Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling

Bob Carpenter, LingPipe, Inc., carp@lingpipe.com

Revision 1.4, September 27, 2010

This note shows how to integrate out the multinomial parameters for latent Dirichlet allocation (LDA) and naive Bayes (NB) models. This allows us to perform Gibbs sampling without taking multinomial parameter samples. Although the conjugacy of the Dirichlet priors makes sampling the multinomial parameters relatively straightforward, sampling on a topic-by-topic basis provides two advantages. First, it means that all samples are drawn from simple discrete distributions with easily calculated parameters. Second, and more importantly, collapsing supports fully stochastic Gibbs sampling where the model is updated after each word (in LDA) or document (in NB) is assigned a topic. Typically, more stochastic sampling leads to quicker convergence to the stationary state of the Markov chain made up of the Gibbs samples.

Both the LDA and NB models are topic models, where words are generated based on topic-specific multinomials. The main difference is that LDA assumes each word in a document is drawn from a mixture of topics, whereas NB assumes each word in a document is drawn from a single topic.

In a hierarchical model, the topic and word priors would themselves be estimated. Here, we assume the priors are fixed hyperparameters in both the NB and LDA models.

1 LDA Model

1.1 LDA Sampling Model

$M \in \mathbb{N}_+$ is the number of documents. $N_m \in \mathbb{N}_+$ is the number of words in the m-th document. J is the number of distinct words. K the number of topics. $y_{m,n} \in 1{:}J$ is the n-th word of the m-th document and $z_{m,n} \in 1{:}K$ is the topic to which it is assigned. $\theta_m \in [0,1]^K$ is the topic distribution for document m. $\varphi_k \in [0,1]^J$ is the word distribution for topic k. $\alpha \in \mathbb{R}_+^K$ is the vector of prior counts (plus 1) for topics in documents and $\beta \in \mathbb{R}_+^J$ is the vector of prior counts (plus 1) for words in a topic.

In sampling notation, we draw the word distribution for topic k by

$$\varphi_k \sim \mathsf{Dir}(\beta) \text{ for } 1 \leq k \leq K \tag{1}$$

For each document m, we draw its topic distribution,

$$\theta_m \sim \mathsf{Dir}(\alpha) \text{ for } 1 \leq m \leq M \tag{2}$$

For each word n in document m, we first draw the topic $z_{m,n}$ from the distribution over topics for the document m,

$$z_{m,n} \sim \mathsf{Disc}(\theta_m) \text{ for } 1 \leq m \leq M \text{ and } 1 \leq n \leq N_m \tag{3}$$

then draw the word $y_{m,n}$ itself from the word distribution for the word's topic, $z_{m,n}$,

$$y_{m,n} \sim \mathsf{Disc}(\varphi_{z_{m,n}}) \text{ for } 1 \leq m \leq M \text{ and } 1 \leq n \leq N_m \tag{4}$$

1.2 LDA Joint Probability

Given the model, the joint probability for all of the parameters in the LDA model is

$$p(y, z, \theta, \varphi|\alpha, \beta) \tag{5}$$

$$= p(\varphi|\beta)\ p(\theta|\alpha)\ p(z|\theta)\ p(y|\varphi, z) \tag{6}$$

$$= \prod_{k=1}^{K} p(\varphi_k|\beta) \times \prod_{m=1}^{M} p(\theta_m|\alpha) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(y_{m,n}|\varphi_{z_{m,n}}) \tag{7}$$

$$= \prod_{k=1}^{K} \mathsf{Dir}(\varphi_k|\beta) \times \prod_{m=1}^{M} \mathsf{Dir}(\theta_m|\alpha) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathsf{Disc}(z_{m,n}|\theta_m) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathsf{Disc}(y_{m,n}|\varphi_{z_{m,n}}) \tag{8}$$

$c_{k,m,j}$ is the number of times word j is assigned to topic k in document m. Summing out various indices, $c_{k,*,j}$ is the number of times word j is assigned to topic k in any document, $c_{k,m,*}$ the number of words in document m assigned to topic k, and $c_{k,*,*}$ the total number of words in all documents assigned to topic k.

