# Parameter Dimensionality

## 1 Data Simulation

We simulated a sparse dataset using LDA generative process. Parameter settings are as follows: the corpus is a collection of 10 documents and the vocabulary has 20 unique words. The number of words per document follows Poisson distribution with parameter 100. Each document is a mixture of 3 topics and the Dirichlet prior $\alpha$ for topics proportions is 0.1. we fitted the LDA model over 100 different random seeds and used $\alpha = 0.001$.

Similar to the analysis of dataset1, for some random seeds the loglikelihoods converge to a optimal value very fast, then remain stable. and we want to look at the seeds where the loglikelihoods does not converge. so firstly we fitted the lda model for 1000 iterations over 100 random seeds. ($seeds$ : $0, 200, 400, \ldots, 19800$) The left panel of fig.1 is the plot of log-likelihoods and the right panel is a closer look at the differences, there are 2 clusters. Lets assume for now the optimal limit is around (-480,-520). For the set of random seeds (named as seedset2) whose loglikelihoods are less than -520, we increased the number of iterations to 2000.

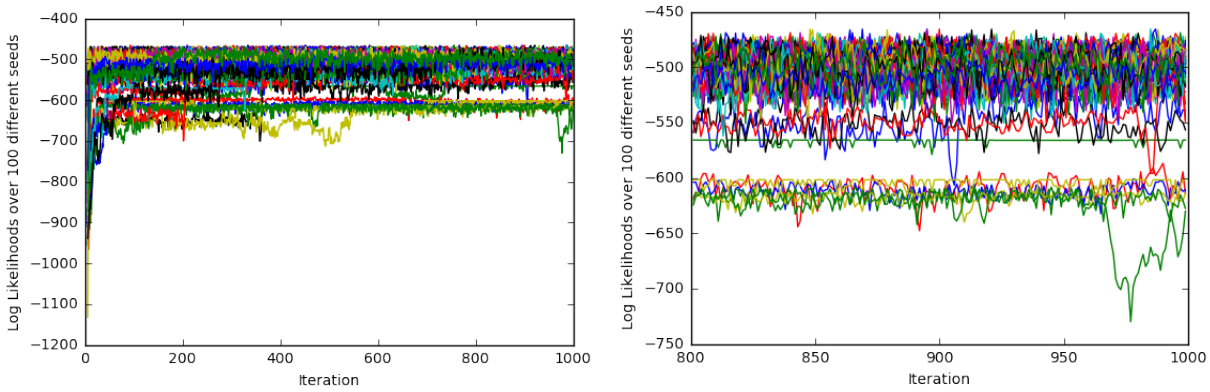


Figure 1: The log-likelihoods plot over 100 random seeds.

The left panel of figure 3 is a plot of loglikelihoods for 2000 interations over seedset2, and the right panel is a closer look at the differences, again there are two clusters. The random seeds used in the higher cluster are [0, 200, 2600, 7800, 13000, 15800](seedset2.1), the random seeds used in the lower cluster are [400, 2800, 3800, 4600, 5800, 12200, 12800, 13000](seedset2.2).

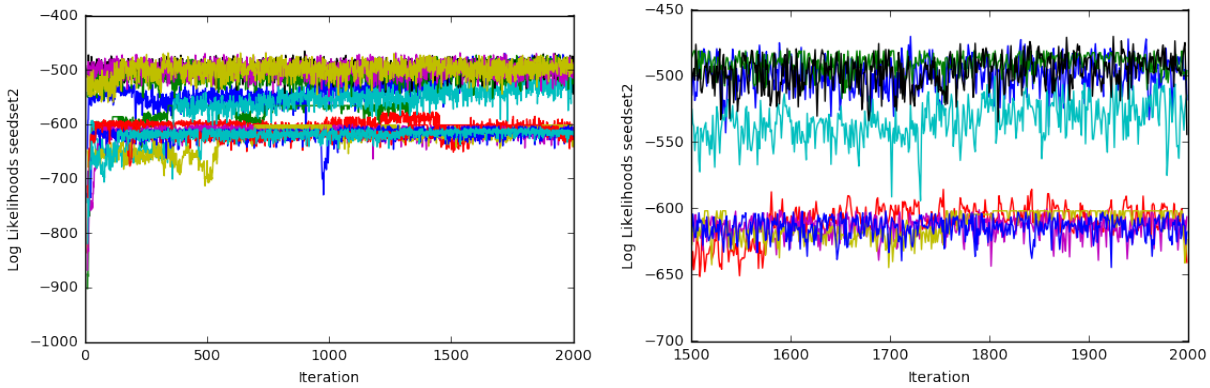


Figure 2: The log-likelihoods plot over seedset2. seedset2= (0, 200, 400, 2600, 2800, 3800, 4600, 5800, 7800, 12200, 12800, 13000, 15800)

