EECS E6720 Bayesian Models for Machine Learning

Columbia University, Fall 2016

Lecture 7, 10/27/2016

Instructor: John Paisley

- Let's look at another example of a standard model that is easily learned with variational inference.

  Latent Dirichlet allocation (LDA)

- LDA is a Bayesian approach to topic modeling and one of the fundamental models in machine learning. It is popular because it has a wide range of applications and is easy to develop and build into larger systems.

  Setup: We have discrete grouped data, $x_d = \{x_{d1}, \ldots, x_{d,N_d}\}$, where d indexes group number. Each $x_{di} \in \{1, \ldots, V\}$, meaning each observation takes one of V values and x_d is a particular collection of values disjoint from all other $x_{d'}$.

  Example: The classic example where LDA is used is in document modeling. We'll assume we're working in that context from now on.

  - In this case, d would index a particular document (e.g., a blog post, an article in a newspaper, etc.) and V would be the size of the vocabulary of words.
  - The value x_{di} is the index of the ith word in the dth document. For example, $x_{di} = 7241$ might mean that the ith word in document d is "government," meaning that the number 7241 maps to the word "government" in a vocabulary that we have constructed.
  - When we say the "ith word in the dth document," this doesn't have to refer to the literal order. LDA doesn't model word order, so any re-ordering (and thus re-indexing) of the words will be viewed in exactly the same way by LDA.
  - Also, as an aside, there is significant pre-processing prior to getting each x_d. The vocabulary size V and the words comprising that vocabulary is selected in advance, with overly common words like "the" removed and words that are very rare also removed. This pre-processing will have an impact on performance, but we will assume that this task has already been done.

Topic modeling: The idea behind topic modeling is to model the D documents $x_1, \ldots, x_D$ as being generated from a set of K "topics," $\beta_1, \ldots, \beta_K$. Every document shares these topics, but has its own document-specific model variable that dictates how they are used.

- β_k : a V-dimensional probability distribution on the V words for topic k
- θ_d : a K-dimensional probability distribution on the topics for document d

Model: Given the topics, $\beta_1, \ldots, \beta_K$ and the distribution θ_d on them for document d, generate all data in document d as follows:

$$c_{di} \sim \text{Discrete}(\theta_d), \quad x_{di} \sim \text{Discrete}(\beta_{c_{di}}). \tag{1}$$

Notice that we have introduced an additional latent variable to the model:

- c_{di} picks out the topic that the ith word in document d belongs to. Notice that if $x_{di} = x_{di'}$ (i.e., the same word appears multiple times in a document) it's not necessarily the case that $c_{di} = c_{di'}$. The same word can have significant probability in multiple topics.
- x_{di} is generated using the topic indexed by c_{di}, and hence the subscript on β.

Priors: We don't know θ_d or β_k (or c_{di} for that matter, but we know it's distribution given θ_d). Therefore, we need to put prior distributions on them. There are many distributions we could pick. LDA uses the following:

$$\theta_d \overset{iid}{\sim} \text{Dirichlet}(\alpha), \quad \beta_k \overset{iid}{\sim} \text{Dirichlet}(\gamma) \tag{2}$$

Since θ_d and β_k are all finite probability vectors used in a discrete (or from another perspective, multinomial) distribution, LDA uses a conditionally conjugate Dirichlet prior for each of them. This makes variational inference very easy.

A prior discussion on the posterior: Before we get into the variational inference algorithm for LDA, what do we hope to find? What will β_k and θ_d tell us that's useful?

- β_k : If we look at the approximate posterior distribution of β_k and see which dimensions of it are expected to be large, then the high probability words should all relate to a coherent theme. For example, the three most probable words might be "government," "politics," and "congress." We would then call this a "politics" topic. In papers on topic modeling, you will often see lists of words. This is exactly what is being done: Each list is showing the 5 or 10 most probable words (dimensions) according to a specific topic (β_k).
- θ_d : This will tell us the fraction of each topic appearing in a particular document. Because we can assign meaning to each β_k after the fact, we can then say, e.g., "this document is 75% politics and 25% technology," etc., because $\theta_{d,1} = 0.75$ and β_1 is the "politics" topic.
- In general, both β_k and θ_d will be highly sparse, meaning only a fraction of values will be significantly nonzero.