$$c_{k,m,j} = \sum_{n=1}^{N_m} \mathrm{I}(z_{m,n} = k\ \&\ y_{m,n} = j) \tag{9}$$

$$c_{k,*,j} = \sum_{m=1}^{M} c_{k,m,j} \qquad c_{k,m,*} = \sum_{j=1}^{J} c_{k,m,j} \qquad c_{k,*,*} = \sum_{m=1}^{M}\sum_{j=1}^{J} c_{k,m,j} \tag{10}$$

1.3 Integrating out Multinomials in LDA

The collapsed sampler needs to compute the probability of topic $z_{a,b}$ being assigned to $y_{a,b}$, the b-th word of the a-th document, given $z_{-(a,b)}$, all the other topic assignments to all the other words.

$$p(z_{a,b}|z_{-(a,b)}, y, \alpha, \beta) \tag{11}$$

By the definition of conditional probability,

$$= \frac{p(z_{a,b}, z_{-(a,b)}, y|\alpha, \beta)}{p(z_{-(a,b)}, y|\alpha, \beta)} \tag{12}$$

Remove the denominator, which does not depend on $z_{a,b}$,

$$\propto p(z_{a,b}, z_{-(a,b)}, y|\alpha, \beta) \tag{13}$$

Note that $z_{a,b}, z_{-(a,b)}$ is just z,

$$= p(y, z|\alpha, \beta) \tag{14}$$

Using the sum rule (or rule of total probability), integrate out the topic distributions for each document, θ, and the word distributions for each topic, φ,

$$= \int \int p(y, z, \theta, \varphi | \alpha, \beta) \; d\theta \; d\varphi \tag{15}$$

Expand the integrand given the model defined in (6),

$$= \int \int p(\varphi|\beta) \; p(\theta|\alpha) \; p(z|\theta) \; p(y|\varphi, z) \; d\theta \; d\varphi \tag{16}$$

Because $\int \int x \times y \; dx \; dy = (\int x \; dx) \times (\int y \; dy)$, we may separate the integrals based on the terms being integrated,

$$= \int p(z|\theta) \; p(\theta|\alpha) \; d\theta \;\; \times \;\; \int p(y|\varphi, z) \; p(\varphi|\beta) \; d\varphi \tag{17}$$

And then expand out the terms again according to the independence assumptions in (7),

$$= \int \prod_{m=1}^{M} p(z_m|\theta_m) \; p(\theta_m|\alpha) \; d\theta \;\; \times \;\; \int \prod_{k=1}^{K} p(\varphi_k|\beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(y_{m,n}|\varphi_{z_{m,n}}) \; d\varphi \tag{18}$$

For the same reason as we could separate two products, we may separate multiple products when other terms are constant, so we may distribute the multivariate integrals through the products over the dimensions,

$$= \prod_{m=1}^{M} \int p(z_m|\theta_m) \; p(\theta_m|\alpha) \; d\theta_m \;\; \times \;\; \prod_{k=1}^{K} \int p(\varphi_k|\beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(y_{m,n}|\varphi_{z_{m,n}}) \; d\varphi_k \tag{19}$$

Expand out the Dirichlet priors and the discrete distributions according to their usual definitions,

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \prod_{n=1}^{N_m} \theta_{m,z_{m,n}} \;\; d\theta_m \tag{20}$$

$$\times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\sum_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j - 1} \prod_{m=1}^{M} \prod_{n=1}^{N_m} \varphi_{z_{m,n}, y_{m,n}} \; d\varphi_k$$

Because $x^a x^b = x^{a+b}$, we replace the innermost products over words in a document N_m by exponeniating to the sum of the counts,

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \prod_{k=1}^{K} \theta_{m,k}^{c_{k,m,*}} \;\; d\theta_m \tag{21}$$

$$\times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\sum_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j - 1} \prod_{j=1}^{J} \varphi_{k,j}^{c_{k,*,j}} \; d\varphi_k$$

