CS 7140: Advanced Machine Learning

Lecture 11: Latent Dirichlet Allocation

Instructor

Jan-Willem van de Meent (j.vandemeent@northeastern.edu)

Scribes

Andrea Baisero (baisero.a@husky.neu.edu)
Jordan Helderman (helderman.j@husky.neu.edu)
Yuan Guo (guo.yu@husky.neu.edu)

1 LDA Model Introduction

Latent Dirichlet Allocation (LDA) is a probabilistic model for characterizing discrete data. It was initially developed in a paper by David Blei, et. al. (2003) [1] as a method of performing latent topic analysis on text corpora. The goal of latent topic analysis is to discover groups of documents that exist in a collection of discrete data that share certain latent characteristics. In the context of text characterization, these groups represent textual topics, e.g. sports, statistics, politics, etc. There are several variants of principal components analysis (PCA) that have been applied to the problem (categorical PCA, exponential PCA, multinomial PCA). Additionally, there are a number of methods for this analysis represented in the information retrieval literature, and one such method, called Latent Semantic Indexing (LSI), and a probabilistic extension (probabilistic LSI) are the logical predecessors of LDA [2] [3]. The key difference between LDA and its predecessor, pLSI, is that LDA adds Dirichlet priors to the model. In the following, we will describe this model in detail and derive methods for inferring the model parameters.

1.1 LDA Model Description

LDA models documents by assigning a topic mixtures to each word in the document and computing aggregate statistics on these topic mixtures over the whole document. More formally, the words, y_{dn}, in a document, d, and at a position, n, as being drawn from a discrete distribution over all possible words. This discrete distribution is dependent the topic, z_{dn}, which is also modeled as a discrete distribution and a probability vector, β_k which represents the probability of each word in the vocabulary given the topic of the n-th word and is modeled as a Dirichlet distributed random vector parameterized by ω. Similarly, the topic, z_{dn} is parameterized by a probability vector θ_d which represents the probability of each topic conditioned on a particular document and is parameterized by α. Putting all this together, we can write the graphical model for LDA like the one in figure 1, and we can write the distribution of the variables in the model as follows.

$$\begin{aligned} y_{d,n} \mid z_{d,n} = k &\sim \text{Discrete}(\beta_k) \\ \beta_k &\sim \text{Dirichlet}(\omega) \\ z_{d,n} &\sim \text{Discrete}(\theta_d) \\ \theta_d &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

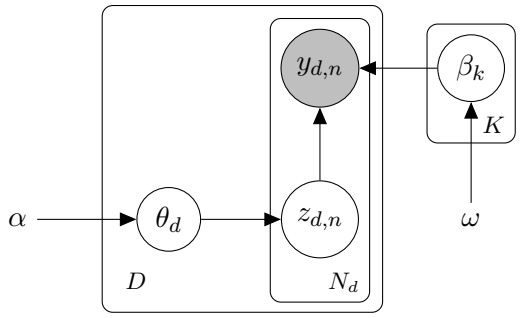

Figure 1: The Latent Dirichlet Allocation Model

Using this graphical model, we can then write the full joint distribution as

$$p(y, z, \beta, \theta) = \prod_k p(\beta_k) \prod_d p(\theta_d) \prod_n p(z_{d,n} \mid \theta_d) p(y_{d,n} \mid z_{d,n}, \beta) \tag{1}$$

Beyond the model definition, we make the following observations about the model. First, the topics are differentiated from one another by their word probability density β_k, and documents are characterized by their topic distribution vector θ_d. LDA differs from many topic analysis methods in that it assigns a topic to each word and models the document as a mixture of topics. Finally, using the concept of D-separation, it is relatively easy to show the following conditional independences are true.

$$(y, z)_d \perp (y, z)_{d' \neq d} \mid \beta \tag{2}$$
$$\beta_k \perp \beta_{k' \neq k} \mid y, z \tag{3}$$
$$\theta_d \perp \theta_{d' \neq d} \mid y, \beta \tag{4}$$

Throughout the rest of this document, we will be referring to the following counts

$$N_k = \sum_{d,n} \mathbb{I}[z_{d,n} = k]\,, \qquad N_k^{-d,n} = N_k - \mathbb{I}[z_{d,n} = k]\,, \tag{5}$$

$$N_{d,k} = \sum_{n} \mathbb{I}[z_{d,n} = k]\,, \qquad N_{d,k}^{-d,n} = N_{d,k} - \mathbb{I}[z_{d,n} = k]\,, \tag{6}$$

$$N_{k,v} = \sum_{d,n} \mathbb{I}[y_{d,n} = v, z_{d,n} = k]\,, \qquad N_{k,v}^{-d,n} = N_{k,v} - \mathbb{I}[y_{d,n} = v, z_{d,n} = k]\,, \tag{7}$$

representing the following quantities:

- N_k is the total number of words, across all documents, associated with topic k, while $N_k^{-d,n}$ is the same count albeit excluding word n in document d;
- $N_{d,k}$ is the total number of words in document d associated with topic k, while $N_{d,k}^{-d,n}$ is the same count albeit excluding word n in document d; and
- $N_{k,v}$ is the total number of times, across all documents, that word v is associated with topic k, while $N_{k,v}^{-d,n}$ is the same count albeit excluding word n in document d.

