4F13 Probabilistic Machine Learning
Coursework 3: Latent Dirichlet Allocation

Candidate Number: 5746G

December 11, 2020

Question a)

Listings 1: Maximum likelihood multinomial over words for the training data in `A`.

```
count_m = np.sum(A[np.where(A[:, 1] == w + 1), 2])
mlm[m] = count_m / total_count
```

Figure 1 shows a graphical representation of a simple generative document model in which documents are modelled by the global word frequency distribution. Observing only the counts of words, and not their sequence in each document, this is a *multinomial* distribution.

The parameters of the multinomial distribution over the vocabulary of M unique words $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^T$ can be fit by maximizing the likelihood of the training data $p(\boldsymbol{w} \mid \boldsymbol{\beta})$ subject to the constraint that $\boldsymbol{\beta}$ must be a valid probability mass function (non-negative, sum to one), yielding the *maximum likelihood estimation* (MLE) of parameters:

$$\beta_m^{\mathrm{ML}} = \frac{c_m}{\sum_{j=1}^{M} c_j} \tag{1}$$

where $c_m = \sum_{d=1}^{D} \sum_{n}^{N_d} \mathbb{I}(w_{nd} = m)$ is the total count of vocabulary word m. Figure 2(a) plots the 20 most probable words given by the MLE of parameters found by fitting to the training data in `A`. Under this model, the prediction for word w^* in the test set being the vocabulary word m uses the value of $\boldsymbol{\beta}^{\mathrm{ML}}$ that maximizes the likelihood of $\boldsymbol{\beta}$ given the training data:

$$p(w^* = m \mid \boldsymbol{\beta}^{\mathrm{ML}}) = \beta_m^{\mathrm{ML}}. \tag{2}$$

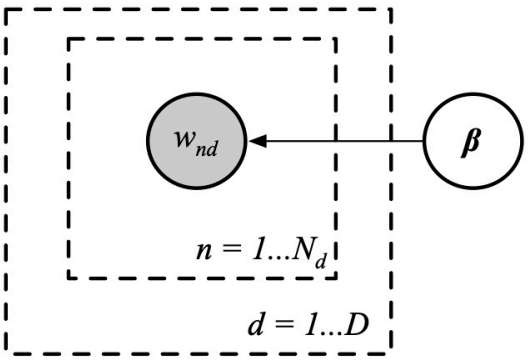
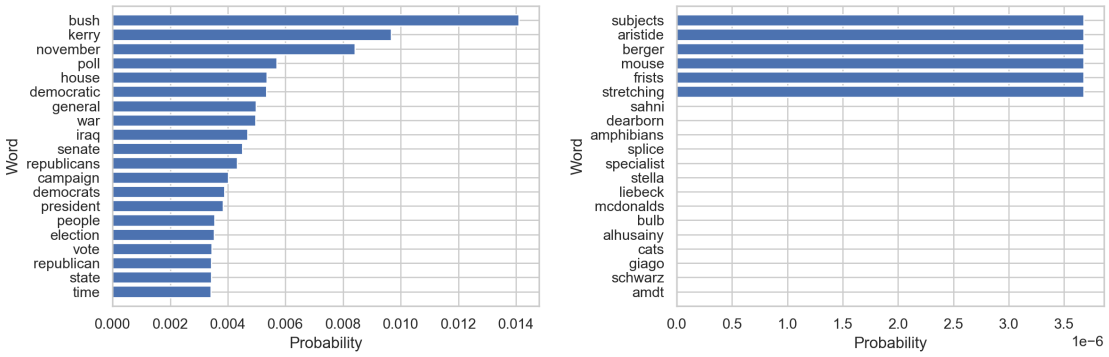


Figure 1: Multinomial document model in which the n-th word from the d-th document w_{nd} is drawn from a discrete multinomial distribution with parameters $\boldsymbol{\beta}$.



(a) **(b)**

Figure 2: Panels (a) and (b) plot the 20 most and least probable words from the MLE of parameters for the multinomial distribution trained on the word count data in `A`.