For the same reason, merge the two products,

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k + c_{k,m,*} - 1} \;\; d\theta_m \;\; \times \;\; \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\sum_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j + c_{k,*,j} - 1} \; d\varphi_k \tag{22}$$

Next, multiply by a constant equal to one (consisting of two inverse fractions), and distribute the integral over the original constant Γ-function fraction for the priors,

$$= \prod_{m=1}^{M} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,*} + \alpha_k)} \int \frac{\Gamma(\sum_{k=1}^{K} c_{k,m,*} + \alpha_k)}{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k + c_{k,m,*} - 1} \; d\theta_m \tag{23}$$

$$\times \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\sum_{j=1}^{J} \Gamma(\beta_j)} \frac{\prod_{j=1}^{J} \Gamma(c_{k,*,j} + \beta_j)}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \int \frac{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)}{\prod_{j=1}^{J} \Gamma(c_{k,*,j} + \beta_j)} \varphi_{k,j}^{\beta_j + c_{k,*,j} - 1} d\varphi_k$$

Note that both integrals are over the entire support of Dirichlet densities, so they both evaluate to 1, and hence drop out of the products,

$$= \prod_{m=1}^{M} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,*} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\sum_{j=1}^{J} \Gamma(\beta_j)} \frac{\prod_{j=1}^{J} \Gamma(c_{k,*,j} + \beta_j)}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \tag{24}$$

Drop out the remaining Γ functions which only depend on the (constant) hyperparameters α and β,

$$\propto \prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,*} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_{j=1}^{J} \Gamma(c_{k,*,j} + \beta_j)}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \tag{25}$$

Next, split apart the products to pull out the terms dependent on the current sample position (a, b),

$$= \prod_{m \neq a} \frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,*} + \alpha_k)} \times \frac{\prod_{k=1}^{K} \Gamma(c_{k,a,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,a,*} + \alpha_k)} \tag{26}$$

$$\times \prod_{k=1}^{K} \frac{\prod_{j \neq y_{a,b}} \Gamma(c_{k,*,j} + \beta_j) \times \Gamma(c_{k,*,y_{a,b}} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)}$$

Then drop terms that don't depend on (a, b),

$$\propto \frac{\prod_{k=1}^{K} \Gamma(c_{k,a,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,a,*} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(c_{k,*,y_{a,b}} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \tag{27}$$

Let $c^{-(a,b)}$ be defined the same way as c, only without the counts for position (a, b), then note that for counts that don't include position (a, b), that $c^{-(a,b)} = c$, and for ones that do include a count, the value is 1 plus the value given by $c^{-(a,b)}$,

$$\propto \frac{\prod_{k \neq z_{a,b}} \Gamma(c_{k,a,*}^{-(a,b)} + \alpha_k) \times \Gamma(c_{z_{a,b},a,*}^{-(a,b)} + \alpha_{z_{a,b}} + 1)}{\Gamma(1 + \sum_{k=1}^{K} c_{k,a,*}^{-(a,b)} + \alpha_k)} \tag{28}$$

$$\times \prod_{k \neq z_{a,b}} \frac{\Gamma(c_{k,*,y_{a,b}}^{-(a,b)} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \times \frac{\Gamma(c_{z_{a,b},*,y_{a,b}}^{-(a,b)} + \beta_{y_{a,b}} + 1)}{\Gamma(1 + \sum_{j=1}^{J} c_{z_{a,b},*,j}^{-(a,b)} + \beta_j)}$$