1.2 Applications and Extensions of LDA

Other than text characterization, these methods have been applied to genetics, health science, and social network analysis, and a number of extensions of LDA have been proposed to adapt the technique for different applications and handle issues that often arise in using it. The following is a summary of the extensions discussed in [4]. First, LDA has been known to be able to do unsupervised discovery of topics, but, as the algorithm does not include any information about the topics present in the text, these automatically discovered topics are sometimes difficult to interpret. To address this issue, there have been a few supervised extensions of LDA that have been proposed in the literature which constrains the topics to correspond to a known set of topics. Additionally, a common problem in machine learning is the problem of model selection, and the same problem can be found in LDA in the selection of the number of topics. Beyond the typical solutions to this problem (like cross validation), a nonparametric extension of LDA has been developed through the application of the Hierarchical Dirichlet Process (Teh, et. al. 2006) [5].

1.3 The Dirichlet Distribution

At a high level, the Dirichlet distribution is a distribution on discrete probability distributions of a fixed support set. Draws from the Dirichlet distribution are vectors of a fixed length, which is the size of the support set, and the values of the vectors are numbers between 0 and 1 with the sum of all the vector elements being equal to 1. The joint probability density function for this distribution can be seen in the following equation.

$$p(x_1, x_2, ..., x_K; \alpha) = \frac{1}{B(\alpha_1, \alpha_2, ..., \alpha_K)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

$$B(\alpha_1, \alpha_2, ..., \alpha_K) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)},$$

where $x_i \in [0, 1]$, $\sum_i x_i = 1$, and $\alpha_i > 0$.

The first two moments of the Dirichlet distribution are given by the following:

$$\mathbb{E}[X_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j},$$

$$\mathrm{Var}[X_i] = \frac{\alpha_i \left(\left(\sum_{j=1}^{K} \alpha_j\right) - \alpha_i\right)}{\left(\sum_{j=1}^{K} \alpha_j\right)^2 \left(\sum_{j=1}^{K} \alpha_j + 1\right)}.$$

Intuitively, if we look at the form of the mean, we can see that the individual alpha parameters represent a concentration (normalized by the sum of all the alpha parameters) of the corresponding component of the probability vector. Additionally, if we plot the distribution for the three dimensional Dirichlet distribution over the simplex (figure 2), we can see that the magnitude (more precisely, the L_1 norm) of these alpha parameters controls the variance of the draws from the distribution. In this figure, blue represents areas of lower relative concentration and red represents areas of higher relative concentration.

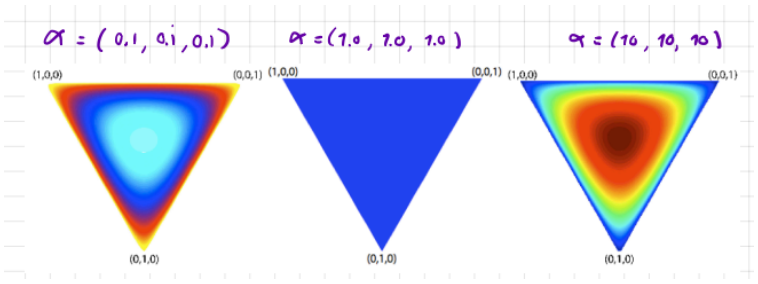

Figure 2: Dirichlet Distribution for Varying Alpha

1.4 Joint Distribution over Discrete Variables

In order to simplify notation, we will represent the joint distribution over N discrete variables in terms of its sufficient statistics:

$$p(z_{d,1}, \ldots, z_{d,N}; \theta) = \prod_{n=1}^{N} p(z_{d,n}; \theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_k^{I[z_{d,n}=k]} = \prod_{k=1}^{K} \theta_k^{N_{d,k}}.$$

where θ_k is the probability of drawing outcome k on any given draw.

2 Exponential Family

Before diving into Gibbs sampling, we freshen up our background knowledge on the exponential family. In this section, we briefly forget the previous notation and will be reusing variable names such as z and θ and α. Although there is sometimes a strong connection between their usage in this section and their meaning in the context of the LDA model, be wary about conflating their meaning and confusing the one for the other.

