**CUED - Engineering Tripos Part IIB 2024-2025** **Module Coursework**

| Module | 4F13 | Title of report | Latent Dirichlet Allocation Model |
|---|---|---|---|
| Date submitted: 5/12/24 | | Assessment for this module is ☑ 100% / ☐ 25% coursework of which this assignment forms 33 % | |

| UNDERGRADUATE and POST GRADUATE STUDENTS | | |
|---|---|---|
| Candidate number: | 5488A | ☑ Undergraduate<br>☐ Post graduate |

| Feedback to the student<br>☐ **See also comments in the text** | | Very good | **Good** | Needs improvmt |
|---|---|---|---|---|
| CONTENT | **Completeness, quantity of content:**<br>Has the report covered all aspects of the lab? Has the analysis been carried out thoroughly? | | | |
| | **Correctness, quality of content**<br>Is the data correct? Is the analysis of the data correct? Are the conclusions correct? | | | |
| | **Depth of understanding, quality of discussion**<br>Does the report show a good technical understanding? Have all the relevant conclusions been drawn? | | | |
| | Comments: | | | |
| PRESENTATION | **Attention to detail, typesetting and typographical errors**<br>Is the report free of typographical errors? Are the figures/tables/references presented professionally? | | | |
| | Comments: | | | |

Marker: Date:

# 4F13 Coursework 3 - Latent Dirichlet Allocation

December 2024

## 1 Task A

We start by using a single bag of words model to represent the entire training dataset, $\mathcal{A}$. The likelihood, Equation 1, is a multinomial over the $M$ words in the vocabulary. $k_m$ is the count of word $m$ in the training dataset and $pi_m$ is its probability. We collect these elements into vectors $\boldsymbol{k}$ and $\boldsymbol{\pi}$ respectively.

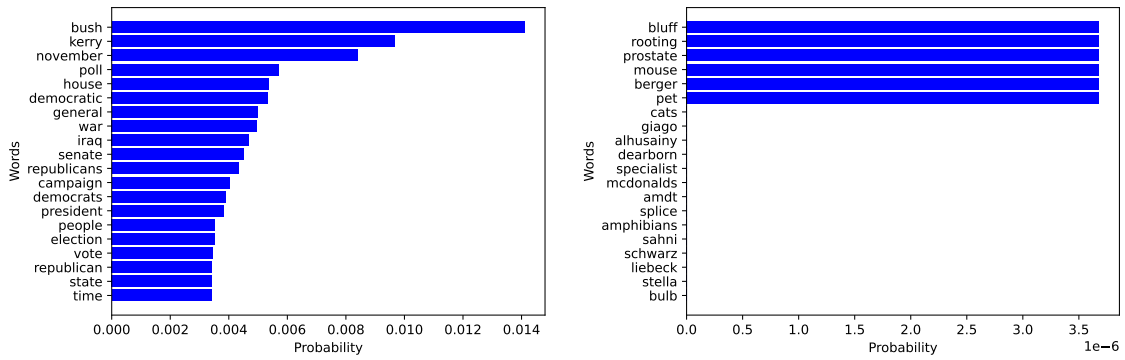
$$P(\mathcal{A}|\boldsymbol{\pi}) = N! \prod_{m=1}^{M} \frac{\pi_m^{k_m}}{k_m!} \tag{1}$$

To compute a maximum likelihood estimate we can maximise the log likelihood, for easier derivative computation, with a Lagrange multiplier to enforce the constraint that the probabilities sum to 1. We find that the maximum likelihood estimate for $\pi_m$, Equation 2, is simply the frequency of word $m$ in the training dataset.

$$\begin{aligned}
\mathcal{L} &= \log(P(\mathcal{A}|\boldsymbol{\pi})) + \lambda(\sum_{m=1}^{M} \pi_m - 1) \\
\left.\frac{\partial \mathcal{L}}{\partial \pi_i}\right|_{\pi_i^{ML}} &= \frac{k_i}{\pi_i^{ML}} + \lambda^{ML} = 0 \quad \therefore \pi_i^{ML} = -\frac{k_i}{\lambda^{ML}} \\
\sum_{m=1}^{M} \pi_m^{ML} &= \sum_{m=1}^{M} -\frac{k_m}{\lambda^{ML}} = 1 \quad \therefore \lambda^{ML} = -\sum_{m=1}^{M} k_m \\
\therefore \pi_i^{ML} &= \frac{k_i}{\sum_{m=1}^{M} k_m}
\end{aligned} \tag{2}$$

We see in Figure 1 that the most likely word is 'bush' with a probability of 0.0141 and there are multiple words in the vocabulary that do not appear in the training dataset and therefore have a probability of 0. Consider a test dataset, $\mathcal{B}$, with length $N_{\mathcal{B}}$. The test set of maximum likelihood would be simply be $N_{\mathcal{B}}$ copies of the most likely word in $\boldsymbol{\pi}^{ML}$, 'bush' when trained on $\mathcal{A}$. The log likelihood of this test set would be $N_{\mathcal{B}} \log(\pi_{\text{bush}}^{ML}) = -4.26 N_{\mathcal{B}}$.

Furthermore we notice that the minimum likelihood test set would be any test set that contains words not in the training dataset. This is because the probability of any word not in the training dataset is 0 and therefore the log likelihood of any test set containing such a word would be $-\infty$. This is a limitation of the ML estimate model as no valid test set should have probability 0.




(a) 20 most likely words.

(b) 20 least likely words

Figure 1: Maximum likelihood estimates of multinomial probabilities from dataset $\mathcal{A}$.

## 2 Task B

To combat this issue we can perform Bayesian inference. We assume a symmetric Dirichlet prior for our probability vector $\boldsymbol{\pi}$, Equation 3, with concentration parameter $a$.

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} = \mathbf{1}a) = \frac{1}{\text{B}(\mathbf{1}a)} \prod_{m=1}^{M} \pi_m^{a-1} \tag{3}$$

Using the same likelihood, Equation 1, we find that the posterior distribution of $\boldsymbol{\pi}$ is also a Dirichlet distribution, Equation 4, with parameter $\boldsymbol{\alpha} = \boldsymbol{k} + \mathbf{1}a$.

$$\begin{aligned} p(\boldsymbol{\pi}|\mathcal{A}) &\propto p(\mathcal{A}|\boldsymbol{\pi})p(\boldsymbol{\pi}) \propto \prod_{m=1}^{M} \pi_m^{k_m+a-1} \\ \therefore p(\boldsymbol{\pi}|\mathcal{A}) &= \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} = \boldsymbol{k} + \mathbf{1}a) \end{aligned} \tag{4}$$

We compute the predictive word probability, Equation 5, and notice that it is similar to the maximum likelihood estimate but with an initial pseudo count of $a$ for each word. This means that the probability of a words not seen in the training dataset now have non-zero probability, fixing the flaw of the maximum likelihood estimate.

$$\begin{aligned} p(w = i|\mathcal{A}) &= \int_{|\boldsymbol{\pi}|_1=1} p(w = i|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathcal{A})d\boldsymbol{\pi} \\ &= \int_0^1 \pi_i p(\pi_i|\mathcal{A})d\pi_i \\ &= \mathbb{E}_{\pi_i|\mathcal{A}}[\pi_i] \\ &= \frac{a + k_i}{Ma + \sum_{m=1}^{M} k_m} = \hat{\pi}_i \end{aligned} \tag{5}$$

For small values of $a$ predictive probabilities with the Bayesian estimate approach those of the maximum likelihood estimate. As $a$ increases words with probability greater than $\frac{1}{M}$ will decrease in probability and words with probability less than $\frac{1}{M}$ will increase in probability. This is because for large values of $a$ our observations become insignificant compared to the prior and the posterior distribution will approach the prior distribution.

We have added a hyper-parameter $a$ to our model which we wish to set in an analytical way. We can do this by maximising the marginal likelihood of the training dataset, $\mathcal{A}$, Equation 6. We plot the log marginal likelihood for different values of $a$ in Figure 2 and find that the maximum marginal likelihood occurs at $a = 0.753$. We will use this value in the following task.

$$\begin{aligned} p(\mathcal{A}) &= \frac{p(\mathcal{A}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{p(\boldsymbol{\pi}|\mathcal{A})} \\ &= \frac{\text{B}(\boldsymbol{k} + \mathbf{1}a)}{\text{B}(\mathbf{1}a)} \end{aligned} \tag{6}$$

## 3 Task C

We wish to compute the probability of a test document, $\mathcal{T}$. We use a categorical distribution over a multinomial as a document is not uniquely defined by its word counts, the word order matters too. "Bush did 9/11" is a very different statement from "9/11 did Bush".