Using the fact that $\Gamma(x+1) = x \times \Gamma(x)$, expand out the incremented terms depending on (a, b),

$$= \frac{\prod_{k \neq z_{a,b}} \Gamma(c^{-(a,b)}_{k,a,*} + \alpha_k) \ \times \ \Gamma(c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}}) \ \times \ (c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}})}{\Gamma(1 + \sum_{k=1}^{K} c^{-(a,b)}_{k,a,*} + \alpha_k)} \tag{29}$$

$$\times \prod_{k \neq z_{a,b}} \frac{\Gamma(c^{-(a,b)}_{k,*,y_{a,b}} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \times \frac{\Gamma(c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}}) \ \times \ (c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c^{-(a,b)}_{z_{a,b},*,j} + \beta_j) \ \times \ \sum_{j=1}^{J} (c^{-(a,b)}_{z_{a,b},*,j} + \beta_j)}$$

Then refold the leftover Γ terms back into the general products,

$$= \frac{\prod_{k=1}^{K} \Gamma(c^{-(a,b)}_{k,a,*} + \alpha_k) \ \times \ (c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}})}{\Gamma(1 + \sum_{k=1}^{K} c^{-(a,b)}_{k,a,*} + \alpha_k)} \tag{30}$$

$$\times \prod_{k=1}^{K} \frac{\Gamma(c^{-(a,b)}_{k,*,y_{a,b}} + \beta_{y_{a,b}})}{\Gamma(\sum_{j=1}^{J} c_{k,*,j} + \beta_j)} \times \frac{c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}}}{\sum_{j=1}^{J} (c^{-(a,b)}_{z_{a,b},*,j} + \beta_j)}$$

Which along with the topic denominator, may be dropped because they are constant,

$$\propto \frac{(c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}}) \times (c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}})}{\sum_{j=1}^{J} (c^{-(a,b)}_{z_{a,b},*,j} + \beta_j)} \tag{31}$$

And finally we replace the denominator with its shorthand form, nothing that the β_j term is counted once per word,

$$\propto \frac{(c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}}) \times (c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}})}{c^{-(a,b)}_{z_{a,b},*,*} + \sum_{j=1}^{J} \beta_j} \tag{32}$$

If the β_j are all the same value, say γ, (as in the LingPipe implementation), then the term $\sum_{j=1}^{J} \beta_j = J \times \gamma$

The first multiplicand in the numerator, $c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}}$, is just the number of other words in document a that have been assigned to topic $z_{a,b}$ plus the topic prior. The second multiplicand in the numerator, $c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}}$, is the number of times the current word $y_{a,b}$ has been assigned to topic $z_{a,b}$ plus the word prior. The denominator just normalizes the second term to a probability.

Because everyting is only taken up to proportionality, it only remains to normalize,

$$p(z_{a,b} | z_{-(a,b)}, y, \alpha, \beta) = \frac{\left(\frac{(c^{-(a,b)}_{z_{a,b},a,*} + \alpha_{z_{a,b}}) \times (c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + \beta_{y_{a,b}})}{c^{-(a,b)}_{z_{a,b},*,*} + J \times \beta_j} \right)}{\left(\sum_{k=1}^{K} \frac{(c^{-(a,b)}_{k,a,*} + \alpha_k) \times (c^{-(a,b)}_{k,*,y_{a,b}} + \beta_{y_{a,b}})}{c^{-(a,b)}_{k,*,*} + J \times \beta_j} \right)} \tag{33}$$

These values are all easily accumulated and updated during Gibbs sampling; just decrement to compute $c^{-(a,b)}$ before a topic assignment and then increment after the topic assignment. The denominator is computed by summation as defined.

2 Naive Bayes Model

2.1 NB Sampling Model

$M \in \mathbb{N}_+$ is the number of documents. $N_m \in \mathbb{N}_+$ is the number of words in the m-th document. J is the number of distinct words. K is the number of topics. $y_{m,n} \in 1{:}J$ is the n-th word of the mth document. $z_m \in 1{:}K$ is the topic assigned to document m. $\theta \in [0,1]^K$ is the global topic distribution. $\varphi_k \in [0,1]^J$ is the word distribution for topic k. $\alpha \in \mathbb{R}_+^K$ is the vector of prior counts (plus 1) for topics and $\beta \in \mathbb{R}_+^J$ is the vector of prior counts (plus 1) for words in a topic.

