Day 2: Lecture 2

John Paisley

Columbia University

Duke-Tsinghua Machine Learning Summer School

July 26, 2017

Part I

Learning Latent Dirichlet Allocation

We've defined the LDA model, now how do we learn the variables in it?

Latent Dirichlet Allocation

1. Generate K topics: $\beta_k \sim \text{Dirichlet}(\gamma)$
2. Generate document distributions: $\theta_d \sim \text{Dirichlet}(\alpha)$
3. For the nth word in the dth document,
    a) Assign to topic, $c_{dn} \sim \text{Discrete}(\theta_d)$
    b) Generate observation, $x_{dn} \sim \text{Discrete}(\beta_{c_{dn}})$

We don't know $\beta_1, \ldots, \beta_K$, $\theta_1, \ldots, \theta_D$, $c_{1,2}$, $c_{4,7}$, $c_{857,12}$, ...

Since LDA is a Bayesian model, start with Bayes Rule. Define

- $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$
- $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_D\}$
- $\boldsymbol{c} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_D\}$ and $\boldsymbol{c}_d = \{c_{d,1}, \ldots, c_{d,n}\}$
- $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D\}$ and $\boldsymbol{x}_d = \{x_{d,1}, \ldots, x_{d,n}\}$

We want the posterior distribution of these variables,

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})}{p(\boldsymbol{x})}$$

However, the denominator is not something we can solve.

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{c}} \int p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})d\boldsymbol{\beta}d\boldsymbol{\theta} = \textbf{?}$$

Variational inference: We will see that all we need is the joint likelihood.

For LDA, conditional independence lets us write this many ways

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) &= p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) \\
&= p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) p(\boldsymbol{c}|\boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \boldsymbol{\theta}) \\
&= p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{c}) p(\boldsymbol{c}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}) \\
&= p(\boldsymbol{\beta}) \prod_{d=1}^{D} p(\boldsymbol{x}_d|\boldsymbol{\beta}, \boldsymbol{c}_d) p(\boldsymbol{c}_d|\theta_d) p(\theta_d) \\
&= \Big[\prod_k p(\beta_k)\Big] \Big[\prod_d p(\theta_d) \prod_n p(x_{dn}|\boldsymbol{\beta}, c_{dn}) p(c_{dn}|\theta_d)\Big]
\end{aligned}
$$

We can move back and forth between these as we find it convenient.

LDA joint likelihood

$$p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) = \Big[\prod_k p(\beta_k)\Big]\Big[\prod_d p(\theta_d)\prod_n p(x_{dn}|\boldsymbol{\beta}, c_{dn})p(c_{dn}|\theta_d)\Big]$$

By the definition of LDA, we know what distributions to use in the joint likelihood. We list them below for reference.

- $p(\beta_k)$: Dirichlet$(\beta_k|\gamma)$
  $p(\theta_d)$: Dirichlet$(\theta_d|\alpha)$

- $p(c_{dn}|\theta_d) = \theta_{d,c_{dn}} \Longrightarrow \prod_{k=1}^{K}(\theta_{dk})^{\mathbb{1}(c_{dn}=k)}$

- $p(x_{dn}|\boldsymbol{\beta}, c_{dn}) = \beta_{c_{dn},x_{dn}} \Longrightarrow \prod_{k=1}\prod_{v=1}(\beta_{k,v})^{\mathbb{1}(x_{dn}=v)\mathbb{1}(c_{dn}=k)}$

Using indicators is a clever trick that makes the derivation easier.

We know how to write out the joint likelihood distribution $p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})$, but we want the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\, \boldsymbol{x})$.

Using variational inference, we can approximate the posterior by

1. defining a distribution q with which to approximate the posterior

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) \;\approx\; p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\, \boldsymbol{x})$$

2. computing a function of the joint likelihood

$$\mathcal{L} \;=\; \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})] - \mathbb{E}_q[\ln q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]$$

3. modifing the parameters of q to increase $\mathcal{L}$ as much as possible

Steps 2 & 3 minimize the KL-divergence between q and the posterior.

First let's focus on choosing $q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})$.

Requirements, some obvious and some for convenience, are:

- It should be defined on all variables $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}$
- It should be parametric, in that it's a function with parameters
- It should be easy to optimize those parameters

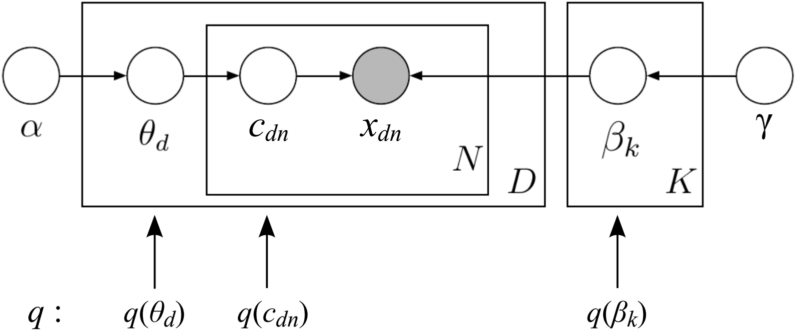
Variational inference achieves this with the **mean-field** assumption.

$$\text{Example:} \quad q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) = q(\boldsymbol{\beta})q(\boldsymbol{\theta})q(\boldsymbol{c})$$

We say q has been "factorized." The question is *what* factorization to use, and then *which* distributions to use in the chosen factorization.

