+\sum_w n_w^k+\beta)} \frac{\Gamma(n_k^d+1+\alpha)}{\Gamma(1+\sum_{k'} n_{k'}^d+\alpha)}}{\frac{\Gamma(n^k_{w_{d,i}}+\beta)}{\Gamma(\sum_w n_w^k+\beta)} \frac{\Gamma(n_k^d+\alpha)}{\Gamma(\sum_{k'} n_{k'}^d+\alpha)}} P(x = 0|\lambda)$$

$$= \frac{n_k^d + \alpha}{n_\star^d + K\alpha} \frac{n_w^k + \beta}{n_\star^k + V\beta}(1 - \lambda)$$

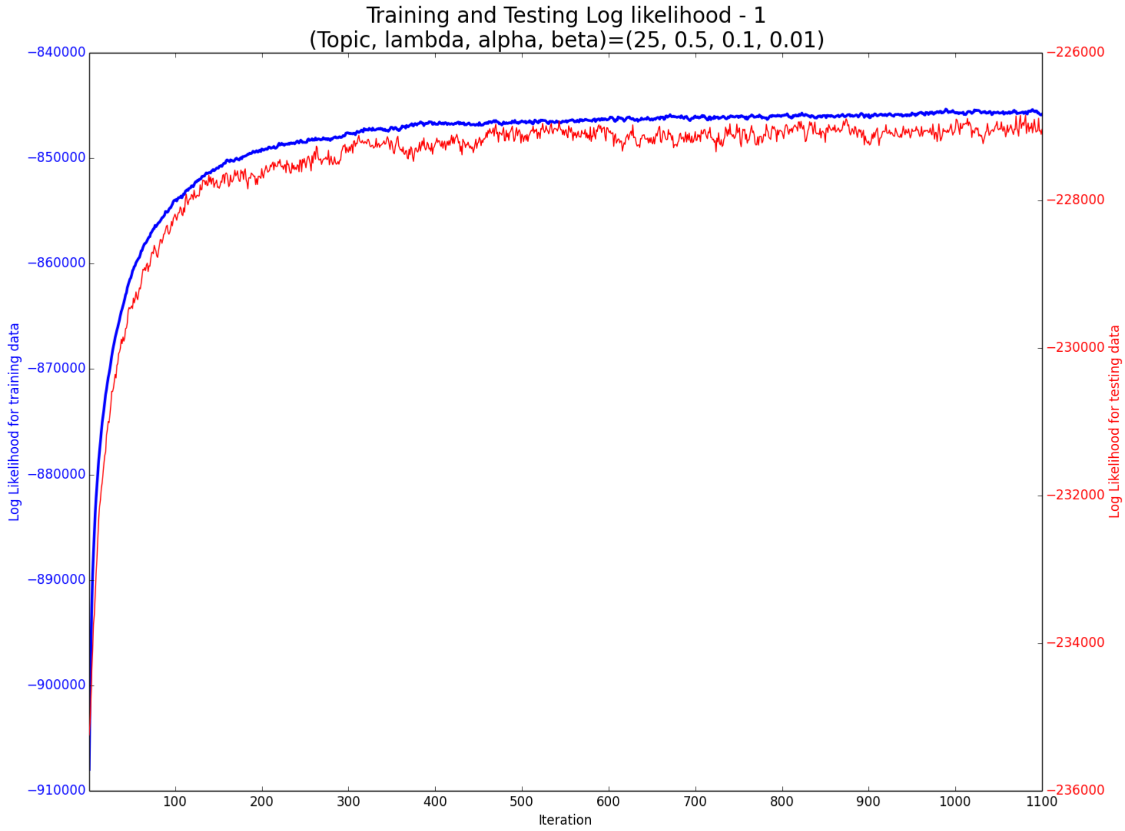
$$P(z_{d,i} = k, x_{d,i} = 1|\mathbf{z} - z_{d,i}, \mathbf{x} - x_{d,i}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \lambda) = \frac{\frac{\Gamma(n^{c,k}_{w_{d,i}}+1+\beta)}{\Gamma(1+\sum_w n_w^{c,k}+\beta)} \frac{\Gamma(n_k^d+1+\alpha)}{\Gamma(1+\sum_{k'} n_{k'}^d+\alpha)}}{\frac{\Gamma(n^{c,k}_{w_{d,i}}+\beta)}{\Gamma(\sum_w n_w^{c,k}+\beta)} \frac{\Gamma(n_k^d+\alpha)}{\Gamma(\sum_{k'} n_{k'}^d+\alpha)}} P(x = 1|\lambda)$$