A parameterized distribution is in the exponential family if it can be expressed in the form

$$p(x \mid \eta) = h(x) \exp\left[\eta^\top t(x) - a(\eta)\right], \text{where} \tag{8}$$

$$a(\eta) = \log \int h(x) \exp\left[\eta^\top t(x)\right] dx \qquad \text{(by normalization)} \tag{9}$$

Joint over Discrete Variables in Exponential Form Let us consider a set of variable $z_1, \ldots, z_D$, and derive its natural parameterization according to the exponential family form:

$$p(z; \theta) = \exp\left[\log\left(\prod_{d,k} \theta_{d,k}^{N_{d,k}}\right)\right] = \exp\left[\sum_{d,k} N_{d,k} \log \theta_{d,k}\right] \tag{10}$$

Equation (10) matches the template of an exponential family distribution as follows:

$$p(z; \eta) = \exp\left[\eta^\top t(z)\right], \text{where} \tag{11}$$

$$h(z) = 1 \tag{12}$$

$$\eta_{d,k} = \log \theta_{d,k} \tag{13}$$

$$t_{d,k}(z) = N_{d,k} \tag{14}$$

$$a(\eta) = 0 \tag{15}$$

Dirichlet in Exponential Form As expert statisticians, we recognize instantly that the conjugate prior of the multinomial distribution is the dirichlet distribution, which is the not-so-serendipitous prior distribution on the LDA model parameters β and θ. Let us derive the natural parameterization of a Dirichlet distribution:

$$p(\theta;\alpha) = \frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1} \tag{16}$$

$$= \exp\left[\log\left(\frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1}\right)\right] \tag{17}$$

$$= \exp\left[\sum_k (\alpha_k - 1)\log\theta_k - \log B(\alpha)\right] \tag{18}$$

Once again, we notice that eq. (18) adheres to the exponential form:

$$p(\eta;\lambda) = \exp\left[\lambda^\top t(\eta) - a(\lambda)\right], \text{where} \tag{19}$$

$$h(\eta) = 1 \tag{20}$$

$$\lambda_k = \alpha_k - 1 \tag{21}$$

$$t_k(\eta) = \log\theta_k \tag{22}$$

$$a(\lambda) = \log B(\alpha) \tag{23}$$

Dirichlet Posterior and Multinomial Marginal Finally, the power (and generality) of conjugacy reveals itself in two results which are directly relevant w.r.t. the LDA model. Foremostly, conjugacy implies a succinct expression for the posterior Dirichlet parameter distribution:

$$p(\eta \mid z) = p(\eta;\tilde{\lambda}), \text{where} \tag{24}$$

$$\tilde{\lambda}_k = \lambda_k + t_k(z) \tag{25}$$

which, in canonical Dirichlet-multinomial form, corresponds to the update

$$\tilde{\alpha}_k = \alpha_k + N_k \tag{26}$$

Further, we find that the *marginal likelihood* $p(z)$ for the Dirichlet-multinomial pair can be expressed exclusively as a direct function of the prior and posterior natural parameters:

$$p(z) = \int p(z \mid \eta) p(\eta)\, d\eta \tag{27}$$

$$= \int \exp\left[\eta^\top t(z) + \lambda^\top \eta - a(\lambda)\right] d\eta \qquad \text{(by eqs. (11) and (19))} \tag{28}$$

$$= \left[\int \exp\left[\eta^\top (t(z) + \lambda)\right] d\eta\right] \exp\left[-a(\lambda)\right] \tag{29}$$

$$= \left[\int \exp\left[\eta^\top \tilde{\lambda}\right] d\eta\right] \exp\left[-a(\lambda)\right] \qquad \text{(by eq. (25))} \tag{30}$$

$$= \exp\left[a(\tilde{\lambda})\right] \exp\left[-a(\lambda)\right] \qquad \text{(by eq. (9))} \tag{31}$$

$$= \frac{B(\tilde{\alpha})}{B(\alpha)} \qquad \text{(by eq. (23))} \tag{32}$$

3 Gibbs Sampling for LDA

We will discuss two versions of Gibbs sampling for the LDA model. The first method corresponds to the standard Gibbs setup, where we alternate sampling topic variables $(z \mid y, \beta, \theta)$ and model parameters $(\beta, \theta \mid y, z)$ iteratively. The second method, called *Collapsed* Gibbs sampling, is a more efficient alternative characterized by the fact that we avoid (via marginalization) having to deal with the model parameters altogether, only sampling individual topic assignments while conditioning on all the others $(z_{d,n} \mid y, z_{-d,n})$, where $z_{-d,n} \doteq z \setminus \{z_{d,n}\}$.

3.1 (Non-collapsed) Gibbs Sampling

To run standard Gibbs sampling, we need to find closed form expressions for the conditional topic assignments $p(z \mid y, \beta, \theta)$ and the conditional model parameters $p(\beta, \theta \mid y, z)$. The model's independence assumptions will play a key role in our ability to compute this quantities efficiently (if at all!).