Posterior calculation: We use Bayes rule to try to calculate the posterior,

$$\begin{aligned} p(\beta, \theta, c|x) &\propto p(x|\beta, \theta, c)p(\beta, \theta, c) \\ &\propto p(x|\beta, \theta, c)p(c|\beta, \theta)p(\beta, \theta) \\ &\propto p(x|\beta, c)p(c|\theta)p(\beta)p(\theta) \end{aligned} \tag{3}$$

The first two lines are simply true statements about Bayes rule and how probabilities factorize. The last line takes into consideration the dependency structure of LDA to remove unnecessary conditioning. By the assumed independence structure, this further breaks down as follows:

$$p(\beta, \theta, c|x) \propto \Big[\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(x_{di}|\beta, c_{di})p(c_{di}|\theta_d)\Big]\Big[\prod_{k=1}^{K} p(\beta_k)\Big]\Big[\prod_{d=1}^{D} p(\theta_d)\Big] \tag{4}$$

- Not surprisingly, this can't be normalized and so we need an inference algorithm to approximate the posterior. Notice that there are quite a few variables we want to learn with this model. For example, for each word x_{di} in the data set, there is an associated topic indicator c_{di} that we need to learn. Therefore the number of variables is much larger than what we've discussed before.

Variational inference for LDA

- We will use our previous discussion of the "optimal method" for variational inference to approximate the posterior of the LDA model.

- Step 1: Using the mean-field assumption, we need to pick a factorization of $q(\beta, \theta, c) \approx p(\beta, \theta, c|x)$. We split these variables according to how they are generated in the prior. This makes learning q much easier.

$$q(\beta, \theta, c) = \Big[\prod_{k=1}^{K} q(\beta_k)\Big]\Big[\prod_{d=1}^{D} q(\theta_d)\Big]\Big[\prod_{d=1}^{D}\prod_{i=1}^{N_d} q(c_{di})\Big] \tag{5}$$

  Step 2: Next we need to select the distribution family for each q. For this model we will be able to find the optimal distributions. Remember from our previous discussion that, for a given variable, we can find the optimal q distribution as follows:

  1. Take the log of the complete joint likelihood
  2. Take the expectation of this using all other q distributions except the one of interest
  3. Exponentiate the result and normalize over the variable of interest

- The potential concern with this was that we don't know any of the q distributions to begin with, so when we take the expectation with respect to "all other" q, we don't know what these expectations are! However, recall that we don't need to know the actual values of the expectations in order to find the family of the q distribution being considered. Therefore, by completing one loop of this procedure we reach the point where we can go back and calculate the expectations that we left undefined previously.

- However, before we can do this, we need to know how to write the joint likelihood for LDA.

Joint likelihood of LDA

- The posterior $p(\beta, \theta, c|x) \propto p(x, \beta, \theta, c)$, which is what we calculated before,

$$p(x, \beta, \theta, c) = \Big[\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(x_{di}|\beta, c_{di})p(c_{di}|\theta_d)\Big]\Big[\prod_{k=1}^{K} p(\beta_k)\Big]\Big[\prod_{d=1}^{D} p(\theta_d)\Big] \tag{6}$$

- These probabilities are all easy to write, but we run into a problem with the term

$$p(x_{di}|\beta, c_{di}) = \prod_{v=1}^{V} \beta_{c_{di}}(v)^{\mathbb{1}(x_{di}=v)} \tag{7}$$

  Notice that this function picks out the correct dimension of β according to the value that x_{di} takes. However, it's hard to do variational inference for c_{di} as written.