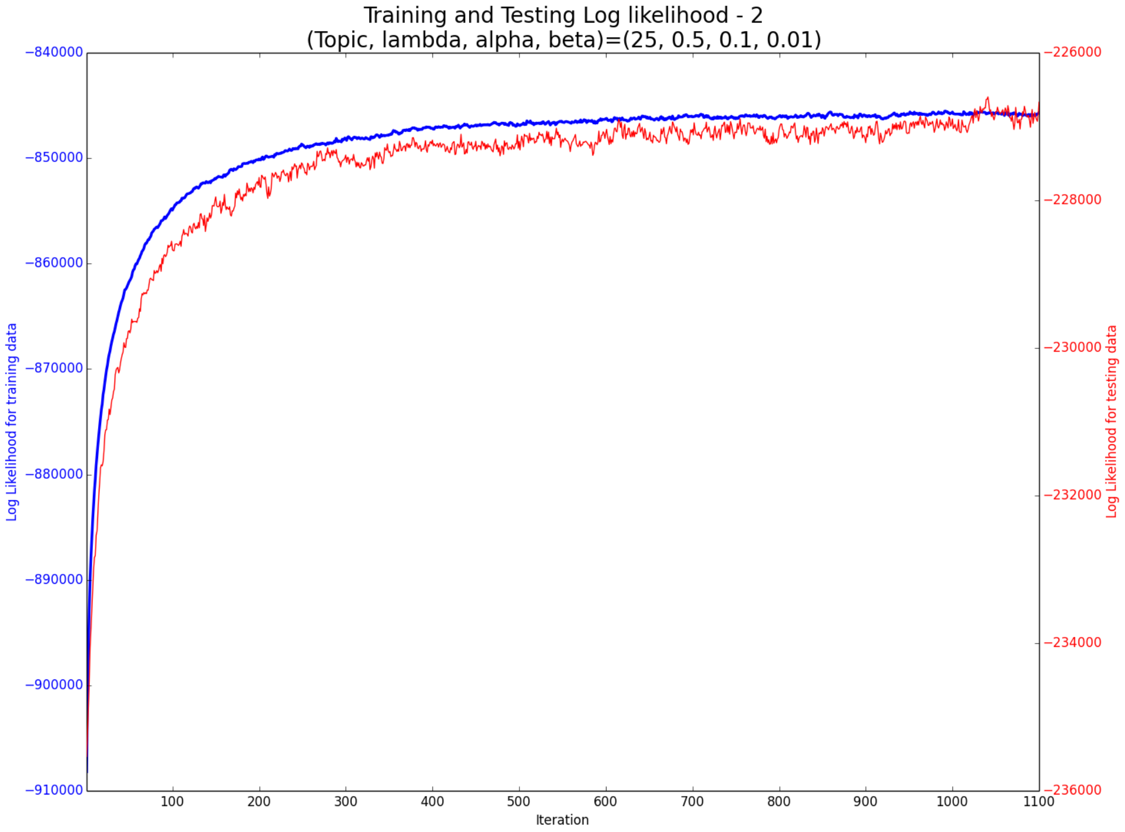
$$= \frac{n_k^d + \alpha}{n_\star^d + K\alpha} \frac{n_w^{c,k} + \beta}{n_\star^{c,k} + V\beta}\lambda$$

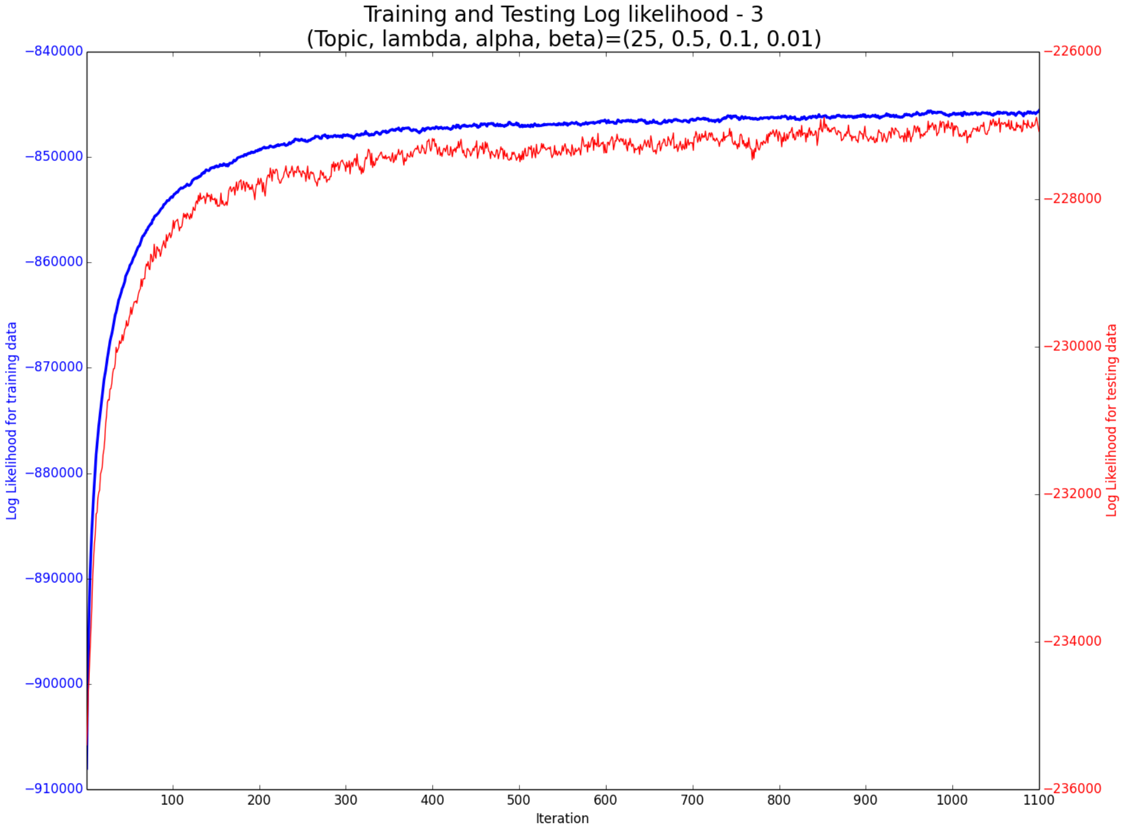
5 Text Analysis with MCLDA

5.1 Empirical Questions

1.