Therefore, the log probability of a test document with a specific sequence of independent words $\boldsymbol{w}^* = \{w_n^*\}_{n=1}^N$ is:

$$\log\left(p(\boldsymbol{w}^* \mid \boldsymbol{\beta}^{\mathrm{ML}})\right) = \log\left(\prod_{n=1}^{N} p(w_n^* \mid \boldsymbol{\beta}^{\mathrm{ML}})\right) \tag{3}$$

$$= \sum_{n=1}^{N} \log\left(p(w_n^* \mid \boldsymbol{\beta}^{\mathrm{ML}})\right) = \sum_{m=1}^{M} c_m^* \log(\beta_m^{\mathrm{ML}}) \tag{4}$$

where $c_m^* = \sum_{n=1}^{N} \mathbb{I}(w_n^* = m)$ is the count for word m in the test set. The test document which maximizes the log probability in eq. (4) contains N occurrences of the most probable word (`bush`):

$$\max_{\boldsymbol{w}^*} \log\left(p(\boldsymbol{w}^* \mid \boldsymbol{\beta}^{\mathrm{ML}})\right) = N \log(\beta_{\texttt{bush}}^{\mathrm{ML}}), \tag{5}$$

where:

$$\beta_{\texttt{bush}}^{\mathrm{ML}} = \max\{\beta_1^{\mathrm{ML}}, \ldots, \beta_M^{\mathrm{ML}}\}. \tag{6}$$

There are 14 words which have zero count for the training data in `A`, as shown in Figure 2(b), giving 14 components in the MLE of $\boldsymbol{\beta}$ with value zero. A test document containing any one of these words would have negative infinity ($-\infty$) log probability.

These two cases highlight the shortcomings of the ML multinomial model, which arise due to a lack of provision in preventing the ML model *overfitting* to the training data. A more robust model would assign the highest log probability to a test document with a distribution of word counts that most closely matches that of the training data in conjunction with any prior beliefs, and non-zero probability to words which do not appear in the training data. This can be achieved through a *Bayesian* approach, by placing priors on the parameters $\boldsymbol{\beta}$.

Question b)

Listings 2: Bayesian posterior predictive distribution using a symmetric Dirichlet prior with concentration parameter α.

```
count_m = np.sum(A[np.where(A[:, 1] == w + 1), 2])
bayesian_p[m] = (alpha + count_m) / (W * alpha + total_count)
```

A Bayesian model places priors $p(\boldsymbol{\beta})$ on the parameters of the multinomial distribution. Using a symmetric Dirichlet prior with a concentration parameter α on the word probabilities:

$$p(\boldsymbol{\beta}) = \mathrm{Dir}(\boldsymbol{\beta} \mid \alpha) \tag{7}$$

$$= \frac{1}{B(\alpha)} \prod_{m=1}^{M} \beta_m^{\alpha-1}, \tag{8}$$

the resulting posterior distribution is:

$$p(\boldsymbol{\beta} \mid \boldsymbol{w}_{\mathtt{A}}) = \frac{p(\boldsymbol{w}_{\mathtt{A}} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta})}{p(\boldsymbol{w}_{\mathtt{A}})} \tag{9}$$

$$\propto \prod_{m=1}^{M} \beta_m^{c_m} \prod_{m=1}^{M} \beta_m^{\alpha-1} \tag{10}$$

$$\propto \mathrm{Dir}\left(\boldsymbol{\beta} \mid c_1 + \alpha, \ldots, c_M + \alpha\right) \tag{11}$$

which is also a Dirichlet distribution; the Dirichlet distribution is a *conjugate prior* to the multinomial distribution.

Bayesian predictions are made by averaging over all possible parameter settings $\boldsymbol{\beta}$:

$$p\left(w^* = m \mid \alpha, \boldsymbol{w}_{\mathtt{A}}\right) = \int p(w^* = m \mid \boldsymbol{\beta}) p\left(\boldsymbol{\beta} \mid \boldsymbol{w}_{\mathtt{A}}\right) d\boldsymbol{\beta} \tag{12}$$

$$= \mathbb{E}_{p(\boldsymbol{\beta} \mid \boldsymbol{w}_{\mathtt{A}})}[\beta_m] = \mathbb{E}_{\mathrm{Dir}(\boldsymbol{\beta} \mid c_1+\alpha, \ldots, c_M+\alpha)}[\beta_m] \tag{13}$$

$$= \frac{\alpha + c_m}{\sum_{l=1}^{M} (\alpha + c_l)} = \frac{\alpha + c_m}{M\alpha + N}. \tag{14}$$

Comparing this expression to the MLE of $\boldsymbol{\beta}$ in eq. (1), introducing the prior distribution $p(\boldsymbol{\beta}) = \mathrm{Dir}(\boldsymbol{\beta} \mid \alpha)$ is equivalent to observing α pseudo-counts of each word. For $\alpha = 0$, eq. (14) reduces to $p\left(w^* = m \mid \alpha, \boldsymbol{w}_{\mathtt{A}}\right) = c_m / N$, and the MLE of parameters is recovered.