- This is another example where notation can help make things easier to derive. Notice that

$$p(x_{di}|\beta, c_{di}) = \prod_{v=1}^{V} \beta_{c_{di}}(v)^{\mathbb{1}(x_{di}=v)} = \prod_{k=1}^{K}\Big[\prod_{v=1}^{V} \beta_k(v)^{\mathbb{1}(x_{di}=v)}\Big]^{\mathbb{1}(c_{di}=k)} \tag{8}$$

  In both cases, c_{di} picks out the correct topic vector β_k, while x_{di} picks out the correct dimension of the selected vector.

- To keep the notation clean, we write

$$p(x, \beta, \theta, c) = \Big[\prod_{d=1}^{D}\prod_{i=1}^{N_d}\prod_{k=1}^{K} \big(p(x_{di}|\beta_k)\theta_{dk}\big)^{\mathbb{1}(c_{di}=k)}\Big]\Big[\prod_{k=1}^{K} p(\beta_k)\Big]\Big[\prod_{d=1}^{D} p(\theta_d)\Big] \tag{9}$$

  Notice that we directly use $\theta_{dk} = p(c_{di} = k|\theta_d)$, but leave the rest in $p(\cdot)$ notation.

- Therefore, the log of the joint likelihood can be written as

$$\begin{aligned}\ln p(x, \beta, \theta, c) \;\; = \;\; & \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} \mathbb{1}(c_{di} = k) \ln p(x_{di}|\beta_k) + \mathbb{1}(c_{di} = k) \ln \theta_{dk} \\ & + \sum_{k=1}^{K} \ln p(\beta_k) + \sum_{d=1}^{D} \ln p(\theta_d)\end{aligned} \tag{10}$$

- The log joint likelihood is a key equation and should never be out of sight when deriving variational inference algorithms.

- As a general rule, when we want to find a q distribution, we can throw away anything not involving the variable being updated. This will make working with the log joint likelihood less intimidating and require much less writing.

- To emphasize this point, let's look at a made-up toy example.

  Toy example

- Imagine a model with joint likelihood $p(x, c, a, b)$ where x is the data and c, a, b are model variables. Then using the factorization

$$q(c, a, b) = q(c)q(a)q(b)$$

  we have that

$$q(c) \propto \mathrm{e}^{\mathbb{E}_{-q}[\ln p(x,c,a,b)]}$$

  We use the shorthand notation "$-q$" to indicate all q distributions except for the one being considered at the moment.

- What if $\ln p(x, c, a, b) = ax + bc + xc^2 + ab$? This is a completely made up log joint likelihood and doesn't actually correspond to anything. However, this is just to highlight the point. In this case,

$$\begin{aligned} q(c) &\propto \mathrm{e}^{\mathbb{E}[a]x+\mathbb{E}[b]c+xc^2+\mathbb{E}[a]\mathbb{E}[b]} \\ &\propto \mathrm{e}^{\mathbb{E}[a]x+\mathbb{E}[a]\mathbb{E}[b]}\mathrm{e}^{\mathbb{E}[b]c+xc^2} \\ &\propto \mathrm{e}^{\mathbb{E}[b]c+xc^2} \end{aligned}$$

- This last line is because

$$q(c) = \frac{\mathrm{e}^{\mathbb{E}[a]x+\mathbb{E}[a]\mathbb{E}[b]}\mathrm{e}^{\mathbb{E}[b]c+xc^2}}{\int \mathrm{e}^{\mathbb{E}[a]x+\mathbb{E}[a]\mathbb{E}[b]}\mathrm{e}^{\mathbb{E}[b]c+xc^2}dc} = \frac{\mathrm{e}^{\mathbb{E}[b]c+xc^2}}{\int \mathrm{e}^{\mathbb{E}[b]c+xc^2}dc}$$

- As a side comment, we can write $\mathbb{E}[ab] = \mathbb{E}[a]\mathbb{E}[b]$ because the expectation uses $q(a, b) = q(a)q(b)$, so they're independent.

- The take-home message here is that, when updating a q distribution for a particular model variable, we only need to look at the terms in the log joint likelihood that involve this variable.