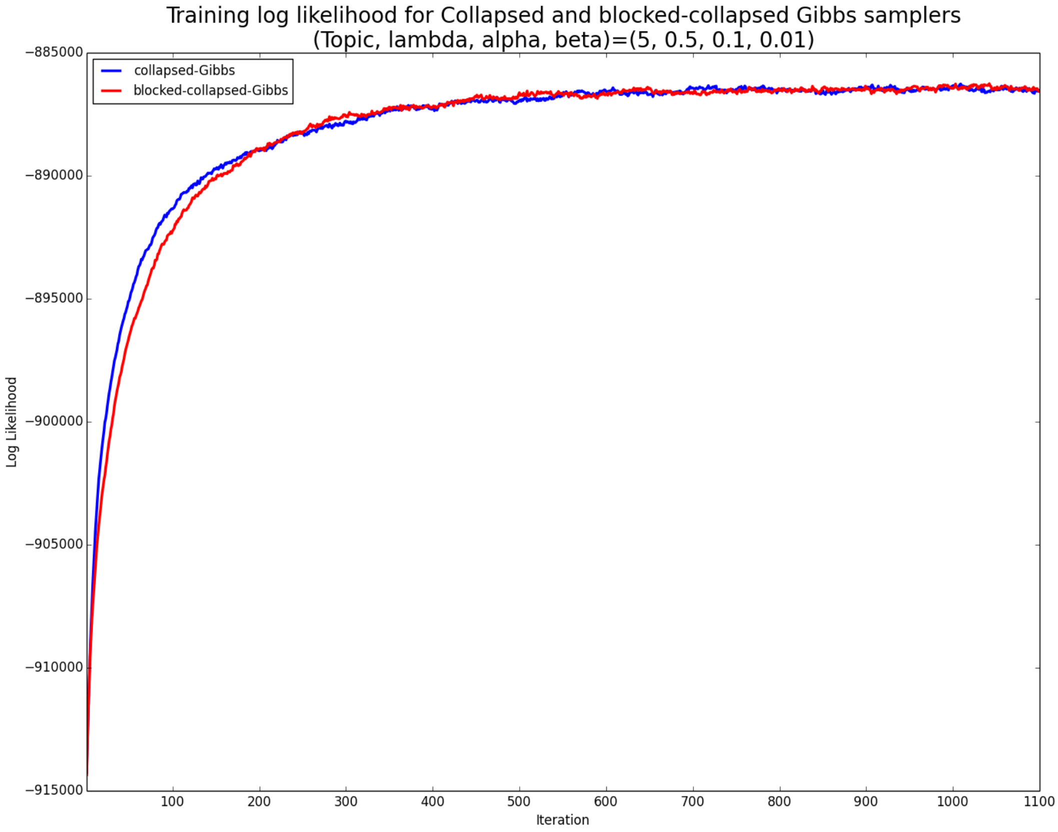


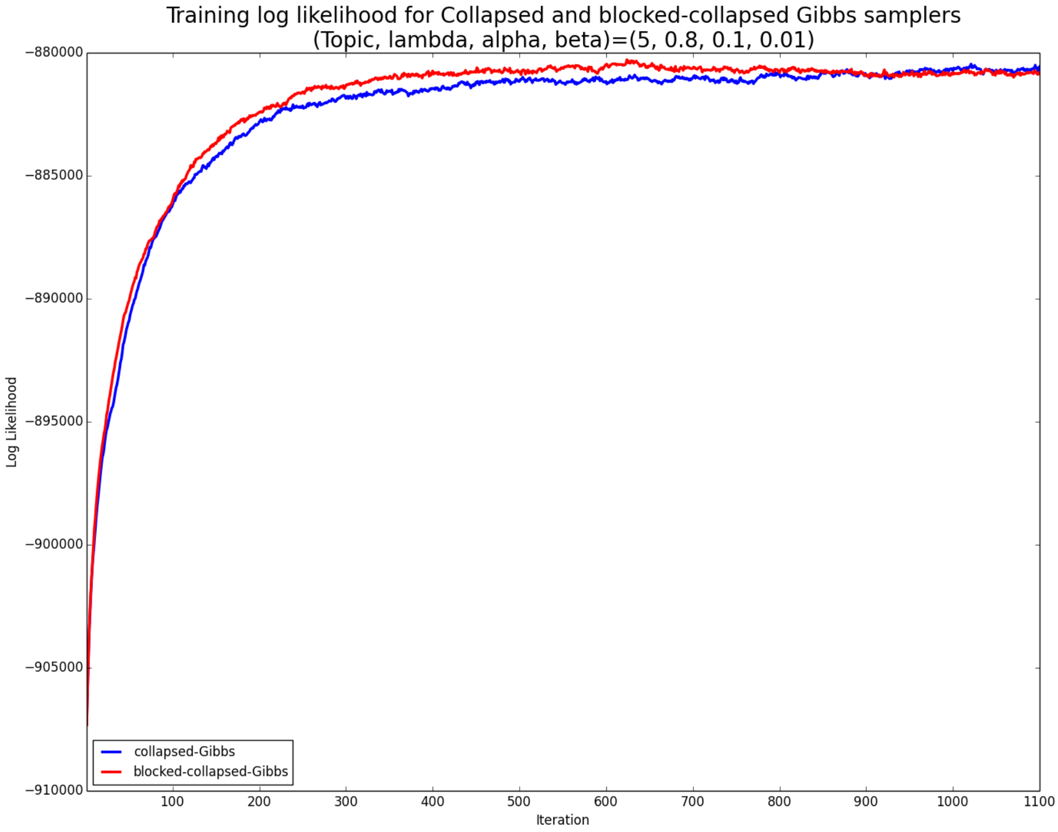


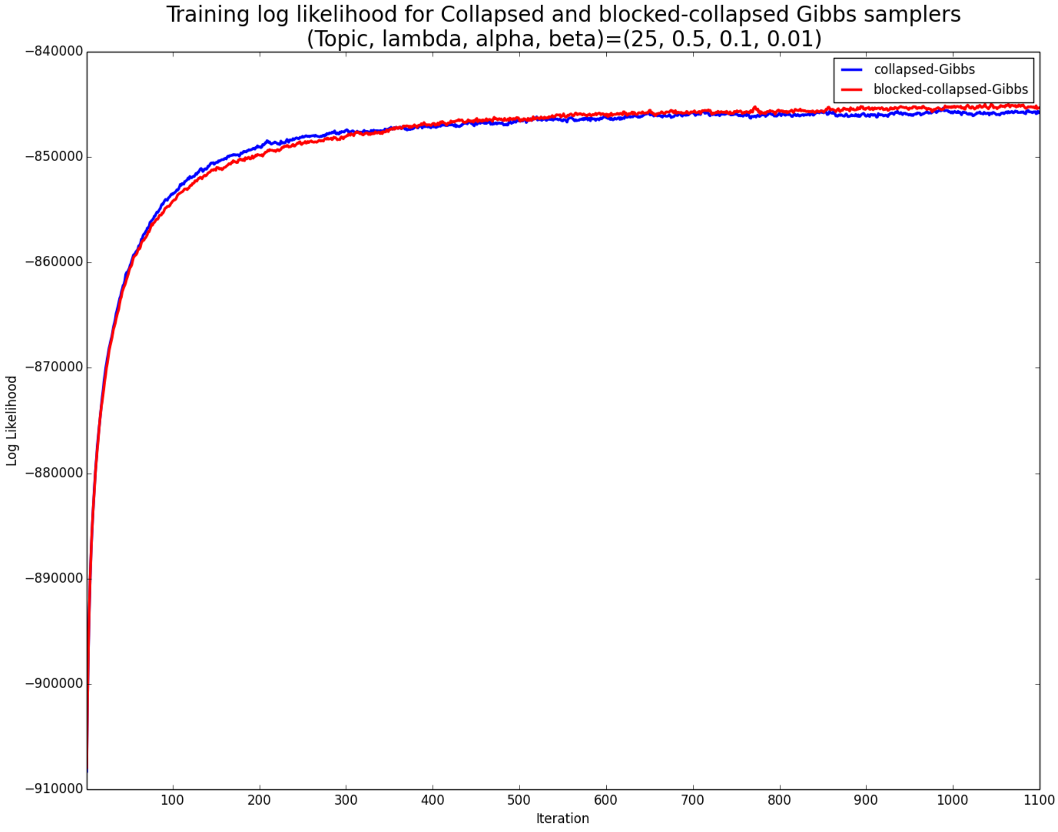


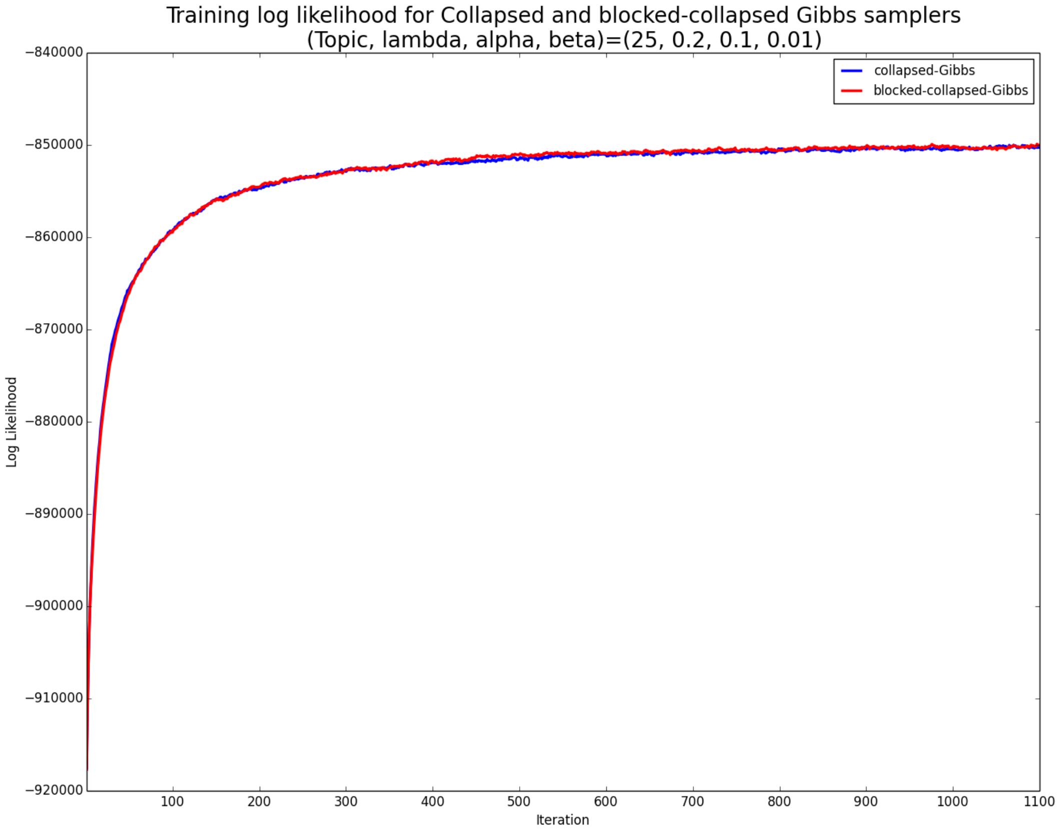
From the three plots we observed that the trend