$$\begin{aligned} p(\mathcal{T}|\mathcal{A}) &= \prod_{n=1}^{N_\mathcal{T}} p(w_n|\mathcal{A}) \\ &= \prod_{m=1}^{M} \hat{\pi}_{w_n}^{k_{w_n}} \\ \log(p(\mathcal{T}|\mathcal{A})) &= \boldsymbol{k} \cdot \log(\hat{\boldsymbol{\pi}}) \end{aligned} \tag{7}$$

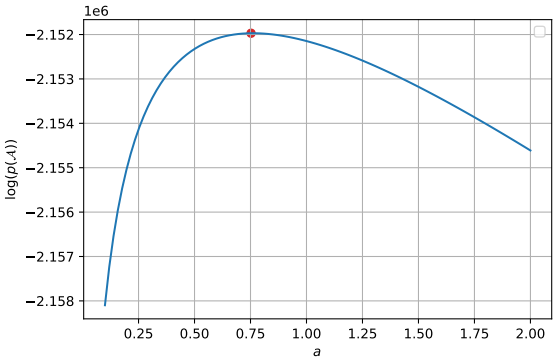



Figure 2: Optimisation of hyper-parameter $a$ using marginal likelihood.

| | Document 2001 - $\mathcal{B}_{2001}$ | Test Set - $\mathcal{B}$ | Uniform Multinomial |
|---|---|---|---|
| per-word perplexity | 4379.6 | 2686.1 | $6906 = M$ |

Table 1: Perplexities of documents.

For document 2001 in the test dataset, $\mathcal{B}_{2001}$, we calculate a log probability of $-3689.3$. We are also interested in the per word perplexity of a given document, Equation 8.

$$\text{Perplexity}(\mathcal{T}) = \exp(-\frac{\log(p(\mathcal{T}|\mathcal{A}))}{N_{\mathcal{T}}}) \tag{8}$$

The per word perplexity of $\mathcal{B}_{2001}$ and the entire test set, $\mathcal{B}$, are given in Table 1. We notice that the per word perplexity of $\mathcal{B}_{2001}$ is greater than that of the whole test set. This indicates that $\mathcal{B}_{2001}$ contains a higher frequency of words that are rare in the training set, $\mathcal{A}$, than the rest of the $\mathcal{B}$. The perplexity of a uniform multinomial is simply the size of the vocabulary $M$ and for long documents this is an upper bound.

# 4 Task D

This model still has a limitation in that it assumes all documents are drawn from the same distribution. However, in the real world documents can be about different topics and therefore have different word distributions. We can use a Mixture of Multinomials model where each document has a latent variable, $z_d$, that determines which of the $K$ topic distributions $\boldsymbol{\beta}_k$ it is drawn from. The latent topic variables are drawn from a categorical distribution with parameter $\boldsymbol{\theta}$. The prior for $\boldsymbol{\theta}$ and each $\boldsymbol{\beta}_k$ are symmetric Dirichlet with parameters $\alpha$ and $\gamma$ respectively. A graphical representation of the model is given in Figure 3.

We perform Gibbs sampling of the latent variable using the training documents, $\mathcal{A}$. We set our prior hyper-parameters to $\alpha = 1$ and $\gamma = 0.1$ and run the Gibbs sampler for 50 iterations. We compute topic posterior, $\hat{\boldsymbol{\theta}}^{(i)}$, at each Gibbs iteration $i$ using Equation 9, where $D_{\mathcal{A}}$ is the number of documents in $\mathcal{A}$. Similar to the previous multinomial posteriors, it is simply the frequency of each topic across the documents adjusted with a pseudo count of the prior, $\alpha$.