- The log joint likelihood will be the sum of many things, and anything not involving this variable can be absorbed in the normalizing constant and so ignored. For this model, that means we can simply ignore $ax + ab$ in the log joint likelihood when we find $q(c)$.

- This is just a toy example to drive home a point. But for complicated models, this way of simplifying things can make deriving the VI algorithm seem like a much less daunting task.

- For quick reference, in deriving the VI algorithm for LDA, we doing this with the joint likelihood

$$\begin{aligned} \ln p(x, \beta, \theta, c) &= \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} \mathbb{1}(c_{di} = k) \ln p(x_{di}|\beta_k) + \mathbb{1}(c_{di} = k) \ln \theta_{dk} \\ &\quad + \sum_{k=1}^{K} \ln p(\beta_k) + \sum_{d=1}^{D} \ln p(\theta_d) \end{aligned} \tag{11}$$

$q(c_{di})$: Indicator of which topic word x_{di} came from

- To find this q distribution, we can focus only on terms in the log joint likelihood involving c_{di}. Therefore,

$$\begin{aligned} q(c_{di}) &\propto \mathrm{e}^{\sum_{k=1}^{K} \mathbb{1}(c_{di}=k)\left(\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]\right)} \\ &\propto \prod_{k=1}^{K}\left[\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]}\right]^{\mathbb{1}(c_{di}=k)} \end{aligned} \tag{12}$$

- We want to normalize this over c_{di}. Since $c_{di} \in \{1, \ldots, K\}$, the integral in the denominator turns into a sum,

$$q(c_{di}) = \frac{\prod_{k=1}^{K}\left[\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]}\right]^{\mathbb{1}(c_{di}=k)}}{\sum_{c_{di}=1}^{K}\prod_{k=1}^{K}\left[\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]}\right]^{\mathbb{1}(c_{di}=k)}} \tag{13}$$

- This is another way of writing

$$q(c_{di}) = \prod_{k=1}^{K}\left[\frac{\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]}}{\sum_{j=1}^{K}\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_j)]+\mathbb{E}_{-q}[\ln \theta_{d,j}]}}\right]^{\mathbb{1}(c_{di}=k)} \tag{14}$$

- Notice that this is simply a discrete distribution,

$$q(c_{di}) = \text{Discrete}(\phi_{di}), \quad \phi_{di}(k) = \frac{\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]+\mathbb{E}_{-q}[\ln \theta_{dk}]}}{\sum_{j=1}^{K}\mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_j)]+\mathbb{E}_{-q}[\ln \theta_{d,j}]}} \tag{15}$$

- Compare this with Gibbs sampling. Remember that we sample from the conditional posterior distribution,

$$p(c_{di} = k|\beta, \theta_d) = \frac{p(x_{di}|\beta_k)\theta_{dk}}{\sum_{j=1}^{K} p(x_{di}|\beta_j)\theta_{d,j}}$$

  where β and θ_d are the most recent samples of these variables.

- For variational inference, we swap

$$p(x_{di}|\beta_k) \Rightarrow \mathrm{e}^{\mathbb{E}_{-q}[\ln p(x_{di}|\beta_k)]} \quad \text{and} \quad \theta_{dk} \Rightarrow \mathrm{e}^{\mathbb{E}_{-q}[\ln \theta_{dk}]}$$

  Rather than sampling, we then simply keep this "approximate conditional posterior" as the q distribution for c_{di}.

- Notice that while $a = \mathrm{e}^{\ln a}$, using expectations, $\mathbb{E}[a] \neq \mathrm{e}^{\mathbb{E}[\ln a]}$.

- We don't know what $\mathbb{E}[\ln \theta_{dk}]$ and $\mathbb{E}[\ln p(x_{di}|\beta_k)] = \mathbb{E}[\ln \beta_{k,x_{di}}]$ are yet, but that doesn't change what the optimal form of $q(c_{di})$ is. Notice that, if we can use the same logic to find $q(\beta_k)$ and $q(\theta_d)$, then after one cycle through the model variables we will be able to come back to $q(c_{di})$ and explicitly calculate these expectations.