For $\alpha > 0$, words that are unobserved in the training data are associated a non-zero probability $\alpha/(M\alpha + N)$ in the posterior predictive distribution. As α is increased, the contribution of the prior relative to the likelihood is increased, and the probability of each word draws closer to $1/M$. This results in a decrease in the variance of the predictive distribution: the probability of words that are less frequent in the training data are increased, whilst the confidence of the model for more common words is reduced. The effect of a small and large α is observed in Figure 3, which plots the 20 most and least probable words for $\alpha = 0.01$ and $\alpha = 1000$.

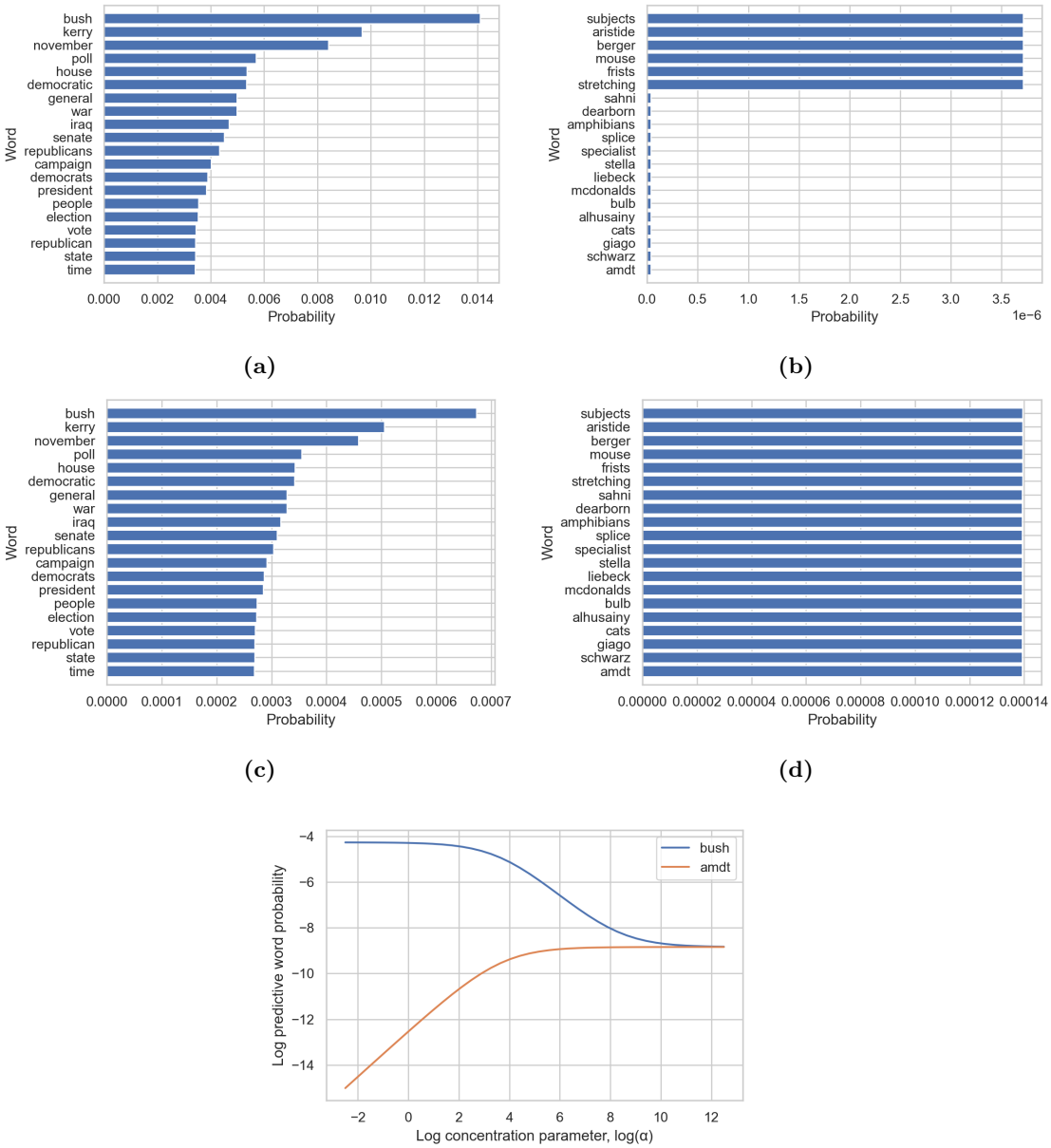


Figure 3: Panels (a) and (b) plot the 20 most and least probable words based on the posterior probability obtained through Bayesian inference using a symmetric Dirichlet prior with a small shape parameter $\alpha = 0.01$. Panels (c) and (d) plot the same rankings, but for a large shape parameter $\alpha = 1000$. A small shape parameter yields non-zero posterior probabilities close to the maximum likelihood estimates, shown in Figure 2. Increasing the shape parameter inflates the posterior probability of words less frequent in the training data, and reduces the probability of words which are more frequent. Panel (e) plots the log posterior probabilities of the most and joint least frequent words from the training data in `A` (`bush` and `amdt`) as functions of $\log(\alpha)$. As $\alpha \to \infty$, the log probability for both words tend to $\log(1/M) = \log(1/6906) = -8.840$.

Question c)

Listings 3: Log probability and per word perplexity of the test document with ID 2001.

```
doc_2001 = B[np.where(B[:, 0] == 2001)]
lp = np.dot(np.log(bayesian_m[doc_2001[:, 1] - 1]), doc_2001[:, 2])  # log