## 1.1 seedset2.1

The left panel of figure 4 is a plot of loglikelihoods over seedset2.1, they all converge to one limit. we can observe that for some seeds in seedset2.1, the loglikelihoods jump from the lower cluster to the higher one (optimal range) at certain iteration. so to examine what happens at those jumps, we listed the all topics over 3000 iterations for each seed in seedset2.1. and we discovered a common topic assignment for all seeds.(see $\text{draft}_l da_s ampling.py formore information$)

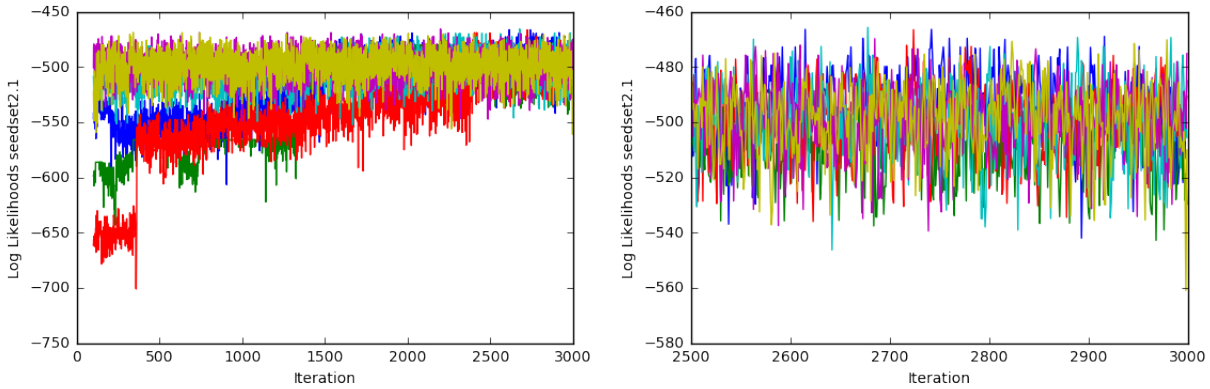


Figure 3: The log-likelihoods plot over seedset2.1. seedset2.1= [0, 200, 2600, 7800, 13000, 15800]

## 1.2 seedset2.2

for seedset2.2, we increased the number of iterations to 6000, and plot its loglikelihoods (figure 5). Again we observed 2 clusters. The set seedset221 contains random seeds [400,4600,5800] and the set seedset222 contains seeds [2800,3800,12200,12800]. we listed all topic assignments for 3 random seeds over 6000 iterations in seedset2.2.1. and same topic assignment in section2.1 is observed again.

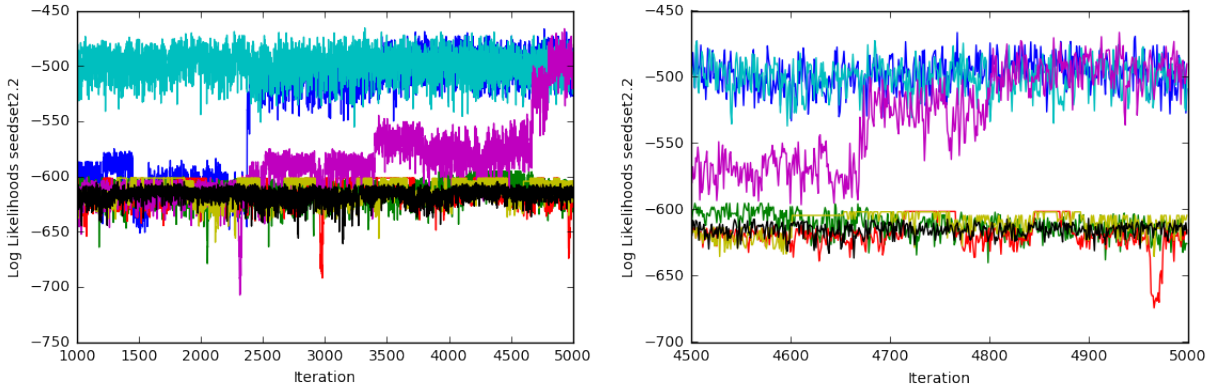


Figure 4: The log-likelihoods plot over seedset2.2. seedset2.2= [400, 2800, 3800, 4600, 5800, 12200, 12800, 13000]