of the training and test likelihood over the iteration are almost the same. They both converge to a steady state. But the test likelihood is larger than the training one.

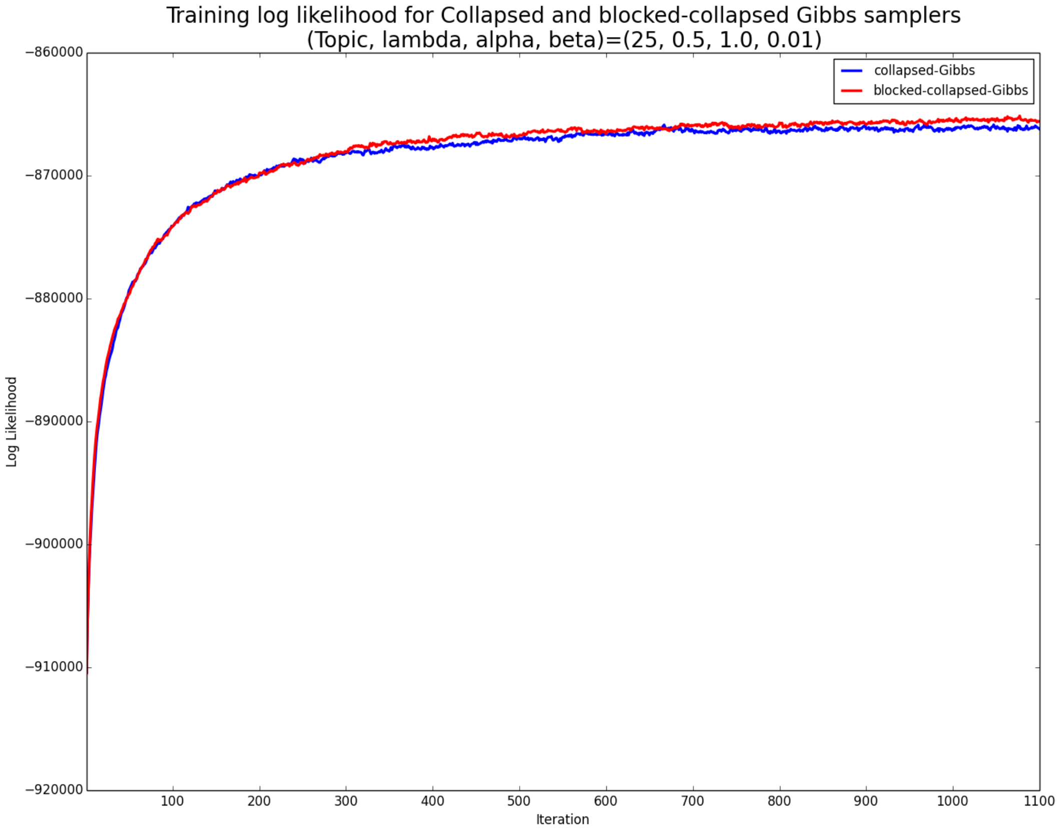
2.


Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers
(Topic, lambda, alpha, beta)=(5, 0.5, 0.1, 0.01)
collapsed-Gibbs
blocked-collapsed-Gibbs
Log Likelihood
−885000
−890000
−895000
−900000
−905000
−910000
−915000
100
200
300
400
500
600
700
800
900
1000
1100
Iteration


Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers
(Topic, lambda, alpha, beta)=(5, 0.8, 0.1, 0.01)
Log Likelihood
−880000
−885000
−890000
−895000
−900000
−905000
−910000
100
200
300
400
500
600
700
800
900
1000
1100
Iteration
collapsed-Gibbs
blocked-collapsed-Gibbs

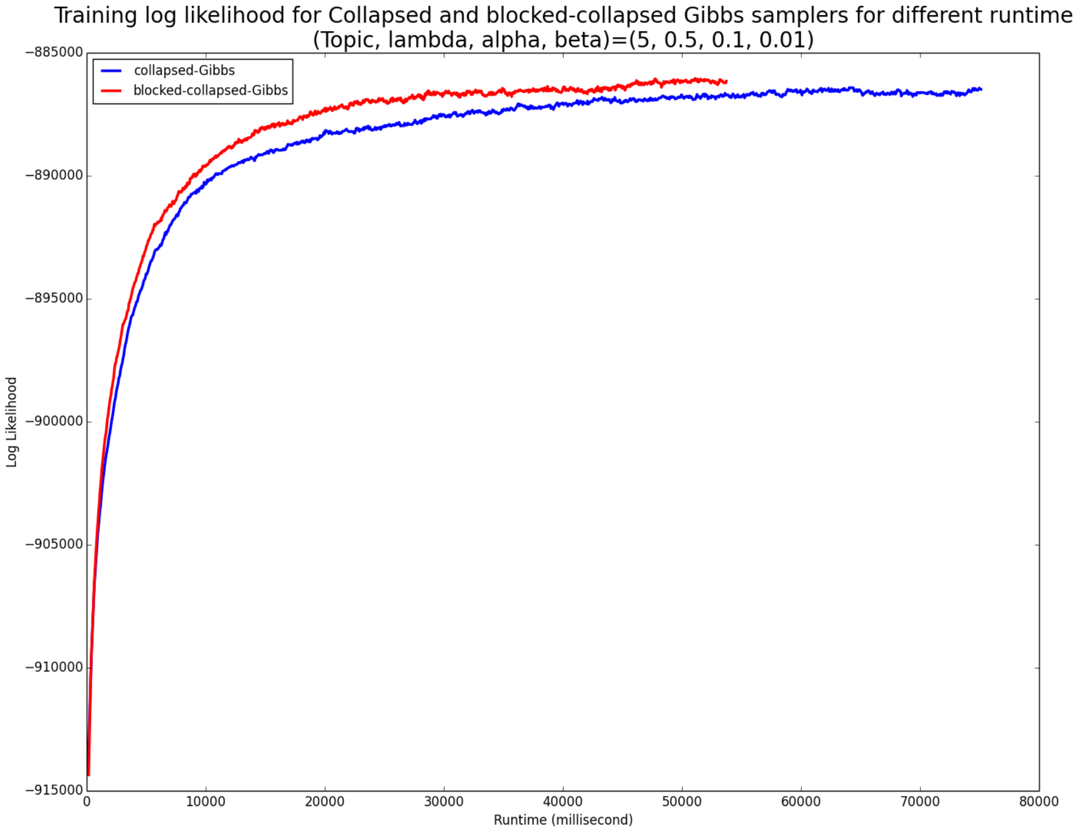

Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers
(Topic, lambda, alpha, beta)=(25, 0.5, 0.1, 0.01)
Log Likelihood
−840000
−850000
−860000
−870000
−880000
−890000
−900000
−910000
100
200
300
400
500
600
700
800
900
1000
1100
Iteration
collapsed-Gibbs
blocked-collapsed-Gibbs