↪  probability, doc 2001
nd = np.sum(doc_2001[:, 2])  # number of words, doc 2001
perplexity = np.exp(-lp / nd)  # perplexity
```

The combinatorial factor in the multinomial distribution represents the number of sequences that are consistent with the word counts. Since words in a test document are arranged in a specific order, this factor is not required when computing their log probability, and so the categorical distribution is used. In effect, this is a multinomial distribution with just one trial. For the Bayesian model with parameters $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^T$ learnt from the training data in `A` (according to eq. (14)), this log probability is calculated as:

$$\log\left(\boldsymbol{w}^* \mid \boldsymbol{w}_{\mathtt{A}}, \alpha\right)\right) = \log\left(\prod_{n=1}^{N} p(w_n^* \mid \boldsymbol{\beta})\right) \tag{15}$$

$$= \sum_{n=1}^{N} \log\left(p(w_n^* \mid \boldsymbol{\beta})\right) = \sum_{m=1}^{M} c_m^* \log(\beta_m) \tag{16}$$

where the $c_m^* = \sum_{n=1}^{N} \mathbb{I}\left(w_n^* = m\right)$ is the total count of word m in the test set. For $\alpha = 0.1$, the log probability for the test document with ID 2001 is -3691. Figure 4(a) plots the log probability for this document as a function of the logarithm of the shape parameter. The log probability remains fairly flat for $\log(\alpha) < 2$. A maximum is reached for $\log(\alpha) = 2.23$, beyond which point increasing α drastically decreases the log likelihood of the test document; the pseudo-counts inflate the variance of the predictive distribution, and the predictive probability for words more common in the training data is reduced.

The *per-word perplexity* is a measure of how well a probabilistic model predicts unseen test data. It is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the (geometric) mean per-word likelihood [1]:

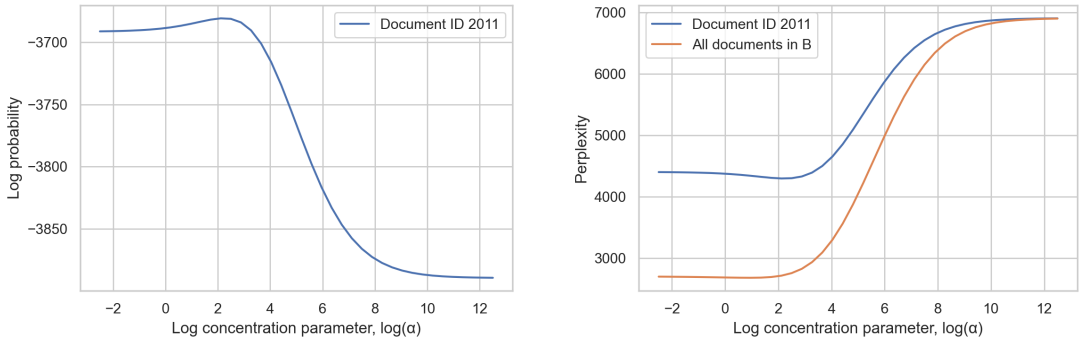
$$\text{perplexity}\left(\boldsymbol{w}^* \mid \boldsymbol{w}_{\mathtt{A}}, \alpha\right)\right) = \exp\left(\frac{\log\left(p(\boldsymbol{w}^* \mid \boldsymbol{w}_{\mathtt{A}}, \alpha)\right)}{\sum_{m=1}^{M} c_m^*}\right) \tag{17}$$

$$= \sqrt[N]{\frac{1}{\prod_{m=1}^{M} \beta_m^{c_m^*}}}. \tag{18}$$

Thus, a lower perplexity score indicates better generalization performance. The per-word perplexities for both the test document with ID 2001 and over all documents in `B` are quoted for a fixed $\alpha = 0.1$ in Table 1. Figure 4(b) plots perplexity as a function of $\log(\alpha)$. With a minimum perplexity for the document with ID 2001 at $\log(\alpha) = 2.23$, and showing an increase in perplexity for α beyond this, the plot reinforces the conclusions drawn from Figure 4(a).

Documents from the test data in `B` are of different length and have different distributions over frequency of words. Longer documents containing a larger number of more probable words have a higher likelihood, and so lower perplexity. On the other hand, shorter documents consisting of less probable words have a lower likelihood, and hence greater perplexity. The test document with ID 2001 has a higher perplexity than the that over all documents in `B`, and so contains a greater proportion of less probable words than the documents in `B` as a whole.

A uniform multinomial model is obtained in the limit $\alpha \to \infty$. It assigns equal probability to



(a) **(b)**

Figure 4: Panel (a) plots the log probability of test document with ID 2001 with varying α. Panel (b) plots the per word perplexity for the same test document and over all documents in `B`.

Table 1: Per word perplexity for the symmetric Bayesian model with concentration parameter $\alpha = 0.1$.

	Document ID 2001	All documents in `B`
Perplexity	4398	2697

every word in the vocabulary, and so is the most uncertain distribution over the vocabulary:

$$\boldsymbol{\beta} = \left[\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right]. \tag{19}$$

The perplexity of such model is $\exp(\log(M)) = M = 6906$.

Question d)

Listings 4: Mixing proportions $\boldsymbol{\theta}$ for the i-th iteration of the collapsed Gibbs sampler for the Bayesian mixture model.

```
theta[:, iter] = (sk_docs[:, 0] + alpha) / np.sum(sk_docs[:, 0] + alpha)
```

Figure 5 shows a graphical representation of a Bayesian mixture model. Allowing for a mixture of K categoricals $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_K$, in which each of the categorical distributions corresponds to a document category, enables different documents to be modelled by different topics.

Topic assignments for each document are randomly initialised, and the collapsed Gibbs sampler used to sample the latent assignments by marginalising out the mixing proportions $\boldsymbol{\theta}$:

$$p\left(z_d = k \mid \mathbf{w}_d, z_{-d}, \boldsymbol{\beta}, \alpha\right) \propto p\left(\mathbf{w}_d \mid \beta_k\right) \frac{\alpha + c_{-d,k}}{\sum_{j=1}^{K}(\alpha + c_{-d,j})} \tag{20}$$

where $c_k = \sum_{d=1}^{D} \mathbb{I}\left(z_d = k\right)$ are the counts for mixture k, and the index $-d$ means *all except* d. The posterior probability of each of the mixture components is then:

$$\theta_k = \frac{c_k + \alpha}{\sum_{j=1}^{K}\left(c_j + \alpha\right)}. \tag{21}$$

Figures 6(a) and 6(b) plot the evolution of the mixing proportions for 20 document categories ($K = 20$) from two different initialisations (seeds) as a function of the number of Gibbs sweeps,

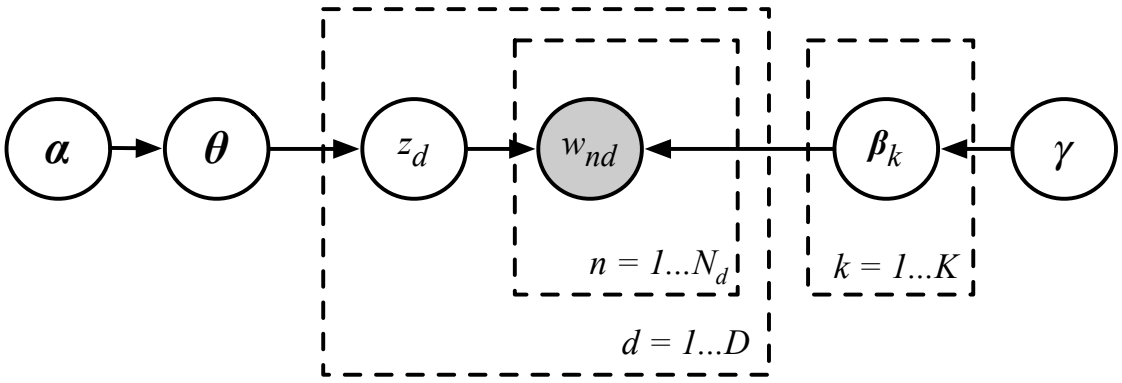
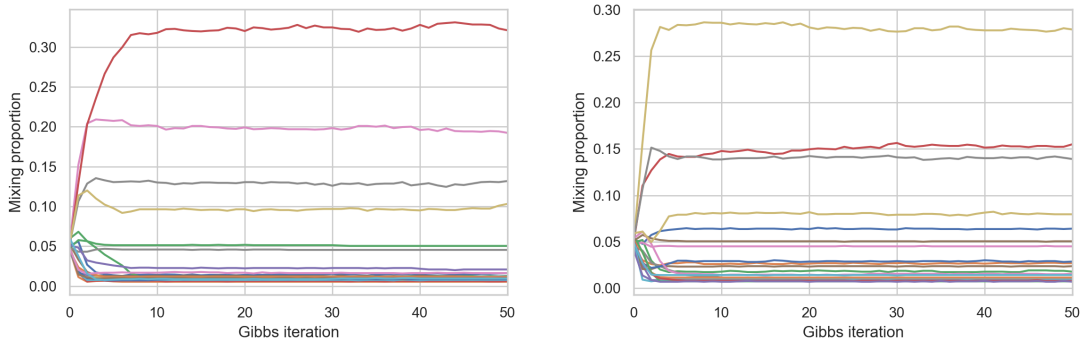


Figure 5: Bayesian mixture document model in which the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are inferred starting from symmetric Dirichlet priors with concentration parameters α (over categories) and γ (over words). $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$ is the parameter of a categorical distribution over K categories, from which the latent variable z_d is drawn. The latent variable $z_d \in \{1, \ldots K\}$ assigns document d to one of the K categories, such that w_{nd} is drawn from a categorical parameterised by $\boldsymbol{\beta}_{z_d}$.



(a) Seed 0 **(b)** Seed 1

Figure 6: Panels (a) and (b) plot the evolution of mixing proportions for the Bayesian mixture model as functions of the number of Gibbs sweeps for two different initialisations.

using symmetric Dirichlet priors with concentration parameters $\alpha = 10$ and $\gamma = 0.1$ respectively. In both cases, the early samples vary significantly, before settling on the *stationary distribution* of the transition kernel which defines the Markov Chain. Although the samples continue to fluctuate, suggesting the sampler is moving inside the posterior, the differently initialised samplers settle on different local areas of the stationary distribution (the posterior). Since neither sampler fully explores the invariant distribution, the Gibbs sampler has not converged after 50 iterations.

The posterior distribution is dominated by a small number of topics, with many topics assigned near zero probabilities. This behaviour arises as a result of the *rich get richer* property of the collapsed Gibbs sampler in eq. (20).

Question e)

Listings 5: Topic posteriors and word entropies $H(\boldsymbol{\beta}_k)$ for the i-th iteration of the collapsed Gibbs sampler for the LDA model.

```
theta_d = (skd + alpha) / np.sum(skd + alpha, axis=0)  # posterior
↪  probabilities over topics for document d
theta[:, iter] = np.mean(theta_d, axis=1)  # average posterior probabilities
↪  over topics across documents
beta = (swk + gamma) / np.sum(swk + gamma, axis=0)
word_entropy[:, iter] = -np.sum(np.multiply(beta, np.log(beta)), axis=0)
```

The Bayesian mixture model draws every word in each document from one specific topic distribution. The latent Dirichlet allocation (LDA) model, shown graphically in Figure 7, facilitates for documents which span more than one topic; every word in each document is drawn from a different topic, and every document has its own distribution over topics. For this model, topic posteriors are defined as the posterior probabilities over topics, averaged across documents.

The update step in the collapsed Gibbs sampler consists of two components, one from the topic distribution and one from the word distribution:

$$p\left(z_{nd}=k \mid \{z_{-nd}\}, \{w\}, \gamma, \alpha\right) \propto \underbrace{\frac{\alpha + c^k_{-nd}}{\sum_{j=1}^{k}\left(\alpha + c^j_{-nd}\right)}}_{\text{topic distribution}} \underbrace{\frac{\gamma + \tilde{c}^k_{-w_{nd}}}{\sum_{m=1}^{M}\left(\gamma + \tilde{c}^k_{-m}\right)}}_{\text{word distribution}} \tag{22}$$

where c^k_{-nd} is the count of words from document d, excluding n, assigned to topic k, and $\tilde{c}^k_{-m}$ is the number of times word m was generated from topic k (again, excluding the observation nd).

Figures 8(a) and 8(b) plot topic posteriors for 20 document categories ($K = 20$) from two different intialisations as a function of Gibbs sweeps, using symmetric Dirichlet priors with concentration parameters $\alpha = 0.1$ and $\gamma = 0.1$. As with the Bayesian mixture model, following an initial period of substantial variation the posteriors settle on fairly constant values. Once again, the samplers remain in local regions of the posterior dependent on their initialisations, indicating the stationary distribution is not fully explored, and that the samples have not converged.

The per word perplexity over all documents in test set `B` for the state after 50 Gibbs sweeps is 1648. This is 21.9% and 38.8% lower than the Bayesian and Bayesian mixture models, the perplexities for which are given in Table 2, indicating that the LDA model best predicts the unseen test data. Of the four models investigated, the more complex a document model, the better it generalises, and so the likelihood of the test data increases.

Entropy is a measure of the *uncertainty* associated with an observation of random variables [2]. Defined in terms of the natural logarithm, the *word entropy* of the random variable w drawn from the posterior categorical distribution for k-th document category $\text{Cat}(\boldsymbol{\beta}_k)$ is given by:

$$H(w \mid z=k) = -\mathbb{E}\left[\log(\boldsymbol{\beta}_k)\right] = -\sum_{m=1}^{M} \boldsymbol{\beta}_{mk} \log(\boldsymbol{\beta}_{mk}), \tag{23}$$

and has units 'nats'. Use of the natural logarithm provides a convenient link with perplexity.

Table 2: Per word perplexity over all documents in `B`.

	Maximum Likelihood	Bayesian Model ($\alpha = 0.1$)	BMM ($\alpha = 10, \gamma = 0.1$)	LDA ($\alpha = 0.1, \gamma = 0.1$)
Perplexity	∞	2697	2111	1648

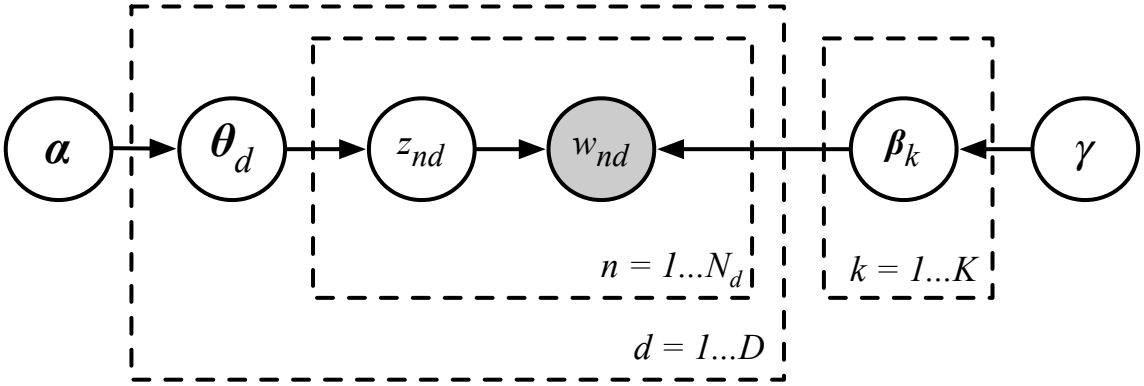
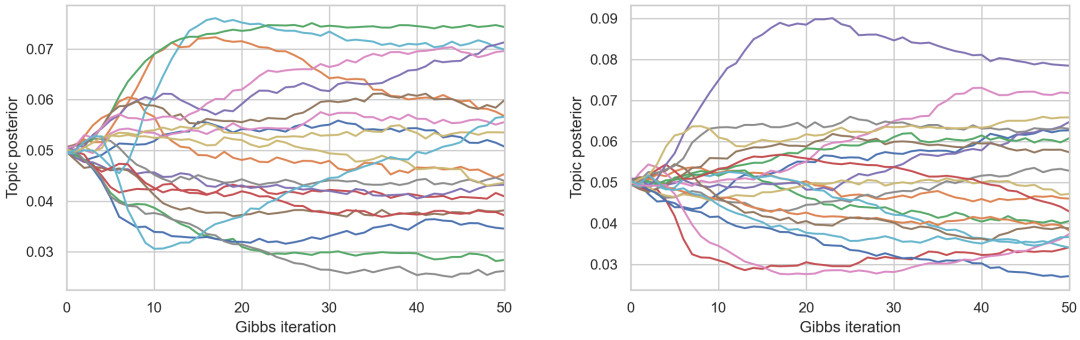
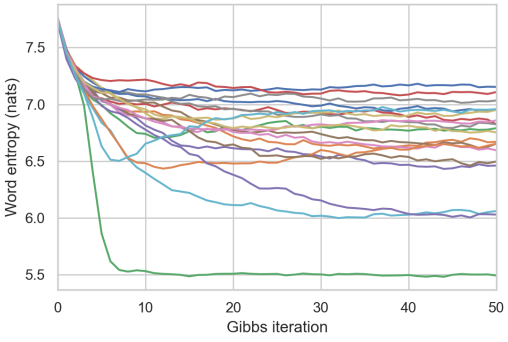


Figure 7: Latent Dirichlet allocation (LDA) document model in which the topic z_{nd} for the n-th word in document d is drawn from the document's categorical distribution over topics $\text{Cat}(\boldsymbol{\theta}_d)$. The latent variable $z_{nd} \in \{1, \ldots, K\}$ assigns the word w_{nd} to one of K categories, such that w_{nd} is drawn from the categorical distribution $\text{Cat}(\boldsymbol{\beta}_{z_{nd}})$.



(a) Seed 0 **(b)** Seed 1



(c)

Figure 8: Panels (a) and (b) plot the evolution of mixing proportions for the LDA model as functions of the number of Gibbs sweeps for two different initialisations. Panel (c) plots the word entropy for each topic's posterior categorical distribution with Gibbs iterations.

Figure 8(c) plots the evolution of the LDA word entropies as a function of Gibbs sweeps. Initialised by assigning each occurrence of word w to a document at random, entropies all begin very high. As iterations are run, and samples approach the stationary distribution, the categorical distribution for each topic $\text{Cat}(\boldsymbol{\beta_k})$ concentrates over a subset of words. Becoming more specialised, the uncertainty in the categorical distribution decreases, resulting in a decrease in word entropy. The word entropies largely stabilise after 30 iterations, indicating little further change to the categorical distributions beyond this; although the Gibbs sampler has not fully converged, 50 iterations is sufficient to make predictions based on the posterior distributions obtained.

The log perplexity of the test data in `B` is $\log(1648) = 7.41$, higher than the word entropy for any individual topic. The documents in `B` are modelled by a combination of topics, and so have higher uncertainty than any single topic.

Word Count (excluding Listings): 999

References

[1] David M. Blei, Andrew Y. Ng, and Michael T. Jordan. Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, 3:993–1022, 2002. ISSN 10495258.

[2] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 (4):623–656, 1948. ISSN 15387305. doi: 10.1002/j.1538-7305.1948.tb00917.x.