- Also notice that, since d and i are arbitrary, we have solved $q(c_{di})$ for all d and i.

$q(\theta_d)$: The distribution on topics for document d

- To find $q(\theta_d)$, we focus only on terms in the log joint likelihood involving this variable

$$\begin{aligned} q(\theta_d) &\propto \mathrm{e}^{\sum_{i,k} \mathbb{E}_{-q}[\mathbb{1}(c_{di}=k)] \ln \theta_{dk} + \ln p(\theta_d)} \\ &\propto \prod_{k=1}^{K} \theta_{dk}^{\sum_{i=1}^{N_d} \mathbb{E}_{-q}[\mathbb{1}(c_{di}=k)] + \alpha - 1} \end{aligned} \tag{16}$$

- The term $\alpha - 1$ comes from the Dirichlet prior $p(\theta_d)$. We want to normalize this over θ_d subject to $\theta_{dk} \geq 0$ and $\sum_k \theta_{dk} = 1$. This was a problem on the first homework, where we saw that the solution is

$$q(\theta_d) = \text{Dirichlet}(\alpha_{d1}, \ldots, \alpha_{dK}), \quad \alpha_{dk} = \alpha + \sum_{i=1}^{N_d} \underbrace{\mathbb{E}_{-q}[\mathbb{1}(c_{di} = k)]}_{\phi_{di}(k)} \tag{17}$$

- The expectation of an indicator of an event is simply the probability of that event. Therefore,

$$\mathbb{E}_{-q}[\mathbb{1}(c_{di} = k)] \;=\; \sum_{j=1}^{K} q(c_{di} = j)\mathbb{1}(c_{di} = k) \;=\; q(c_{di} = k) \tag{18}$$

  Since we previously calculated this q distribution, we are able to solve this expectation and use $\phi_{di}(k)$ as defined above for $q(c_{di})$.

- Notice that the parameters of the Dirichlet use the expected histogram of the allocations of all words from document d. Therefore, if 75% of words are expected to come from topic 1, then $q(\theta_d)$ will reflect this by having $\mathbb{E}_q[\theta_{d,1}] \approx 0.75$. However, q will also capture an approximation of the uncertainty of this value under its posterior distribution.

- Again compare with Gibbs sampling, where we have the conditional posterior

$$p(\theta_d | c_d) \propto \underbrace{\Big[\prod_{i=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{\mathbb{1}(c_{di}=k)}\Big]}_{p(c_d|\theta_d)} \underbrace{\Big[\prod_{k=1}^{K} \theta_{dk}^{\alpha-1}\Big]}_{p(\theta_d)} \tag{19}$$

- In this case,

$$p(\theta_d | c_d) = \text{Dirichlet}(\alpha_{d1}, \ldots, \alpha_{dK}), \quad \alpha_{dk} = \alpha + \sum_{i=1}^{N_d} \mathbb{1}(c_{di} = k)$$

  where c_{di} is the most recent sample of this variable. In this case, we use the *empirical* histogram (rather than the expected histogram) constructed from the most recent samples. Gibbs sampling then samples a new vector θ_d from this distribution, while variational inference keeps the approximate conditional posterior as the approximation to the full posterior of this variable.

- Again, because d is arbitrary, we've solved for all θ_d.

$q(\beta_k)$: The topics

- Finally, we learn the q distributions for the topics themselves. Following the same procedure as for $q(\theta_d)$ and $q(c_{di})$, we have

$$\begin{aligned} q(\beta_k) &\propto \mathrm{e}^{\sum_{d,i} \mathbb{E}_{-q}[\mathbb{1}(c_{di}=kk)] \ln p(x_{di}|\beta_k) + \ln p(\beta_k)} \\ &\propto p(\beta_k) \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(x_{di}|\beta_k)^{\mathbb{E}_{-q}[\mathbb{1}(c_{di}=k)]} \end{aligned} \tag{20}$$