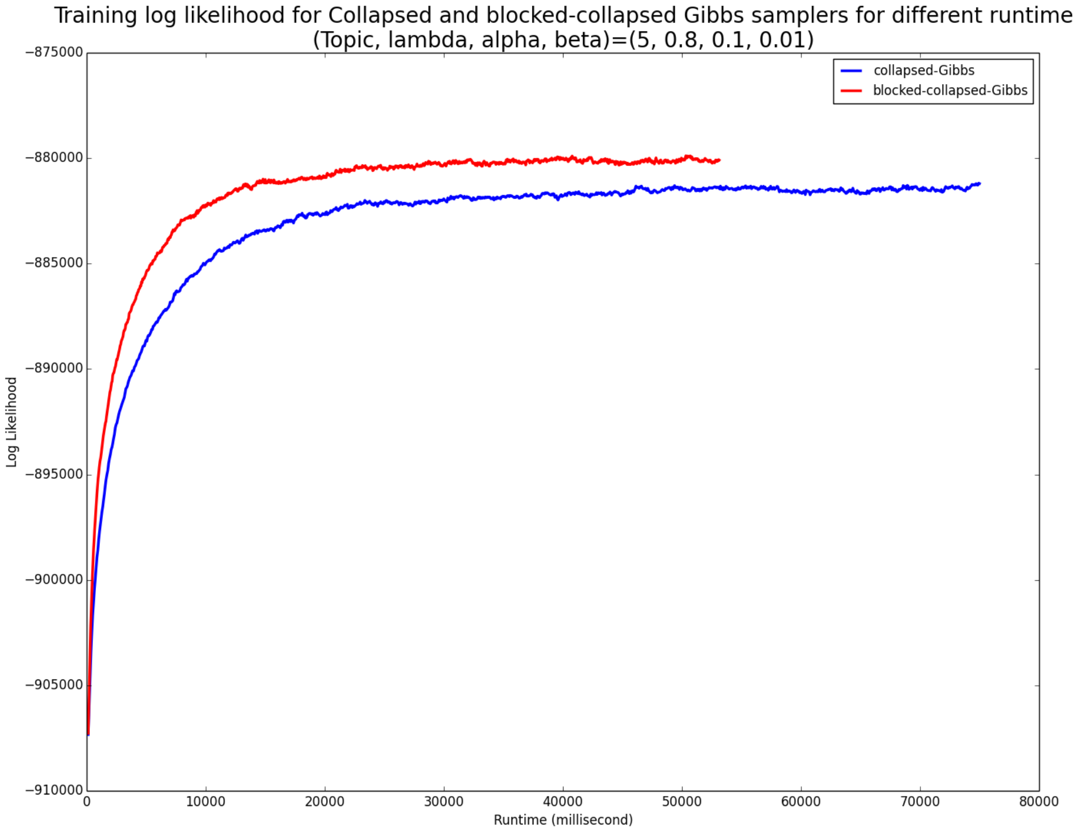


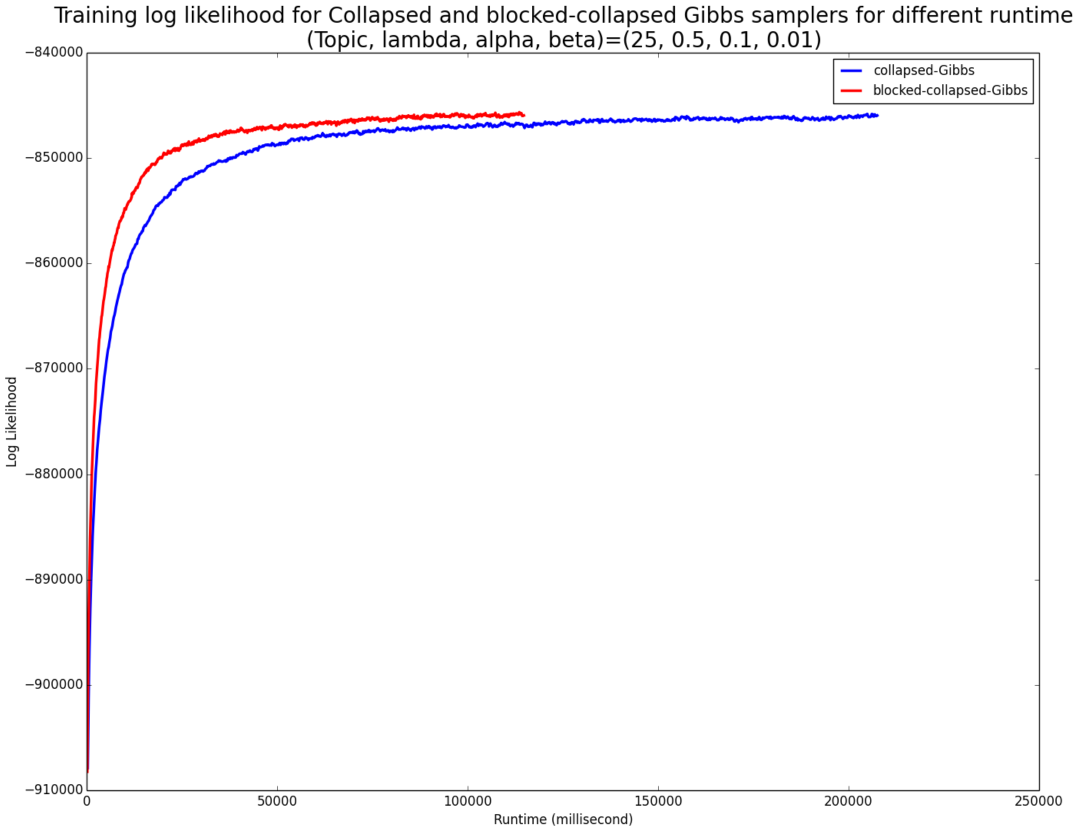


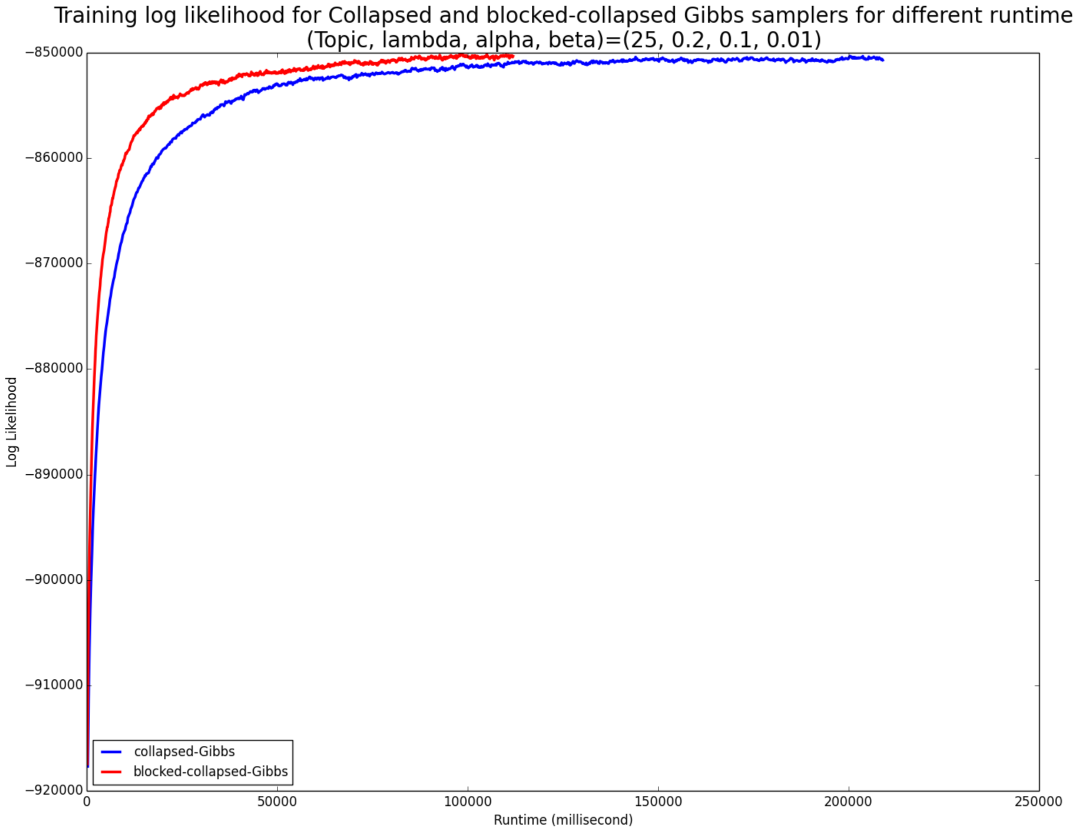
From the plot we can see bolcked collapsed sampler seems to mix faster.

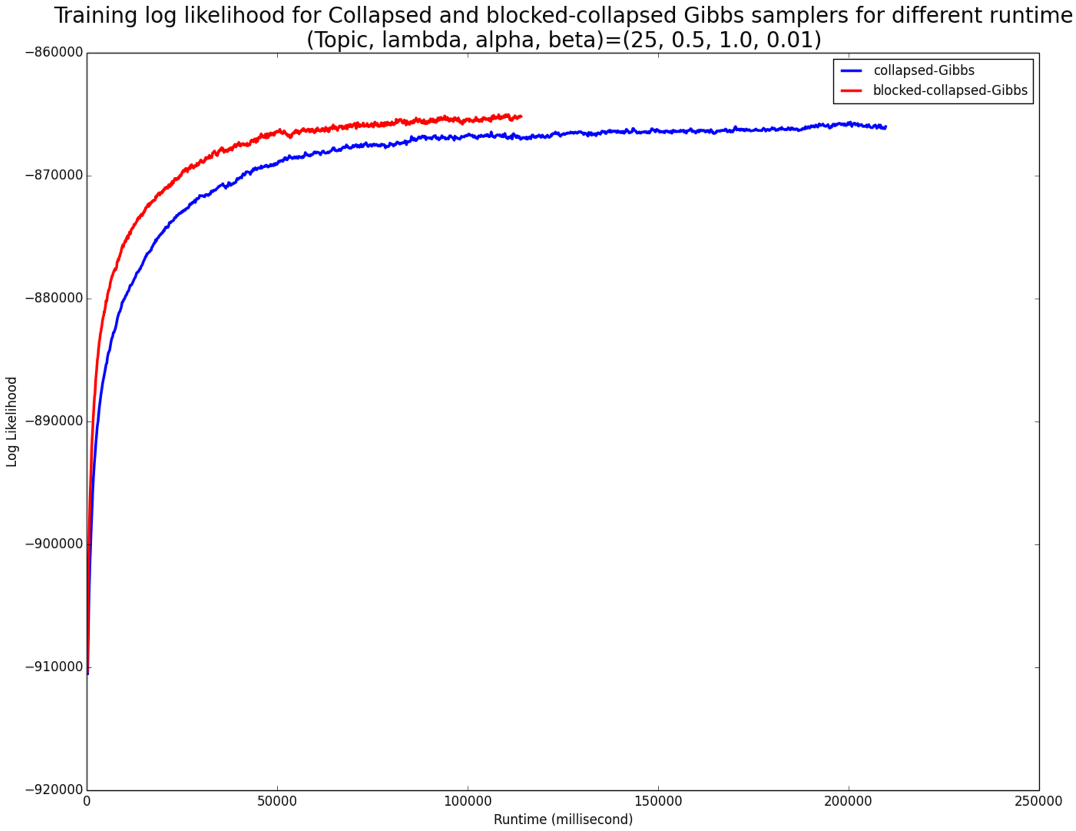
3.


Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers for different runtime
(Topic, lambda, alpha, beta)=(5, 0.5, 0.1, 0.01)
collapsed-Gibbs
blocked-collapsed-Gibbs
Log Likelihood
−885000
−890000
−895000
−900000
−905000
−910000
−915000
0
10000
20000
30000
40000
50000
60000
70000
80000
Runtime (millisecond)


Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers for different runtime
(Topic, lambda, alpha, beta)=(5, 0.8, 0.1, 0.01)
collapsed-Gibbs
blocked-collapsed-Gibbs
Log Likelihood
−875000
−880000
−885000
−890000
−895000
−900000
−905000
−910000
0
10000
20000
30000
40000
50000
60000
70000
80000
Runtime (millisecond)

