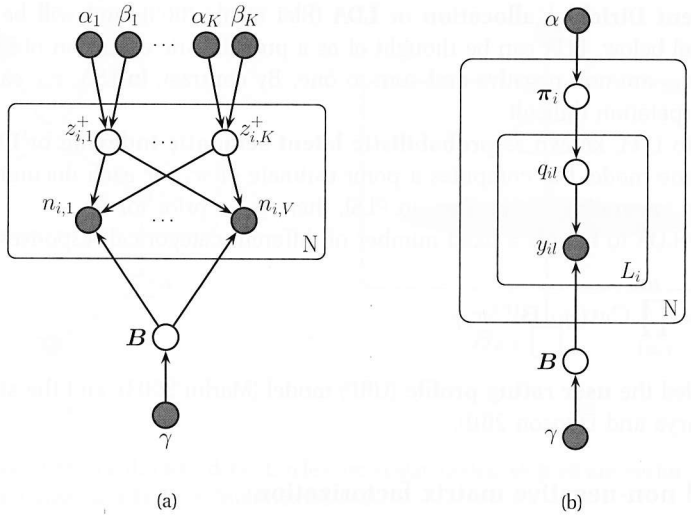

Figure 27.2 (a) Gaussian-Poisson (GAP) model. (b) Latent Dirichlet allocation (LDA) model.

27.3 Latent Dirichlet allocation (LDA)

In this section, we explain the **latent Dirichlet allocation** or **LDA** (Blei et al. 2003) model in detail.

27.3.1 Basics

In a mixture of multinoullis, every document is assigned to a single topic, $q_i \in \{1, \ldots, K\}$, drawn from a global distribution $\boldsymbol{\pi}$. In LDA, every word is assigned to its own topic, $q_{il} \in \{1, \ldots, K\}$, drawn from a document-specific distribution $\boldsymbol{\pi}_i$. Since a document belongs to a distribution over topics, rather than a single topic, the model is called an **admixture mixture** or **mixed membership model** (Erosheva et al. 2004). This model has many other applications beyond text analysis, e.g., genetics (Pritchard et al. 2000), health science (Erosheva et al. 2007), social network analysis (Airoldi et al. 2008), etc.

Adding conjugate priors to the parameters, the full model is as follows:[1]

$$\boldsymbol{\pi}_i|\alpha \sim \text{Dir}(\alpha \mathbf{1}_K) \tag{27.18}$$

$$q_{il}|\boldsymbol{\pi}_i \sim \text{Cat}(\boldsymbol{\pi}_i) \tag{27.19}$$

$$\mathbf{b}_k|\gamma \sim \text{Dir}(\gamma \mathbf{1}_V) \tag{27.20}$$

$$y_{il}|q_{il} = k, \mathbf{B} \sim \text{Cat}(\mathbf{b}_k) \tag{27.21}$$

This is illustrated in Figure 27.2(b). We can marginalize out the q_i variables, thereby creating a

1. Our notation is similar to the one we use elsewhere in this book, but is different from that used by most LDA papers. They typically use w_{nd} for the identity of word n in document d, z_{nd} to represent the discrete indicator, $\boldsymbol{\theta}_d$ as the continuous latent vector for document d, and $\boldsymbol{\beta}_k$ as the k'th topic vector.

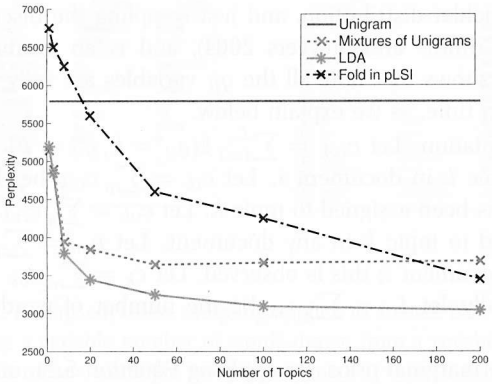

Figure 27.6 Perplexity vs number of topics on the TREC AP corpus for various language models. Based on Figure 9 of (Blei et al. 2003). Figure generated by `bleiLDAperplexityPlot`.

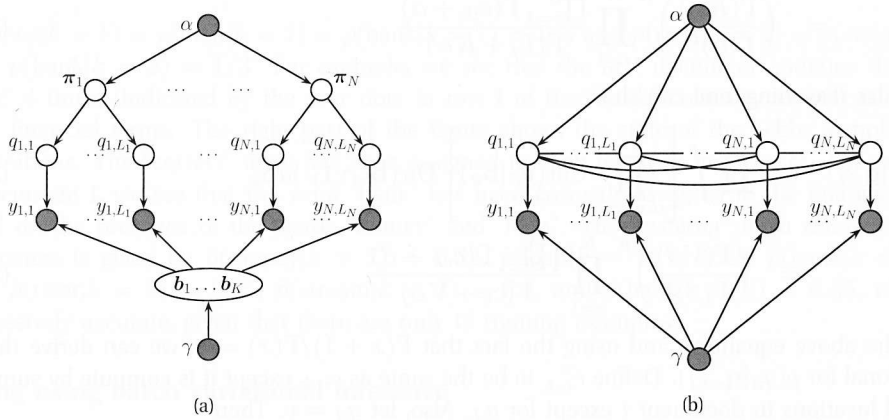

Figure 27.7 (a) LDA unrolled for N documents. (b) Collapsed LDA, where we integrate out the $\boldsymbol{\pi}_i$ and the $\mathbf{b}_k$.

27.3.4 Fitting using (collapsed) Gibbs sampling

It is straightforward to derive a Gibbs sampling algorithm for LDA. The full conditionals are as follows:

$$p(q_{il} = k|\cdot) \propto \exp[\log \pi_{ik} + \log b_{k,x_{il}}] \tag{27.30}$$

$$p(\boldsymbol{\pi}_i|\cdot) = \text{Dir}(\{\alpha_k + \sum_l \mathbb{I}(z_{il} = k)\}) \tag{27.31}$$

$$p(\mathbf{b}_k|\cdot) = \text{Dir}(\{\gamma_v + \sum_i \sum_l \mathbb{I}(x_{il} = v, z_{il} = k)\}) \tag{27.32}$$

However, one can get better performance by analytically integrating out the $\boldsymbol{\pi}_i$'s and the $\mathbf{b}_k$'s,

both of which have a Dirichlet distribution, and just sampling the discrete q_{il}'s. This approach was first suggested in (Griffiths and Steyvers 2004), and is an example of **collapsed Gibbs sampling**. Figure 27.7(b) shows that now all the q_{il} variables are fully correlated. However, we can sample them one at a time, as we explain below.

First, we need some notation. Let $c_{ivk} = \sum_{l=1}^{L_i} \mathbb{I}(q_{il} = k, y_{il} = v)$ be the number of times word v is assigned to topic k in document i. Let $c_{ik} = \sum_v c_{ivk}$ be the number of times any word from document i has been assigned to topic k. Let $c_{vk} = \sum_i c_{ivk}$ be the number of times word v has been assigned to topic k in any document. Let $n_{iv} = \sum_k c_{ivk}$ be the number of times word v occurs in document i; this is observed. Let $c_k = \sum_v c_{vk}$ be the number of words assigned to topic k. Finally, let $L_i = \sum_k c_{ik}$ be the number of words in document i; this is observed.

We can now derive the marginal prior. By applying Equation 5.24, one can show that

$$p(\mathbf{q}|\alpha) = \prod_i \int \left[\prod_{l=1}^{L_i} \text{Cat}(q_{il}|\boldsymbol{\pi}_i)\right] \text{Dir}(\boldsymbol{\pi}_i|\alpha \mathbf{1}_K) d\boldsymbol{\pi}_i \tag{27.33}$$

$$= \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^N \prod_{i=1}^N \frac{\prod_{k=1}^K \Gamma(c_{ik} + \alpha)}{\Gamma(L_i + K\alpha)} \tag{27.34}$$

By similar reasoning, one can show

$$p(\mathbf{y}|\mathbf{q}, \gamma) = \prod_k \int \left[\prod_{il:q_{il}=k} \text{Cat}(y_{il}|\mathbf{b}_k)\right] \text{Dir}(\mathbf{b}_k|\gamma \mathbf{1}_V) d\mathbf{b}_k \tag{27.35}$$

$$= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^K \frac{\prod_{v=1}^V \Gamma(c_{vk} + \beta)}{\Gamma(c_k + V\beta)} \tag{27.36}$$

From the above equations, and using the fact that $\Gamma(x+1)/\Gamma(x) = x$, we can derive the full conditional for $p(q_{il}|\mathbf{q}_{-i,l})$. Define c^-_{ivk} to be the same as c_{ivk} except it is compute by summing over all locations in document i except for q_{il}. Also, let $y_{il} = v$. Then

$$p(q_{i,l} = k|\mathbf{q}_{-i,l}, \mathbf{y}, \alpha, \gamma) \propto \frac{c^-_{v,k} + \gamma}{c^-_k + V\gamma} \frac{c^-_{i,k} + \alpha}{L_i + K\alpha} \tag{27.37}$$

We see that a word in a document is assigned to a topic based both on how often that word is generated by the topic (first term), and also on how often that topic is used in that document (second term).