In sampling notation, draw the word distribution for topic k by

$$\varphi_k \sim \mathsf{Dir}(\beta) \text{ for } 1 \leq k \leq K \tag{34}$$

Draw the overall topic distribution by

$$\theta \sim \mathsf{Dir}(\alpha) \tag{35}$$

For each document m, first draw its topic,

$$z_m \sim \mathsf{Disc}(\theta) \text{ for } 1 \leq m \leq M \tag{36}$$

and then draw each word n from document m based on the document's topic,

$$y_{m,n} \sim \mathsf{Disc}(\varphi_{z_m}) \text{ for } 1 \leq m \leq M \text{ and } 1 \leq n \leq N_m \tag{37}$$

2.2 NB Joint Probability

$$p(y, z, \theta, \varphi | \alpha, \beta) \tag{38}$$

$$= p(\varphi|\beta)\ p(\theta|\alpha)\ p(z|\theta)\ p(y|z,\varphi) \tag{39}$$

$$= \prod_{k=1}^{K} p(\varphi_k|\beta) \times p(\theta|\alpha) \times \prod_{m=1}^{M} p(z_m|\theta) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(y_{m,n}|\theta_{z_m}) \tag{40}$$

$$= \prod_{k=1}^{K} \mathsf{Dir}(\varphi_k|\beta) \times \mathsf{Dir}(\theta|\alpha) \times \prod_{m=1}^{M} \mathsf{Disc}(z_m|\theta) \times \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathsf{Disc}(y_{m,n}|\theta_{z_m}) \tag{41}$$

d_k is the number of documents assigned to topic k, $e_{k,j}$ the number of times word j is assigned to topic k across all documents, and $f_{m,j}$ the number of times word j appears in the m-th document.

$$d_k = \sum_{m=1}^{M} \mathrm{I}(z_m = k) \tag{42}$$

$$e_{k,j} = \sum_{m=1}^{M}\sum_{n=1}^{N_m} \mathrm{I}(z_m = k \ \&\ y_{m,n} = j) \tag{43}$$

$$f_{m,j} = \sum_{n=1}^{N_m} \mathrm{I}(y_{m,n} = j) \tag{44}$$

Note that if we sum over all the words in a document, we get its length,

$$f_{m,*} = \sum_{j=1}^{J} f_{m,j} = N_m \tag{45}$$

2.3 Integrating out Multinomials in NB

The collapsed sampler needs to compute the probabiilty of topic z_a being assigned to document a given z^{-a}, the assignment of topics to all documents other than a.

$$p(z_a | z_{-a}, y, \alpha, \beta) \tag{46}$$

We begin by expanding according to the definition of conditional probability

$$= \frac{p(z_a, z^{-a}, y | \alpha, \beta)}{p(z^{-a}, y | \alpha, \beta)} \tag{47}$$

And then note that the denominator does not depend on z_a,

$$\propto p(z_a, z_{-a}, y | \alpha, \beta) \tag{48}$$

Because z_a and z^{-a} together make up all of z,

$$= p(z, y | \alpha, \beta) \tag{49}$$

Compute this value by integrating out the multinomial parameters θ and φ from the joint probability,

$$= \int \int p(z, y, \theta, \varphi | \alpha, \beta) \; d\theta \; d\varphi \tag{50}$$

Expand the joint probability by its definition (39),

$$= \int \int p(\varphi | \beta) \; p(\theta | \alpha) \; p(z | \theta) \; p(y | z, \varphi) \; d\theta \; d\varphi \tag{51}$$

Separate out the integrals based on bound variables, noting that others are constant,

$$= \int p(z | \theta) \; p(\theta | \alpha) \; d\theta \times \int p(y | z, \varphi) \; p(\varphi | \beta) \; d\varphi \tag{52}$$