Rule of thumb: Use the factorization that arises in the generative model.

Warning: This is not always the best choice, but it works for LDA.



By factorizing q at this level, we are saying

$$
\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\boldsymbol{x}) &\approx q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) \\
&\approx \Big[\prod_k q(\beta_k)\Big]\Big[\prod_d q(\theta_d)\Big]\Big[\prod_d \prod_n q(c_{dn})\Big]
\end{aligned}
$$

We are approximating the variables to be independent *in the posterior*.

Now we just need to define the distribution for each, and then calculate

$$\mathcal{L} = \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})] - \mathbb{E}_q[\ln q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]$$

By factorizing q at this level, we are saying

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\, \boldsymbol{x}) &\approx q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) \\ &\approx \Big[\prod_k q(\beta_k)\Big]\Big[\prod_d q(\theta_d)\Big]\Big[\prod_d \prod_n q(c_{dn})\Big] \end{aligned}$$

We are approximating the variables to be independent *in the posterior*.

Now we just need to define the distribution for each, and then calculate

$$\mathcal{L} = \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})] - \mathbb{E}_q[\ln q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]$$

By factorizing q at this level, we are saying

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c} \,|\, \boldsymbol{x}) &\approx q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) \\ &\approx \Big[\prod_k q(\beta_k)\Big]\Big[\prod_d q(\theta_d)\Big]\Big[\prod_d \prod_n q(c_{dn})\Big] \end{aligned}$$

We are approximating the variables to be independent *in the posterior*.

Now we just need to define the distribution for each, and then calculate

$$\mathcal{L} = \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})] - \mathbb{E}_q[\ln q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]$$

or do we?

LDA belongs to the class of **conjugate exponential family** models.

- All distributions are in the exponential family
- All model variables are *conditionally* conjugate

Pick any unknown variable in the model and pretend that you know all other variables. The posterior distribution is the same family as the prior.

LDA example: $$\underbrace{p(\theta_d|\boldsymbol{c}_d)}_{\text{Dirichlet posterior}} \propto \underbrace{p(\boldsymbol{c}_d|\theta_d)}_{\text{Discrete likelihood}}\underbrace{p(\theta_d)}_{\text{Dirichlet prior}}$$

We *could* define each q, calculate $\mathcal{L}$, then optimize the parameters of q.

However, because LDA is a conjugate exponential family model, we don't have to. Instead, we can use techniques from Bishop (2006) to find each q.

1. Take the complete q distribution

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c}) = \left[\, \prod_k q(\beta_k)\right] \left[\, \prod_d q(\theta_d)\right] \left[\, \prod_d \prod_n q(c_{dn})\right]$$

   and define $-q_u$ to be q *without* including variable "u."

2. Then *given all other q distributions*, the optimal $q(u)$ is

$$q(u) \propto \exp\{\mathbb{E}_{-q_u}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\}$$

   where "u" is replaced with a variable in the LDA model.

Christopher Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

For LDA, we have to find $q(\beta_k)$, $q(\theta_d)$ and $q(c_{dn})$. Notice that this covers every variable in the model because we derive for all index values.

Let's start with $q(c_{dn})$:

$$q(c_{dn}) \propto \exp\{\mathbb{E}_{-q_{c_{dn}}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\}$$

On an earlier slide we saw how to actually write out $p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})$.

It's a complicated likelihood, but the $\propto$ symbol lets us remove anything that doesn't involve c_{dn} in the exponent (since it cancels when normalizing).

$$q(c_{dn}) \propto \exp\left\{\sum_{k=1}^{K} \mathbb{1}(c_{dn} = k)\big(\mathbb{E}_q[\ln \beta_{k,x_{dn}}] + \mathbb{E}_q[\ln \theta_{dk}]\big)\right\}$$

Before discussing each q in more depth, what about $q(\beta_k)$ and $q(\theta_d)$?

Finding $q(\beta_k)$:

$$
\begin{aligned}
q(\beta_k) &\propto \exp\{\mathbb{E}_{-q_{\beta_k}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\
&\propto \exp\Big\{\sum_v \Big[\gamma - 1 + \sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\mathbb{1}(x_{dn} = v)\Big] \ln \beta_{kv}\Big\}
\end{aligned}
$$

Finding $q(\theta_d)$:

$$
\begin{aligned}
q(\theta_d) &\propto \exp\{\mathbb{E}_{-q_{\theta_d}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\
&\propto \exp\Big\{\sum_k \Big[\alpha - 1 + \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\Big] \ln \theta_{dk}\Big\}
\end{aligned}
$$

Before discussing each q in more depth, what about $q(\beta_k)$ and $q(\theta_d)$?

Finding $q(\beta_k)$:

$$\begin{aligned} q(\beta_k) &\propto \exp\{\mathbb{E}_{-q_{\beta_k}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\ &\propto \exp\Big\{\sum_v \Big[\gamma - 1 + \sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\mathbb{1}(x_{dn} = v)\Big] \ln \beta_{kv}\Big\} \end{aligned}$$

Finding $q(\theta_d)$:

$$\begin{aligned} q(\theta_d) &\propto \exp\{\mathbb{E}_{-q_{\theta_d}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\ &\propto \exp\Big\{\sum_k \Big[\alpha - 1 + \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\Big] \ln \theta_{dk}\Big\} \end{aligned}$$

Before discussing each q in more depth, what about $q(\beta_k)$ and $q(\theta_d)$?

Finding $q(\beta_k)$:

$$\begin{aligned} q(\beta_k) &\propto \exp\{\mathbb{E}_{-q_{\beta_k}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\ &\propto \exp\Big\{ \sum_v \Big[\gamma - 1 + \sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\mathbb{1}(x_{dn} = v)\Big] \ln \beta_{kv} \Big\} \end{aligned}$$

Finding $q(\theta_d)$:

$$\begin{aligned} q(\theta_d) &\propto \exp\{\mathbb{E}_{-q_{\theta_d}}[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]\} \\ &\propto \exp\Big\{ \sum_k \Big[\alpha - 1 + \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn} = k)]\Big] \ln \theta_{dk} \Big\} \end{aligned}$$

Thus far: We've derived each q distribution

1. as a function of the variable of interest
2. as an unnormalized function that needs to be normalized
3. as a function of expectations involving other q distributions

We need to resolve $\#2$ and $\#3$ before we are finished with the variational inference algorithm for LDA.

The plan is to first complete $\#2$, and then use this to answer $\#3$.

Continuing the derivation from a previous slide:

$$
\begin{aligned}
q(c_{dn}) &\propto \prod_{k=1}^{K} \left(\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}\right)^{\mathbb{1}(c_{dn}=k)} \\
&= \mathsf{Discrete}(\phi_{dn}), \quad \phi_{dn}(k) = \frac{\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}}{\sum_{j=1}^{K} \mathrm{e}^{\mathbb{E}_q[\ln \beta_{j,x_{dn}}]+\mathbb{E}_q[\ln \theta_{d,j}]}}
\end{aligned}
$$

$$
\begin{aligned}
q(\beta_k) &\propto \prod_{v=1}^{V} (\beta_{k,v})^{\gamma-1+\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)} \\
&= \mathsf{Dir}(\gamma_k), \quad \gamma_k(v) = \gamma + \textstyle\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)
\end{aligned}
$$

$$
\begin{aligned}
q(\theta_d) &\propto \prod_{k=1}^{K} (\theta_{dk})^{\alpha-1+\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]} \\
&= \mathsf{Dir}(\alpha_d), \quad \alpha_d(k) = \alpha + \textstyle\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]
\end{aligned}
$$

Continuing the derivation from a previous slide:

$$
\begin{aligned}
q(c_{dn}) &\propto \prod_{k=1}^{K} \left(\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}\right)^{\mathbb{1}(c_{dn}=k)} \\
&= \text{Discrete}(\phi_{dn}), \quad \phi_{dn}(k) = \frac{\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}}{\sum_{j=1}^{K} \mathrm{e}^{\mathbb{E}_q[\ln \beta_{j,x_{dn}}]+\mathbb{E}_q[\ln \theta_{d,j}]}}
\end{aligned}
$$

$$
\begin{aligned}
q(\beta_k) &\propto \prod_{v=1}^{V} (\beta_{k,v})^{\gamma-1+\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)} \\
&= \text{Dir}(\gamma_k), \quad \gamma_k(v) = \gamma + \sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)
\end{aligned}
$$

$$
\begin{aligned}
q(\theta_d) &\propto \prod_{k=1}^{K} (\theta_{dk})^{\alpha-1+\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]} \\
&= \text{Dir}(\alpha_d), \quad \alpha_d(k) = \alpha + \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]
\end{aligned}
$$

Continuing the derivation from a previous slide:

$$
\begin{aligned}
q(c_{dn}) &\propto \prod_{k=1}^{K} \left(\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}\right)^{\mathbb{1}(c_{dn}=k)} \\
&= \mathsf{Discrete}(\phi_{dn}), \quad \phi_{dn}(k) = \frac{\mathrm{e}^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}}{\sum_{j=1}^{K} \mathrm{e}^{\mathbb{E}_q[\ln \beta_{j,x_{dn}}]+\mathbb{E}_q[\ln \theta_{d,j}]}}
\end{aligned}
$$

$$
\begin{aligned}
q(\beta_k) &\propto \prod_{v=1}^{V} (\beta_{k,v})^{\gamma-1+\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)} \\
&= \mathsf{Dir}(\gamma_k), \quad \gamma_k(v) = \gamma + \textstyle\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v)
\end{aligned}
$$

$$
\begin{aligned}
q(\theta_d) &\propto \prod_{k=1}^{K} (\theta_{dk})^{\alpha-1+\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]} \\
&= \mathsf{Dir}(\alpha_d), \quad \alpha_d(k) = \alpha + \textstyle\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]
\end{aligned}
$$

The variational parameters are

1. ϕ_{dn} : A K-dimensional distribution for a discrete distribution
2. γ_k : A V-dimensional parameter for a Dirichlet distribution
3. α_d : A K-dimensional parameter for a Dirichlet distribution

We could have defined *a priori* that

$$q(c_{dn}) = \text{Disc}(\phi_{dn}), \quad q(\beta_k) = \text{Dir}(\gamma_k) \;\; \text{and} \;\; q(\theta_d) = \text{Dir}(\alpha_d)$$

however, following the steps of the previous derivation

1. shows that these are the optimal distributions
2. shows how to update the parameters for these distributions

To appreciate the usefulness of this, try calculating $\mathcal{L}$ with these q, taking derivatives (e.g., w.r.t. γ_k) and solving for the root. (Not your homework!)

Finally, we need to know what to plug in for the expectations

$$
\begin{aligned}
q(c_{dn}) &= \text{Discrete}(\phi_{dn}), \quad \phi_{dn}(k) = \frac{e^{\mathbb{E}_q[\ln \beta_{k,x_{dn}}]+\mathbb{E}_q[\ln \theta_{dk}]}}{\sum_{j=1}^{K} e^{\mathbb{E}_q[\ln \beta_{j,x_{dn}}]+\mathbb{E}_q[\ln \theta_{d,j}]}} \\
q(\beta_k) &= \text{Dir}(\gamma_k), \quad \gamma_k(v) = \gamma + \textstyle\sum_d \sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]\mathbb{1}(x_{dn}=v) \\
q(\theta_d) &= \text{Dir}(\alpha_d), \quad \alpha_d(k) = \alpha + \textstyle\sum_n \mathbb{E}_q[\mathbb{1}(c_{dn}=k)]
\end{aligned}
$$

We can look these up in a textbook to find that

$$
\begin{aligned}
\mathbb{E}_q[\mathbb{1}(c_{dn}=k)] &= \phi_{dn}(k) \\
\mathbb{E}_q[\ln \beta_{k,x_{dn}}] &= \psi(\gamma_k(x_{dn})) - \psi(\textstyle\sum_v \gamma_k(v)) \\
\mathbb{E}_q[\ln \theta_{dk}] &= \psi(\alpha_d(k)) - \psi(\textstyle\sum_j \alpha_d(j))
\end{aligned}
$$

$\psi(\cdot)$ is the digamma function. Call a built-in function when coding.

Variational inference for LDA

Input documents $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ and number of topics K

Output variational parameters ϕ_{dn}, γ_k and α_d for all d, k, n

Initialize each γ_k in some way and α_d to a vector of 1's

For iteration t

1. For each n and d, update ϕ_{dn} by setting

$$\phi_{dn}(k) = \frac{\mathrm{e}^{\psi(\gamma_k(x_{dn})) - \psi(\sum_v \gamma_k(v)) + \psi(\alpha_d(k))}}{\sum_{k'} \mathrm{e}^{\psi(\gamma_{k'}(x_{dn})) - \psi(\sum_v \gamma_{k'}(v)) + \psi(\alpha_d(k'))}}, \quad k = 1, \ldots, K$$

2. For each d, update α_d by setting

$$\alpha_d(k) = \alpha + \textstyle\sum_n \phi_{dn}(k), \quad k = 1, \ldots, K$$

3. For each k, update γ_k by setting

$$\gamma_k(v) = \gamma + \textstyle\sum_d \sum_n \phi_{dn}(k) \mathbb{1}(x_{dn} = v), \quad v = 1, \ldots, V$$

4. Calculate $\mathcal{L} = \mathbb{E}_q[\ln p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})] - \mathbb{E}_q[\ln q(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{c})]$ for convergence