- Since $p(x_{di}|\beta_k) = \prod_{v=1}^{V} \beta_{kv}^{\mathbb{1}(x_{di}=v)}$,

$$q(\beta_k) \propto \prod_{v=1}^{V} \beta_{kv}^{\sum_{d,i} \mathbb{E}_{-q}[\mathbb{1}(c_{di}=k)]\mathbb{1}(x_{di}=v)+\gamma-1} \tag{21}$$

- Normalizing over the probability vector β_k,

$$q(\beta_k) = \text{Dirichlet}(\gamma_{k,1}, \ldots, \gamma_{k,V}), \quad \gamma_{k,v} = \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{E}_{-q}[\mathbb{1}(c_{di} = k)]\mathbb{1}(x_{di} = v) + \gamma \tag{22}$$

- Once again, we set $\mathbb{E}_{-q}[\mathbb{1}(c_{di} = k)] = \phi_{di}(k)$ as defined in the update of $q(c_{di})$.

- We can interpret $\sum_{d=1}^{D} \sum_{i=1}^{N_d} \phi_{di}(k)\mathbb{1}(x_{di} = v)$ as follows:

  1. $\phi_{di}(k)$: The probability that word x_{di} came from topic k according to $q(c_{di})$
  2. $\mathbb{1}(x_{di} = v)$: An indicator of what the ith word in document d corresponds to

- So if $v \rightarrow$ "government" then this sum is the expected total number of times we see the word "government" come from topic k given the model's q distributions at the current iteration.

- Because we solved for $q(c_{di})$ first, and the updates of $q(\theta_d)$ and $q(\beta_k)$ only used $q(c_{di})$, we were able to input the correct variational parameters to update these last two distributions. Finally, we need to address how to solve the expectations in $q(c_{di})$ in order to update this distribution, and therefore explicitly say what each ϕ_{di} equals in the updates of $q(\theta_d)$ and $q(\beta_k)$.

- That is, we have the question, how do we actually calculate

$$\mathbb{E}_{-q}[\ln \theta_{dk}] \quad \text{and} \quad \mathbb{E}_{-q}[\ln \beta_{kv}] \; ?$$

- The quick answer is, since we know these expectations are both with respect to Dirichlet distributions on θ and β, we can go to Wikipedia and look up the solution. However, we can also derive this using a clever technique.

- To do so, we will find that understanding properties of exponential family distributions gives a general way to find many interesting expectations.

Exponential family distributions

- We take a quick detour to discuss exponential family distributions. Almost all distributions we have (and will) discuss are in the "exponential family," for example, Gaussian, beta, Dirichlet, gamma, Poisson, multinomial, etc. distributions.

- This means they can be written in the form

$$p(x|\eta) = h(x)\mathrm{e}^{\eta^T t(x) - A(\eta)} \tag{23}$$

  where these terms are called as follows:

  1. η is the natural parameter vector
  2. $t(x)$ is the sufficient statistic vector
  3. $h(x)$ is the base measure (a function of x)
  4. $A(\eta)$ is the log-normalizer (a function of η)

- Q: Is there a general way to find $\mathbb{E}[t(x)]$?

- A: Yes, the derivation follows: Since

$$\int p(x|\eta)dx = 1 \quad \Rightarrow \quad \nabla_\eta \int p(x|\eta)dx = 0 \tag{24}$$

  we have that

$$\begin{aligned}
\nabla_\eta \int p(x|\eta)dx &= 0 && (25)\\
&\Downarrow \\
\int (t(x) - \nabla_\eta A(\eta))p(x|\eta)dx &= 0 && (26)\\
&\Downarrow \\
\int t(x)p(x|\eta)dx &= \int \nabla_\eta A(\eta)p(x|\eta)dx && (27)\\
&\Downarrow \\
\mathbb{E}[t(x)] &= \nabla_\eta A(\eta) && (28)
\end{aligned}$$

- We will find that, in many cases, the expectations we want to take in variational inference are of a sufficient statistic. We show this for the Dirichlet example.