Expand out the products implicit in the vector notation, following (40),

$$= \int p(\theta | \alpha) \prod_{m=1}^{M} p(z_m | \theta) \; d\theta \times \int \prod_{k=1}^{K} p(\varphi_k | \beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(y_{m,n} | \varphi_{z_m}) \; d\varphi \tag{53}$$

Then distribute the integral over the word probabilities per topic k,

$$= \int p(\theta|\alpha) \prod_{m=1}^{M} p(z_m|\theta)\ d\theta \times \prod_{k=1}^{K} \int p(\varphi_k|\beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(y_{m,n}|\varphi_{z_m})\ d\varphi_k \tag{54}$$

And then expand each probability formula based on its density,

$$= \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{m=1}^{M} \theta_{z_m}\ d\theta \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j - 1} \prod_{m=1}^{M} \prod_{n=1}^{N_m} \varphi_{z_m, y_{m,n}}\ d\varphi_k \tag{55}$$

Because $x^a x^b = x^{a+b}$, we may replace the product over documents m in the first term by total topic counts d_k, and similarly for the second term's product over documents m and words n with word-topic counts $e_{k,j}$,

$$= \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{d_k}\ d\theta \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j - 1} \prod_{j=1}^{J} \varphi_{k,j}^{e_{k,j}}\ d\varphi_k \tag{56}$$

Based on the same algebra, we push the counts together from the separate products over topics k in the first term, and words j in the second term,

$$= \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k + d_k - 1}\ d\theta\ \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j + e_{k,j} - 1}\ d\varphi_k \tag{57}$$

We then multiply by 1 expressed in a convenient form, and distribute through the integrals,

$$= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + d_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + d_k)} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k + d_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k + d_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k + d_k - 1}\ d\theta \tag{58}$$

$$\times \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j})} \int \frac{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j})}{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j})} \prod_{j=1}^{J} \varphi_{k,j}^{\beta_j + e_{k,j} - 1}\ d\varphi_k$$

Because the remaining integrals are of Dirichlet densities over their complete support, they evaluate to 1, and drop out of products,

$$= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + d_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + d_k)} \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_j)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j})} \tag{59}$$

We then eliminate constant terms that don't depend on the current topic choice z_a,

$$\propto \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + d_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + d_k)} \times \prod_{k=1}^{K} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j})} \tag{60}$$

Pull apart the products based on whether the topic k is the assignment z_a to the current document, noting that counts d and d^{-a} are the same for documents not equal to a, and noting that e and e^{-a} have the same count for topics k other than z_a,

$$= \frac{\prod_{k \neq z_a} \Gamma(\alpha_k + d_k^{-a})}{\Gamma(1 + \sum_{k=1}^{K} \alpha_k + d_k^{-a})} \times \Gamma(\alpha_{z_a} + d_{z_a}^{-a} + 1) \tag{61}$$

$$\times \prod_{k \neq z_a} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j}^{-a})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j}^{-a})} \times \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{z_a,j}^{-a} + f_{a,j})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{z_a,j}^{-a} + f_{a,j})}$$

Then expand out the increment terms, based on the expansion $\Gamma(x+q) = \Gamma(x) \times \prod_{i=1}^{q}(x+i)$; empty products evaluate to 1, so that if $q = 0$, we retain $\Gamma(x+0) = \Gamma(x)$,

$$= \frac{\prod_{k \neq z_a} \Gamma(\alpha_k + d_k^{-a})}{\Gamma(1 + \sum_{k=1}^{K} \alpha_k + d_k^{-a})} \times \Gamma(\alpha_{z_a} + d_{z_a}^{-a}) \times (\alpha_{z_a} + d_{z_a}^{-a}) \tag{62}$$

$$\times \prod_{k \neq z_a} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j}^{-a})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j}^{-a})} \times \frac{\prod_{j=1}^{J} \left(\Gamma(\beta_j + e_{z_a,j}^{-a}) \times \prod_{i=1}^{f_{a,j}} (\beta_j + e_{z_a,j}^{-a} + i) \right)}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{z_a,j}^{-a}) \times \prod_{i=1}^{f_{a,*}} (\sum_{j=1}^{J} \beta_j + e_{z_a,j}^{-a} + i)}$$