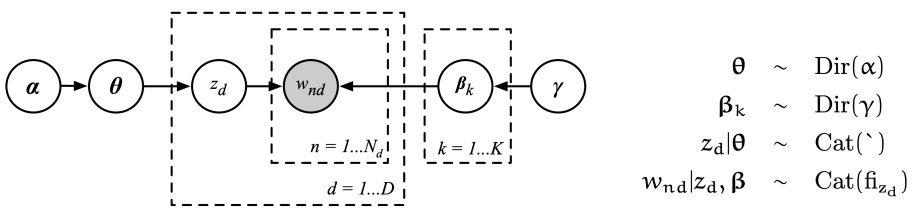
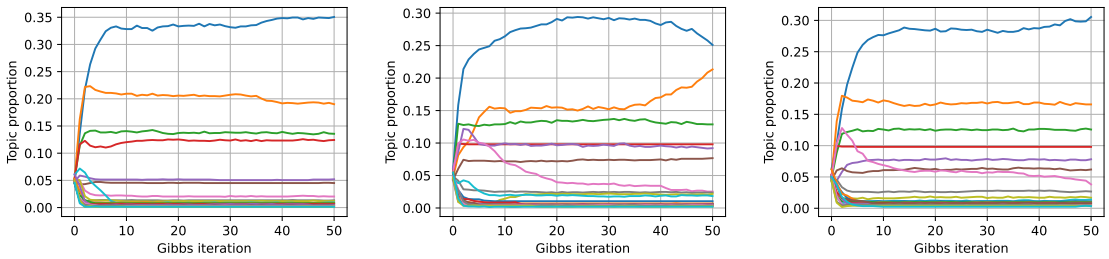



Figure 3: Mixture of Multinomials model.




(a) Seed = 0 (b) Seed = 10 (c) Seed = 100

Figure 4: Topic proportions at each Gibbs for different seeds, coloured by final proportion.

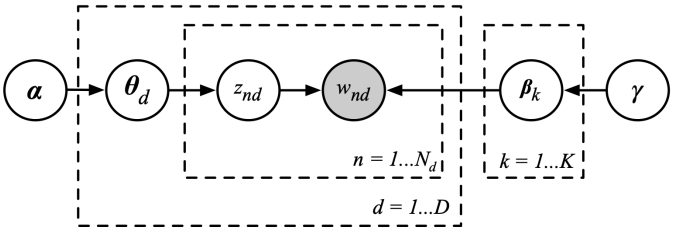



Figure 5: Latent Dirichlet Allocation Model.

$$(\hat{\boldsymbol{\theta}}^{(i)})_k = \frac{\alpha + \sum_{d=1}^{D_{\mathcal{A}}} \mathbb{1}(z_d = k)}{K\alpha + D_{\mathcal{A}}} \tag{9}$$

We plot the evolution of the topic posterior for 3 different seeds in Figure 4. We notice a burn in period of about 20 iterations before the sampler reaches the high probability region of our posterior. We expect the the same that the same "concepts" will not be assigned to the same topic variable across different seeds as the topic variables are arbitrary. However, we see that the final topic proportions are different across each seed. This indicates that that the topic posterior distribution is multimodal and the Gibbs sampler is not doing a good job of fully exploring the posterior distribution. In fact, in Figure 4b we see the Gibbs sampler switching to a different mode after iteration 40. The mode of the posterior distribution the we end up converging to is dependant on the seed. We could consider alternative MCMC methods to more fully explore the posterior distribution.

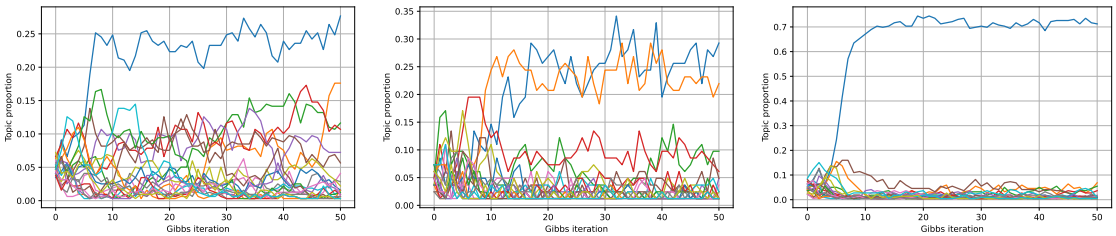
## 5 Task E

Sometimes a single document can talk about more than one topic, therefore, we make a final modification to our model. We allow the topic proportions, $\boldsymbol{\theta}_d$, to vary for each document. The $n$th word in document $d$ is then given a latent topic variable, $z_{nd}$, sampled from $\boldsymbol{\theta}_d$. We still use a symmetric Dirichlet prior with parameter $\alpha$ for each $\boldsymbol{\theta}_d$. This allows each document to be described by and arbitrary mixture of topics and is called Latent Dirichlet Allocation (LDA). The graphical model is given in Figure 5.