Conditional Topic Assignments Equation (2) allows us to sample each topic assignment individually, according to

$$p(z \mid y, \beta, \theta) = \prod_{d,n} p(z_{d,n} \mid y, \beta, \theta) \tag{33}$$

$$p(z_{d,n} = k \mid y_{d,n} = v, y_{-d,n}, \beta, \theta) \propto p(y, z, \beta, \theta) \tag{34}$$

$$\propto p(y_{d,n} = v \mid z_{d,n} = k, \beta) p(z_{d,n} = k \mid \theta_d) \tag{35}$$

$$= \beta_{k,v} \theta_{d,k} \tag{36}$$

Conditional Model Parameters We use the product rule and the model's independence assumptions eqs. (3) and (4) to factorize the conditional distribution of model parameters as

$$p(\beta, \theta \mid y, z) = p(\beta \mid y, z) p(\theta \mid y, z, \beta) \tag{37}$$

$$= \prod_k p(\beta_k \mid y, z) \prod_d p(\theta_d \mid y, z, \beta)\,, \tag{38}$$

which can be solved individually:

$$p(\beta_k \mid y, z) \propto p(\beta_k) \prod_{d,n} p(y_{d,n} \mid z_{d,n} = k, \beta)^{\mathbb{I}[z_{d,n}=k]} \tag{39}$$

$$\propto \left[\prod_v \beta_{k,v}^{\omega_v - 1}\right] \left[\prod_{d,n,v} \underbrace{p(y_{d,n} = v \mid z_{d,n} = k, \beta)}_{\beta_{k,v}}{}^{\mathbb{I}[y_{d,n}=v, z_{d,n}=k]}\right] \tag{40}$$

$$= \left[\prod_v \beta_{k,v}^{\omega_v - 1}\right] \left[\prod_v \beta_{k,v}^{\sum_{d,n} \mathbb{I}[y_{d,n}=v, z_{d,n}=k]}\right] \tag{41}$$

$$= \prod_v \beta_{k,v}^{\omega_v - 1 + N_{k,v}} \tag{42}$$

$$\propto \text{Dirichlet}(\beta_k; \{\omega_v + N_{k,v}\}_v)\,, \tag{43}$$

and

$$p(\theta_d \mid y, z, \beta) \propto p(\theta_d) \prod_n p(z_{d,n} \mid \theta_d) \tag{44}$$

$$\propto \left[\prod_k \theta_{d,k}^{\alpha_k - 1}\right] \left[\prod_{n,k} \underbrace{p(z_{d,n} = k \mid \theta_d)}_{\theta_{d,k}}^{\mathbb{I}[z_{d,n}=k]}\right] \tag{45}$$

$$= \left[\prod_k \theta_{d,k}^{\alpha_k - 1}\right] \left[\prod_k \theta_{d,k}^{\sum_n \mathbb{I}[z_{d,n}=k]}\right] \tag{46}$$

$$= \prod_k \theta_{d,k}^{\alpha_k - 1 + N_{d,k}} \tag{47}$$

$$\propto \text{Dirichlet}(\theta_d; \{\alpha_k + N_{d,k}\}_k)\,. \tag{48}$$

To summarize the standard Gibbs sampling approach, eqs. (36), (43) and (48) give closed form solutions for sampling the topic assignments and model parameters iteratively.

3.2 Collapsed Gibbs Sampling

In Collapsed Gibbs sampling, the model parameters are taken out of the sampling process via marginalization, and the individual topic assignments $z_{d,n}$ are iteratively sampled directly by conditioning on the other assignments $z_{-d,n}$, i.e. we aim at sampling directly from $p(z_{d,n} \mid y, y_{-d,n})$.

$$p(z_{d,n} = k \mid y_{d,n} = v, y_{-d,n}, z_{-d,n}) \tag{49}$$

$$= \frac{p(y_{-d,n} \mid y_{d,n} = v, z_{d,n} = k, z_{-d,n})}{p(y_{-d,n} \mid y_{d,n} = v, z_{-d,n})} p(z_{d,n} = k \mid y_{d,n} = v, z_{-d,n}) \tag{50}$$

$$= \frac{p(y_{-d,n} \mid y_{d,n} = v, z_{d,n} = k, z_{-d,n})}{p(y_{-d,n} \mid \cancel{y_{d,n} = v}, z_{-d,n})} \frac{p(z_{-d,n} \mid \cancel{y_{d,n} = v}, z_{d,n} = k)}{p(z_{-d,n} \mid \cancel{y_{d,n} = v})} p(z_{d,n} = k \mid y_{d,n} = v) \tag{51}$$

$$\propto \underbrace{\frac{p(y_{-d,n} \mid y_{d,n} = v, z_{d,n} = k, z_{-d,n})}{p(y_{-d,n} \mid z_{-d,n})}}_{A} \underbrace{\frac{p(z_{-d,n} \mid z_{d,n} = k)}{p(z_{-d,n})}}_{B} \tag{52}$$