Given Equation 27.37, we can implement the collapsed Gibbs sampler as follows. We randomly assign a topic to each word, $q_{il} \in \{1, \ldots, K\}$. We can then sample a new topic as follows: for a given word in the corpus, decrement the relevant counts, based on the topic assigned to the current word; draw a new topic from Equation 27.37, update the count matrices; and repeat. This algorithm can be made efficient since the count matrices are very sparse.

27.3.5 Example

This process is illustrated in Figure 27.8 on a small example with two topics, and five words. The left part of the figure illustrates 16 documents that were sampled from the LDA model using

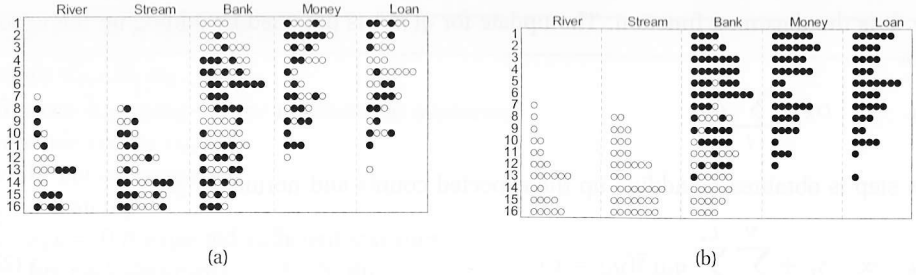

Figure 27.8 Illustration of (collapsed) Gibbs sampling applied to a small LDA example. There are $N = 16$ documents, each containing a variable number of words drawn from a vocabulary of $V = 5$ words, There are two topics. A white dot means word the word is assigned to topic 1, a black dot means the word is assigned to topic 2. (a) The initial random assignment of states. (b) A sample from the posterior after 64 steps of Gibbs sampling. Source: Figure 7 of (Steyvers and Griffiths 2007). Used with kind permission of Tom Griffiths.

$p(\text{money}|k = 1) = p(\text{loan}|k = 1) = p(\text{bank}|k = 1) = 1/3$ and $p(\text{river}|k = 2) = p(\text{stream}|k = 2) = p(\text{bank}|k = 2) = 1/3$. For example, we see that the first document contains the word "bank" 4 times (indicated by the four dots in row 1 of the "bank" column), as well as various other financial terms. The right part of the figure shows the state of the Gibbs sampler after 64 iterations. The "correct" topic has been assigned to each token in most cases. For example, in document 1, we see that the word "bank" has been correctly assigned to the financial topic, based on the presence of the words "money" and "loan". The posterior mean estimate of the parameters is given by $\hat{p}(\text{money}|k = 1) = 0.32$, $\hat{p}(\text{loan}|k = 1) = 0.29$, $\hat{p}(\text{bank}|k = 1) = 0.39$, $\hat{p}(\text{river}|k = 2) = 0.25$, $\hat{p}(\text{stream}|k = 2) = 0.4$, and $\hat{p}(\text{bank}|k = 2) = 0.35$, which is impressively accurate, given that there are only 16 training examples.

27.3.6 Fitting using batch variational inference

A faster alternative to MCMC is to use variational EM. (We cannot use exact EM since exact inference of $\boldsymbol{\pi}_i$ and $\mathbf{q}_i$ is intractable.) We give the details below.

27.3.6.1 Sequence version

Following (Blei et al. 2003), we will use a fully factorized (mean field) approximation of the form

$$q(\boldsymbol{\pi}_i, \mathbf{q}_i) = \text{Dir}(\boldsymbol{\pi}_i|\tilde{\boldsymbol{\pi}}_i) \prod_l \text{Cat}(q_{il}|\tilde{\mathbf{q}}_{il}) \tag{27.38}$$

We will follow the usual mean field recipe. For $q(q_{il})$, we use Bayes' rule, but where we need to take expectations over the prior:

$$\tilde{q}_{ilk} \propto b_{y_{i,l},k} \exp(\mathbb{E}[\log \pi_{ik}]) \tag{27.39}$$

where

$$\mathbb{E}[\log \pi_{ik}] = \psi_k(\tilde{\boldsymbol{\pi}}_{i\cdot}) \triangleq \Psi(\tilde{\pi}_{ik}) - \Psi(\sum_{k'} \tilde{\pi}_{ik'}) \tag{27.40}$$