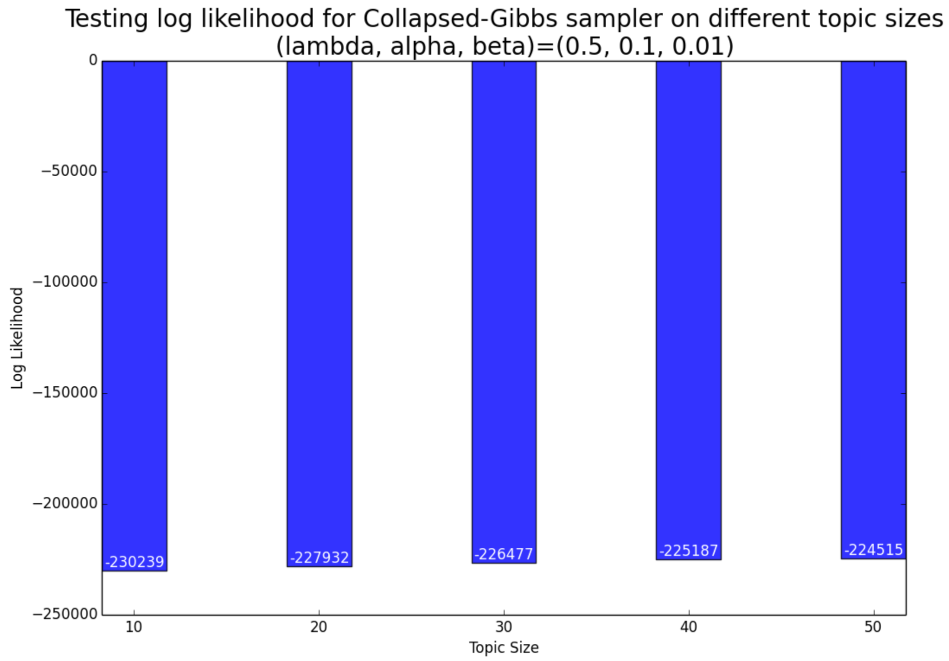

Training log likelihood for Collapsed and blocked-collapsed Gibbs samplers for different runtime
(Topic, lambda, alpha, beta)=(25, 0.5, 0.1, 0.01)
collapsed-Gibbs
blocked-collapsed-Gibbs
Log Likelihood
−840000
−850000
−860000
−870000
−880000
−890000
−900000
−910000
0
50000
100000
150000
200000
250000
Runtime (millisecond)



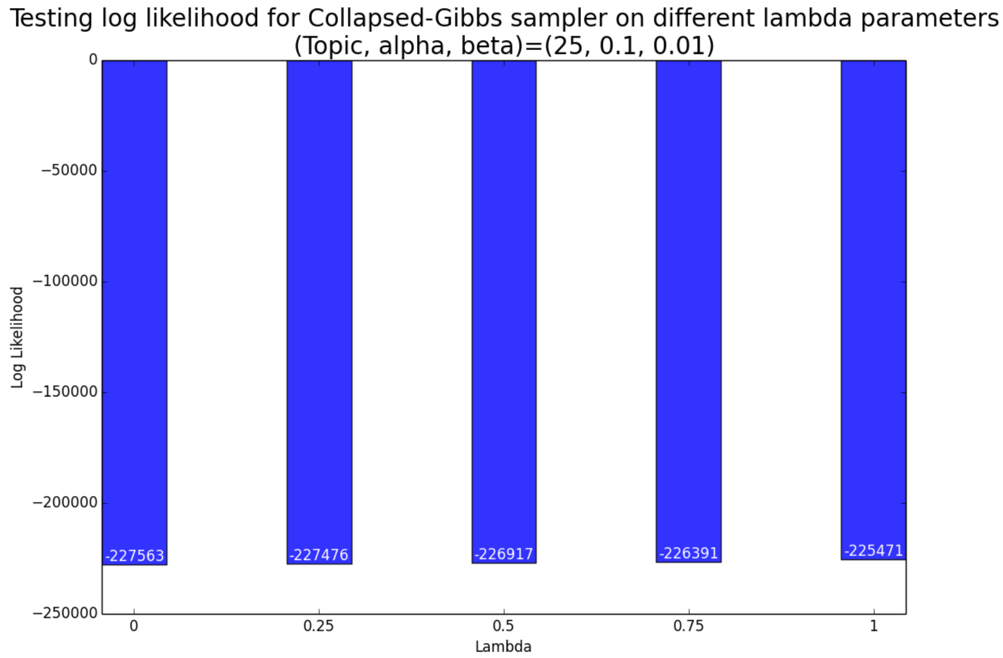


From the plot we can see bolcked collapsed sampler is faster to reach a stable point.

4.



5.



6.

(a) When we use 10 topics, we find that for topic 4 there are 11 common words appeared for all three word distributions. In this case we conclude that for topic 4 the global, NIPS and ACL word distributions could have a common theme. However for topic 0 only 3 words appeared in common for all distributions, ”information”,”paper” and ”analysis”. These are very common words in scientific documents and couldn’t form a theme. So for this topic 0 the three word distributions may be very different.

We can also look at this with 15 topics. We find that for topic 3 there are 16 common words among the three! We can guess the common theme is about recognition problem with classification method. While for topic 8 only ”model” and ”models” are common words so this topic index appears very different among the three word distributions. For 25 topics we can also detect that topic 18 among the three shares a common theme (maybe also about classification) and topic 6 is very different among the three.

(b) We can see that for $\lambda = 1$ the global and corpus-specific topics are very different. For example, topic 0 in the three distributions have no common words. While for $\lambda = 0.25$ the global and corpus-specific topic are very similar as they have many common words. So we conclude that for smaller values of λ the global and corpus-specific topics become more similar while the larger one being the opposite. This is reasonable because from the definition of the parameters, words are more possible to be sampled from global topic distribution if λ is small, so topics can be similar. And if λ is large, words are sampled according to collection-specific distribution, which results in different topics with high probability.

(c) For $\alpha = 0.001$, a word appears more frequently in different topics while for $\alpha = 10$ it appears less frequently. For example, "network" appears 2 times for $\alpha = 10$ and appears 13 times for $\alpha = 0.001$. This is because α is the hyperparameter of θ, which is Dirichlet distributed. For smaller α, the Dirichlet distribution is more smooth, and since θ is the mixing proportion of the topics, so a word could appear more frequently in all topics. In the opposite, for larger α , the Dirichlet distribution is not flat so a word may only appear in very few topics. The case is similar for β, i.e., for larger β a word appears less frequently across topics and for smaller β it appears more frequently. The reason is similar because β is the parameter of ϕ, which is Dirichlet distributed also.