Notice that, $z_{-d,n}$ and $(y_{-d,n} \mid z_{-d,n})$ are both multinomial-distributed with Dirichlet prior, so can exploit eq. (32) to compute

$$p(y_{-d.n} \mid z_{-d,n}) = \frac{B(\{\tilde{\omega}_{k,v}^{-d,n}\}_{k,v})}{B(\{\omega_v^{-d,n}\}_{k,v})}\,, \tag{53}$$

$$p(z_{-d,n}) = \frac{B(\{\tilde{\alpha}_{d,k}^{-d,n}\}_{d,k})}{B(\{\alpha_k^{-d,n}\}_{d,k})}\,, \tag{54}$$

and, compared to the above two, only an additional count needs to be added for

$$p(y_{-d.n} \mid y_{d,n} = v, z_{d,n} = k, z_{-d,n}) = \frac{B(\{\tilde{\omega}_{k',v'}^{-d,n} + \mathbb{I}[k' = k, v' = v]\}_{k',v'})}{B(\{\omega_v'^{-d,n}\}_{k',v'})}\,, \tag{55}$$

$$p(z_{-d,n} \mid z_{d,n} = k) = \frac{B(\{\tilde{\alpha}_{d',k'}^{-d,n} + \mathbb{I}[d' = d, k' = k]\}_{d',k'})}{B(\{\alpha_{k'}^{-d,n}\}_{d',k'})}\,. \tag{56}$$

Equations (53) to (56) allow us to simplify (A) and (B) into

$$\frac{p(y_{-d,n} \mid y_{d,n}=v, z_{d,n}=k, z_{-d,n})}{p(y_{-d,n} \mid z_{-d,n})} = \frac{B(\{\tilde{\omega}^{-d,n}_{k',v'} + \mathbb{I}[k'=k, v'=v]\})}{\cancel{B(\{\omega^{-d,n}_{k',v'}\})}} \frac{\cancel{B(\{\omega^{-d,n}_{k',v'}\})}}{B(\{\tilde{\omega}^{-d,n}_{k',v'}\})} \tag{57}$$

$$= \frac{\prod_{k',v'} \Gamma\left(\tilde{\omega}^{-d,n}_{k',v'} + \mathbb{I}[k'=k, v'=v]\right)}{\Gamma\left(\sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'} + \mathbb{I}[k'=k, v'=v]\right)} \frac{\Gamma\left(\sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}\right)}{\prod_{k',v'} \Gamma\left(\tilde{\omega}^{-d,n}_{k',v'}\right)} \tag{58}$$

$$= \frac{\Gamma\left(\tilde{\omega}^{-d,n}_{k,v} + 1\right) \prod_{k' \neq k, v' \neq v} \Gamma\left(\tilde{\omega}^{-d,n}_{k',v'}\right)}{\Gamma\left(1 + \sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}\right)} \frac{\Gamma\left(\sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}\right)}{\prod_{k',v'} \Gamma\left(\tilde{\omega}^{-d,n}_{k',v'}\right)} \tag{59}$$

$$= \frac{\Gamma\left(\tilde{\omega}^{-d,n}_{k,v} + 1\right)}{\Gamma\left(1 + \sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}\right)} \frac{\Gamma\left(\sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}\right)}{\Gamma\left(\tilde{\omega}^{-d,n}_{k,v}\right)} \tag{60}$$

$$= \frac{\tilde{\omega}^{-d,n}_{k,v}}{\sum_{k',v'} \tilde{\omega}^{-d,n}_{k',v'}} \tag{61}$$

and

$$\frac{p(z_{-d,n} \mid z_{d,n}=k)}{p(z_{-d,n})} = \frac{B(\{\tilde{\alpha}^{-d,n}_{d',k'} + \mathbb{I}[d'=d, k'=k]\})}{\cancel{B(\{\alpha^{-d,n}_{d',k'}\})}} \frac{\cancel{B(\{\alpha^{-d,n}_{d',k'}\})}}{B(\{\tilde{\alpha}^{-d,n}_{d',k'}\})} \tag{62}$$

$$= \frac{\prod_{d',k'} \Gamma\left(\tilde{\alpha}^{-d,n}_{d',k'} + \mathbb{I}[d'=d, k'=k]\right)}{\Gamma\left(\sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'} + \mathbb{I}[d'=d, k'=k]\right)} \frac{\Gamma\left(\sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}\right)}{\prod_{d',k'} \Gamma\left(\tilde{\alpha}^{-d,n}_{d',k'}\right)} \tag{63}$$

$$= \frac{\Gamma\left(\tilde{\alpha}^{-d,n}_{d,k} + 1\right) \prod_{d' \neq d, k' \neq k} \Gamma\left(\tilde{\alpha}^{-d,n}_{d',k'}\right)}{\Gamma\left(1 + \sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}\right)} \frac{\Gamma\left(\sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}\right)}{\prod_{d',k'} \Gamma\left(\tilde{\alpha}^{-d,n}_{d',k'}\right)} \tag{64}$$

$$= \frac{\Gamma\left(\tilde{\alpha}^{-d,n}_{d,k} + 1\right)}{\Gamma\left(1 + \sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}\right)} \frac{\Gamma\left(\sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}\right)}{\Gamma\left(\tilde{\alpha}^{-d,n}_{d,k}\right)} \tag{65}$$

$$= \frac{\tilde{\alpha}^{-d,n}_{d,k}}{\sum_{d',k'} \tilde{\alpha}^{-d,n}_{d',k'}} \tag{66}$$

Finally, we combine the results in eqs. (61) and (66) back into eq. (52) to obtain the final collapsed Gibbs sampling expression

$$p(z_{d,n}=k \mid y_{d,n}=v, y_{-d,n}, z_{-d,n}) \propto \tilde{\alpha}^{-d,n}_{d,k} \tilde{\omega}^{-d,n}_{k,v}, \quad \text{where} \tag{67}$$

$$\tilde{\alpha}^{-d,n}_{d,k} = \alpha_k + N^{-d,n}_{d,k} \qquad\qquad \tilde{\omega}^{-d,n}_{k,v} = \frac{\omega_v + N^{-d,n}_{k,v}}{\sum_{v'} \omega_{v'} + N^{-d,n}_{k}}. \tag{68}$$

4 Expectation Maximization for LDA (MAP)

Expectation Maximization describes an iterative process to estimate the most likely model parameters given the existence of unobserved hidden variables in the model. In this section, we will

consider the MAP variant, where maximize the parameters' posterior likelihood:

$$\theta^*, \beta^* = \arg\max_{\theta,\beta} \log p(\theta, \beta|y) \tag{69}$$

$$= \arg\max_{\theta,\beta} \log p(y, \theta, \beta) \tag{70}$$

EM uses a parameterized distribution over the hidden variables, which is intended to serve as an approximation to their true posterior likelihood,

$$q(z; \phi) \simeq p(z \mid y, \theta, \beta) \tag{71}$$

and which, being discrete, we can parameterize directly as a categorical distribution

$$q(z_{d,n} = k; \phi) = \phi_{d,n,k} \tag{72}$$

The maximization objective in eq. (70) has a lower bound whose gap is the KL-divergence between the true hidden variable likelihood and its approximation:

$$\mathcal{L}(\{\theta, \beta\}, \phi) = \mathbb{E}_{q(z;\phi)} \left[\log \frac{p(y, z, \theta, \beta)}{q(z; \phi)} \right] \tag{73}$$

$$= \mathbb{E}_{q(z;\phi)} \left[\log \frac{p(y, \theta, \beta) p(z \mid, y, \theta, \beta)}{q(z; \phi)} \right] \tag{74}$$

$$= \log p(y, \theta, \beta) - \mathrm{KL}(q(z; \phi) \mid\mid p(z \mid y, \theta, \beta)) \tag{75}$$

$$\leq \log p(y, \theta, \beta) \tag{76}$$

Expectation Step In this step, we need to update the approximation parameters ϕ such that $p(z; \phi) = p(z \mid y, \theta, \beta)$. Given our parameterization choice in eq. (72), this amounts to

$$\phi_{d,n,k} = p(z_{d,n} = k \mid y, \theta, \beta) \tag{77}$$

$$= \mathbb{E}_{p(z|y,\theta,\beta)} \left[\mathbb{I}[z_{d,n} = k] \right] \tag{78}$$

Maximization Step For the M-step, given the distribution for hidden variable z, we want to find β and θ that maximize the log-likelihood below:

$$\begin{aligned} \theta, \beta &= \arg\max_{\theta,\beta} \mathcal{L}(\{\theta, \beta\}, \phi) \\ &= \arg\max_{\theta,\beta} \mathbb{E}_{q(z;\phi)} \left[\log \frac{p(y, z, \beta, \theta)}{q(z; \phi)} \right] \\ &= \arg\max_{\theta,\beta} \mathbb{E}_{q(z;\phi)} \left[\log p(y \mid z, \beta) + \log p(z \mid \theta) \right] + \log p(\theta) + \log p(\beta) \end{aligned} \tag{79}$$

For each part, the log-likelihood can be written as below:

$$\mathbb{E}_{q(z;\phi)}\left[\log p(y \mid z, \beta)\right] = \mathbb{E}_{q(z;\phi)}\left[\sum_{k,v} \log \beta_{k,v} \sum_{d,n} \mathbb{I}[y_{d,n} = v]\mathbb{I}[z_{d,n} = k]\right] \tag{80}$$

$$\mathbb{E}_{q(z;\phi)}\left[\log p(z \mid \theta)\right] = \mathbb{E}_{q(z;\phi)}\left[\sum_{k} \log \theta_{d,k} \sum_{n} \mathbb{I}[z_{d,n} = k]\right] \tag{81}$$

$$\log p(\theta_d) = \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{d,k} - \log B(\alpha) \tag{82}$$

$$\log p(\beta_k) = \sum_{v=1}^{V} (\omega_v - 1) \log \beta_{k,v} - \log B(\omega) \tag{83}$$

For the maximization step, as the $\log B(\alpha)$ and $\log B(\omega)$ are constant, we can ignore them. Then for the optimal value calculation, the partial derivative should be zero. As there is a constraint that $\forall d \sum_l \theta_{d,l} = 1$, we can use Lagrange multiplier to find the optimal value as follows:

$$\mathcal{L}(\theta_d, \lambda) = \mathcal{L}(\{\theta, \beta\}, \phi) + \lambda(\theta_{d,1} + \cdots, +\theta_{d,K} - 1) \tag{84}$$

If we want to maximize $\mathcal{L}(\theta_d, \lambda)$, the partial derivatives should satisfy:

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta_d, \lambda)}{\partial \theta_{d,l}} &= \frac{\alpha_l - 1 + \sum_n \phi_{d,n,l}}{\theta_{d,l}} + \lambda = 0 \quad l = 1, 2, \cdots, K \\ \frac{\partial \mathcal{L}(\theta_d, \lambda)}{\partial \lambda} &= \theta_{d,1} + \cdots, +\theta_{d,K} - 1 = 0 \end{aligned} \tag{85}$$

$$\Rightarrow \theta_{d,l} = \frac{\alpha_l - 1 + \sum_n \phi_{d,n,l}}{-\lambda} \quad l = 1, 2, \cdots, K \tag{86}$$

$$\Rightarrow \sum_{l=1}^{K} \theta_{d,l} = 1 \tag{87}$$

$$\Rightarrow \sum_{l=1}^{K} \frac{\alpha_l - 1 + \sum_n \phi_{d,n,l}}{-\lambda} = 1 \tag{88}$$

$$\Rightarrow -\lambda = \sum_{l=1}^{K} \left(\alpha_l - 1 + \sum_n \phi_{d,n,l}\right) \tag{89}$$

So the parameter $\theta_{d,k}$ satisfy:

$$\theta_{d,k} = \frac{\alpha_k - 1 + \sum_n \phi_{d,n,k}}{\sum_l \left(\alpha_l - 1 + \sum_n \phi_{d,n,l}\right)} \tag{90}$$

We also know that $\forall k \sum_v \beta_{k,v} = 1$. We can similarly set up a Lagrange multiplier for $\beta_{k,v}$ as follows:

$$\mathcal{L}(\beta_k, \eta) = \mathcal{L}(\{\theta, \beta\}, \phi) + \eta(\beta_{k,1} + \cdots, +\beta_{k,V} - 1) \tag{91}$$

If we want to maximize $\mathcal{L}(\beta_k, \eta)$, the partial derivatives should satisfy:

$$\begin{aligned}\frac{\partial \mathcal{L}(\beta_k,\eta)}{\partial \beta_{k,l}} &= \frac{\omega_l - 1 + \sum_n \phi_{d,n,l}\mathbb{I}[y_{d,n}=v]}{\beta_{k,l}} + \eta = 0, \quad l = 1, 2, \cdots, V \\ \frac{\partial \mathcal{L}(\beta_k,\eta)}{\partial \eta} &= \beta_{k,1} + \cdots, + \beta_{k,V} - 1 = 0\end{aligned} \tag{92}$$

$$\Rightarrow \beta_{k,l} = \frac{\omega_l - 1 + \sum_n \phi_{d,n,l}\mathbb{I}[y_{d,n}=v]}{-\eta}, \quad l = 1, 2, \cdots, V \tag{93}$$

$$\Rightarrow \sum_{l=1}^{V} \beta_{k,l} = 1 \tag{94}$$

$$\Rightarrow \sum_{l=1}^{V} \frac{\omega_l - 1 + \sum_n \phi_{d,n,l}\mathbb{I}[y_{d,n}=v]}{-\eta} = 1 \tag{95}$$

$$\Rightarrow -\eta = \sum_{l=1}^{V} \left(\omega_l - 1 + \sum_n \phi_{d,n,l}\mathbb{I}[y_{d,n}=v] \right) \tag{96}$$

So the parameter $\beta_{k,v}$ satisfy the following equation, and we have the necessary updates for β and θ.

$$\beta_{k,v} = \frac{\omega_v - 1 + \sum_{d,n}\mathbb{I}[y_{d,n}=v]\phi_{d,n,k}}{\sum_l(\omega_l - 1 + \sum_{d,n}\mathbb{I}[y_{d,n}=v]\phi_{d,n,l})} \tag{97}$$

5 Sequential Monte Carlo for LDA

As we recall, Sequential Monte Carlo divides (and conquers!) the problem of sampling from a high-dimensional space into a sequence of low-dimensional sampling problems. To apply Sequential Monte Carlo, we have to first restructure the LDA domain as a sequential one. We do this by rethinking the body of all documents as a single stream of words which come one after the other; note that we are only defining a global word order, and want to retain the information about each word's original document.