Again we perform Gibbs sampling, using the same hyper-parameter values as previously, to generate samples from our latent variable posteriors. We compute the topic posteriors from the Gibbs samples, $\hat{\boldsymbol{\theta}}_d$, as previously except now for each document individually. These are plotted for a selection of documents against Gibbs iteration in Figure 6.

$$(\hat{\boldsymbol{\theta}}_d^{(i)})_k = \frac{\alpha + \sum_{n=1}^{N_d} \mathbb{1}(z_{nd} = k)}{K\alpha + N_d} \tag{10}$$

For the documents in Figure 6 we see that the burn in period is about 20 iterations. We also notice that there is significant variation in how specific the topic distribution of each document is. Document 21 has a high proportion of a single topic whereas document 1 has a more uniform. The samples of some of the topic probabilities have significant variance




(a) Document 1 (b) Document 11 (c) Document 21

Figure 6: Posterior topic proportions at each Gibbs iteration for different documents.

| | Maximum Likelihood | Bayes Posterior | Mixture of Multinomials | LDA |
|---|---|---|---|---|
| Perplexity | $\infty$ | 2686.1 | 2112.6 | 1681.1 |

Table 2: Entropy of word distributions at each Gibbs iteration.

and therefore we should average over more iterations to obtain an accurate estimate of the topic proportions.

The perplexity of the test set $\mathcal{B}$ computed using the four models we have discussed is given in Table 2. As our models get more complex the perplexity decreases, this indicates that the test set has a higher likelihood under the more complex models. This is good evidence that the more complex models are both giving a better fit and not overfitting as the perplexity is being computed on the unseen test documents.

To assess convergence of the word posteriors for each topic we can compute the entropy for each topic with our Gibbs samples. We compute the word posterior, $\hat{\boldsymbol{\beta}}_k^{(i)}$, at each Gibbs iteration $i$ using Equation 11. $k_{mk}$ is the number of words $m$ assigned to topic $k$ in $\mathcal{A}$. Equation 12 gives the entropy of the word distribution for each topic. We are using natural logarithms and therefore the units of this entropy are nats.

$$(\hat{\boldsymbol{\beta}}_k^{(i)})_m = \frac{\alpha + k_{mk}}{M\gamma + \sum_{i=1}^{M} k_{ik}} \tag{11}$$

$$\begin{aligned} \mathrm{H}(W_k) &= -\sum_{m=1}^{M} p(w = m|\hat{\boldsymbol{\beta}}_k) \log(p(w = m|\hat{\boldsymbol{\beta}}_k)) : W_k \sim \mathrm{Cat}(\hat{\boldsymbol{\beta}}_k) \\ &= -\hat{\boldsymbol{\beta}}_k \cdot \log(\hat{\boldsymbol{\beta}}_k) \end{aligned} \tag{12}$$

We plot the word entropy against Gibbs iteration in Figure 7. A uniform distribution has the highest possible entropy and distributions with most probability mass concentrated in a smaller number of words have lower entropy. Therefore entropy of each topic generally decreases with iteration as the word distributions begin uniform and become increasingly specific as they converge to the posterior. The log perplexity = 7.43 of the test documents is higher than the entropy of any topic. This is expected as the test set is a blend of all topics and thus has higher entropy than any individual topic.

From Figures 6 and 7 we see that about 30 iterations are adequate to reach convergence for in the posteriors we have computed and therefore running 50 iterations is reasonable. In Figure 7 we see that there are only about 5 topics with lower entropy than the rest. This is evidence that our chosen number of topics, $K = 20$, is sufficient to capture the main distinct topics in the dataset.

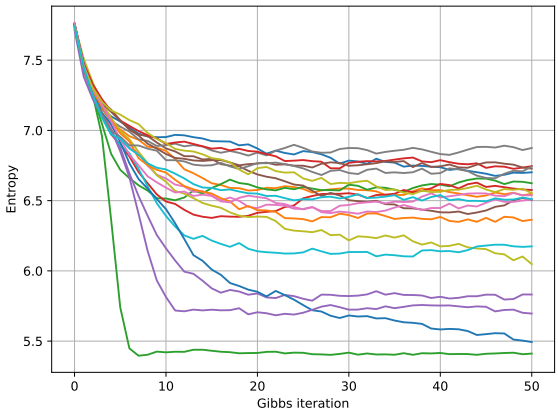



Figure 7: Entropy of word distributions at each Gibbs iteration.