- Dirichlet example: The density of the Dirichlet distribution can be put into the exponential family form as follows

$$\begin{aligned}
p(x|\alpha_1, \ldots, \alpha_k) &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1} && (29)\\
&= \Big(\prod_{i=1}^k x_i^{-1}\Big) \mathrm{e}^{\sum_{i=1}^k \alpha_i \ln x_i - \left(\sum_i \ln \Gamma(\alpha_i) - \ln \Gamma(\sum_i \alpha_i)\right)} && (30)
\end{aligned}$$

- Matching terms with the generic exponential family distribution form, we have
  - $h(x) = \prod_i x_i^{-1}$
  - $t(x) = [\ln x_1, \ldots, \ln x_k]^T$
  - $\eta = [\alpha_1, \ldots, \alpha_k]^T$
  - $A(\eta) = \sum_i \ln \Gamma(\alpha_i) - \ln \Gamma(\sum_i \alpha_i)$

  Therefore, $\mathbb{E}[t(x)] = \nabla_\eta A(\eta)$ implies that, for each i,

$$\mathbb{E}[\ln x_i] = \partial A / \partial \alpha_i \quad \rightarrow \quad \mathbb{E}[\ln x_i] = \frac{\partial \ln \Gamma(\alpha_i)}{\partial \alpha_i} - \frac{\partial \ln \Gamma(\sum_j \alpha_j)}{\partial \alpha_i} \tag{31}$$

  These derivatives appear often enough that they have been given a symbol, $\psi(\cdot)$, called a digamma function. Therefore, $\mathbb{E}[\ln x_i] = \psi(\alpha_i) - \psi(\sum_j \alpha_j)$. This can be evaluated in languages such as Matlab using a built-in function.

- The final variational inference algorithm for LDA is given below.

Variational inference for latent Dirichlet allocation (LDA)

1. Define $q(c_{di}) = \text{Discrete}(\phi_{di})$, $q(\theta_d) = \text{Dirichlet}(\alpha_d)$ and $q(\beta_k) = \text{Dirichlet}(\gamma_k)$. Initialize each $\phi_{di}^{(0)}$, $\alpha_d^{(0)}$ and $\gamma_k^{(0)}$ in some way.
2. for iteration $t = 1, \ldots, T$
   (a) For each d and i, set

$$\phi_{di}^{(t)}(k) = \frac{\mathrm{e}^{\psi(\gamma_{k,x_{di}}^{(t-1)}) - \psi(\sum_v \gamma_{k,v}^{(t-1)}) + \psi(\alpha_{d,k}^{(t-1)}) - \psi(\sum_j \alpha_{d,j}^{(t-1)})}}{\sum_{m=1}^K \mathrm{e}^{\psi(\gamma_{m,x_{di}}^{(t-1)}) - \psi(\sum_v \gamma_{m,v}^{(t-1)}) + \psi(\alpha_{d,m}^{(t-1)}) - \psi(\sum_j \alpha_{d,j}^{(t-1)})}}$$

   (b) For each d and $k = 1, \ldots, K$, set

$$\alpha_{dk}^{(t)} = \alpha + \sum_{i=1}^{N_d} \phi_{di}^{(t)}(k)$$

   (c) For each k and $v = 1, \ldots, V$, set

$$\gamma_{kv}^{(t)} = \gamma + \sum_{d=1}^{D} \sum_{i=1}^{N_d} \phi_{di}^{(t)}(k) \mathbb{1}(x_{di} = v)$$

3. Using all variational parameter updates after iteration t, evaluate

$$\mathcal{L}_t = \mathbb{E}_q[\ln p(x, c, \beta, \theta)] - \sum_{d,i} \mathbb{E}_q[\ln q(c_{di})] - \sum_d \mathbb{E}_q[\ln q(\theta_d)] - \sum_k \mathbb{E}_q[\ln q(\beta_k)]$$

   to assess convergence. This function must be monotonically increasing as a function of t.