The general formulation of Sequential Monte Carlo requires two domain-specific densities to be defined: the sequence of unnormalized densities $\gamma_1(x_1), \ldots, \gamma_T(x_{1:T})$, which are, for the "sequenced" LDA model,

$$\gamma_n(z_{1:n}) = p(y_{1:n}, z_{1:n} \mid \beta)\,, \tag{98}$$

and the proposal density $q\left(x_t \mid x_{1:t-1}^{a_{t-1}^s}\right)$, which is

$$q(z_n \mid z_{1:n-1}) = p(z_n \mid y_{1:n-1}, z_{1:n-1}, \beta)\,. \tag{99}$$

Having determined these quantities for the LDA model, there are only two model-specific computations required to run Sequential Monte Carlo:

1. Sample $z_n^s \sim q(z_n \mid z_{1:n-1})$ from eq. (99). This can be done in a similar fashion to the Collapsed Gibbs sampling in eq. (52), with the only adjustment being—in order to avoid "peeking" into the future—that the counts used in eq. (52) should only consider documents up to the d-th and words up to the n-th in the d-th document.

2. Compute the unnormalized sample weights

$$w_n^s = \frac{\gamma_n\left(z_{1:n}^s\right)}{\gamma_{n-1}\left(z_{1:n-1}^{a_{n-1}^s}\right) q\left(z_n^s \mid y_{1:n-1}, z_{1:n-1}^{a_{n-1}^s}\right)} \tag{100}$$

$$= \frac{p\left(y_{1:n}, z_{1:n-1}^{a_{n-1}^s}, z_n^s \mid \beta\right)}{p\left(y_{1:n-1}, z_{1:n-1}^{a_{n-1}^s} \mid \beta\right) p\left(z_n^s \mid z_{1:n-1}^{a_{n-1}^s}, y_{1:n-1}, \beta\right)} \quad \text{(by eqs. (98) and (99))} \tag{101}$$

$$= \frac{p\left(y_n \mid z_n^s, \beta\right) p\left(z_n^s \mid y_{1:n-1}, z_{1:n-1}^{a_{n-1}^s} \mid \beta\right) p\left(y_{1:n-1}, z_{1:n-1}^{a_{n-1}^s} \mid \beta\right)}{p\left(y_{1:n-1}, z_{1:n-1}^{a_{n-1}^s} \mid \beta\right) p\left(z_n^s \mid z_{1:n-1}^{a_{n-1}^s}, y_{1:n-1}, \beta\right)} \tag{102}$$

$$= p\left(y_n \mid z_n^s, \beta\right) \tag{103}$$

$$= \prod_{k,v} \beta_{k,v}^{\mathbb{I}[y_{d,n}=v]\mathbb{I}[z_{d,n}=k]} \tag{104}$$

6 Thought Experiment: Hamiltonian Monte Carlo for LDA

As a thought experiment, we will consider whether Hamiltonian Monte Carlo (HMC) can be applied to perform inference in LDA. HMC is an auxiliary variable method inspired by Hamiltonian dynamics, which describe a system's kinetic and potential state change over time. The involved dynamics are necessarily continuous, which means we can not apply this method to sample the discrete topic assignments. Luckily, we can split the sampling process in two steps,

$$\theta, \beta \sim p(\theta, \beta \mid y) \tag{105}$$

$$z \sim p(z \mid y, \theta, \beta) \tag{106}$$

and use eq. (36) to sample the conditional topic assignments. So for the rest of this section, we will focus on using Hamiltonian Monte Carlo to sample the model parameters from $p(\theta, \beta \mid y)$ while marginalizing over the topic assignments.

We should emphasize that marginalizing over the topic assignments requires a full pass over all the data, which makes this a slow process for big data sets (e.g. all 5M articles in Wikipedia). However, technically speaking it is not possible to define a HMC algorithm w.r.t. the density,

$$\gamma(\theta, \beta) = p(y, \theta, \beta) \tag{107}$$

The potential function $U(\theta, \beta)$ is defined as being proportional to the neg-log gamma function, and we only need to derive its gradients,

$$U(\theta, \beta) = -\log \gamma(\theta, \beta) \tag{108}$$

$$\nabla U_{\theta,\beta}(\theta, \beta) = -\nabla_{\theta,\beta} \log \gamma(\theta, \beta) \tag{109}$$

The probability for the distribution $p(y, \theta, \beta)$:

$$p(y, \theta, \beta) = \exp(-U(\theta, \beta)) \tag{110}$$

So the Hamiltonian Monte Carlo method needs the gradient of log joint distribution which can be found by marginalizing over z in the joint density found in eq. (1).

$$\nabla_{\theta,\beta} U(\theta, \beta) = -\nabla_{\theta,\beta} \log p(y, \theta, \beta) \tag{111}$$

In the stochastic variational inference section, we will discuss a method for computing the gradient with respect to β. The gradients with respect to θ can be defined similarly.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[2] Scott Deerwester et al. "Indexing by latent semantic analysis". In: *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41.6 (1990), pp. 391–407.

[3] Thomas Hofmann. "Probabilistic Latent Semantic Indexing". In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. Berkeley, California, USA: ACM, 1999, pp. 50–57. ISBN: 1-58113-096-1. DOI: 10.1145/312624.312649. URL: http://doi.acm.org/10.1145/312624.312649.

[4] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[5] Yee Whye Teh et al. "Hierarchical Dirichlet Processes". In: *Journal of the American Statistical Society*. 2006.