We then refold the residual Γ-function terms back into their general products,

$$= \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + d_k^{-a})}{\Gamma(1 + \sum_{k=1}^{K} \alpha_k + d_k^{-a})} \times (\alpha_{z_a} + d_{z_a}^{-a}) \tag{63}$$

$$\times \prod_{k=1}^{K} \frac{\prod_{j=1}^{J} \Gamma(\beta_j + e_{k,j}^{-a})}{\Gamma(\sum_{j=1}^{J} \beta_j + e_{k,j}^{-a})} \times \frac{\prod_{j=1}^{J} \prod_{i=1}^{f_{a,j}} (\beta_j + e_{z_a,j}^{-a} + i)}{\prod_{i=1}^{f_{a,*}} (\sum_{j=1}^{J} \beta_j + e_{z_a,j}^{-a} + i)}$$

And finally remove all the terms that don't depend on z_a, leaving us with

$$\propto (\alpha_{z_a} + d_{z_a}^{-a}) \times \frac{\prod_{j=1}^{J} \prod_{i=1}^{f_{a,j}} (\beta_j + e_{z_a,j}^{-a} + i)}{\prod_{n=1}^{N_a} (\sum_{j=1}^{J} \beta_j + e_{z_a,j}^{-a} + n)} \tag{64}$$

The first term is just the prior α_{z_a} for topic z_a plus the number of documents assigned to topic z_a not counting document a. The right-hand fraction has a natural interpretation in terms of Bayesian updating. The numerator of the fraction may be read procedurally as assigning each word sequentially to a topic, and always counting how many have been applied so far. That is, the third instance of a word is going to be more likely than the first. The denominator, where we've replaced $f_{a,*}$, the count of the number of words in document a, with the equivalent constant N_a, simply provides the normalization for each term, computed incrementally. Note that the model is truly multinomial in that the order of words in a document doesn't matter.

3 Historical Notes

The LDA model was introduced by Blei, Ng and Jordan (2003). The collapsed Gibbs sampler was introduced by Griffiths and Steyvers (2004, 2007). Our derivation was based on the structure of the proof found in the Wikipedia.

There is a similar, though much terser, derivation for the LDA case in (Heinrich 2008) and an even terser sketch exploiting conjugacy of the Dirichlet-multinomial model in (Griffiths 2002).

Acknowledgments

Thanks to my blog commenters for corrections on earlier drafts. Brad (no last name given) supplied a correction for a typo in the LDA derivation and Arwen Twinkle provided corrections for several important corrections in the LDA derivation.

Thanks to Ramnath Balasubramanyan for pointing out (Heinrich 2008) which led me to (Griffiths 2002).

References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**:993–1022.
  http://jmlr.csail.mit.edu/papers/v3/blei03a.html

- Carpenter, Bob. 2010. Collapsed Gibbs sampling for LDA and Bayesian naive Bayes. *LingPipe Blog*.
  http://lingpipe-blog.com/2010/07/13/collapsed-gibbs-sampling-for-lda-bayesian-naive-bayes

- Griffiths, Tom. 2002. Gibbs sampling in the generative model of latent Dirichlet allocation. Unpublished note.
  http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.8022

- Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* **101** (Suppl. 1):5228–5235.
  http://www.pnas.org/content/101/suppl.1/5228

- Heinrich, Gregor. 2008. Parameter estimation for text analysis. Version 2.4. Unpublished note.
  http://www.arbylon.net/publications/text-est.pdf

- Steyvers, Mark and Tom Griffiths. 2007. Probabilistic topic models. In Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch (eds.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.
  http://cocosci.berkeley.edu/tom/papers/SteyversGriffiths.pdf

- Wikipedia. *Latent Dirichlet Allocation*. Downloaded 13 July 2010.
  http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation