Latent Dirichlet Allocation

Spring 2024

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing

THE UNIVERSITY OF UTAH

Outline

- Latent Dirichlet Allocation
- Variational inference

Outline

- Latent Dirichlet Allocation
- Variational inference

Latent Dirichlet Allocation

- A classical text mining model that extract topics from the text corpus
- Broadly used in all kinds of text mining and related tasks: information retrieval, text classification, advertisement keywords extraction, sentimental analysis, ….
- https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4
- A very good example to study Bayesian learning

Latent Dirichlet Allocation

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Latent Dirichlet Allocation

- The original paper

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

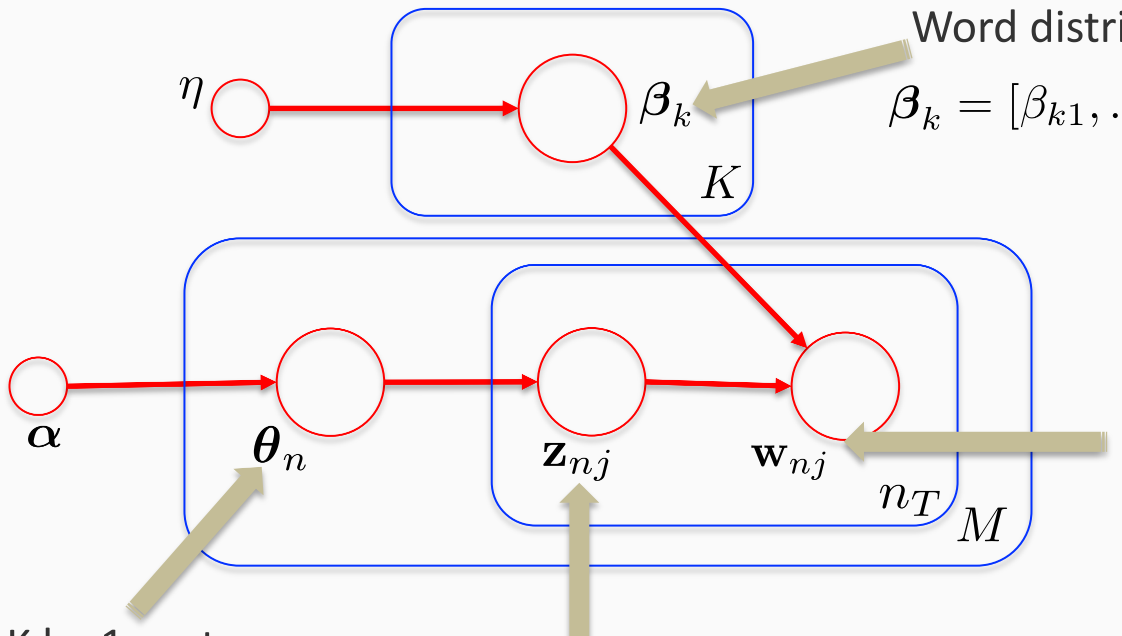
http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

LDA: sampling procedure

- Given *K* topics
- First sample *K* topics (word distributions)
- For each document in the corpus
  - Sample topic mixture distribution
  - For each word in the document
    - Sample the topic index, according to which to sample the word

LDA: graphical model

Suppose we have ***V*** words, ***M*** documents, document ***n*** has $\boldsymbol{n_T}$ words



Word distribution

$$\boldsymbol{\beta}_k = [\beta_{k1}, \ldots, \beta_{kV}]^\top$$

Multinomial parameters

V by 1 one-hot vector
word index

$$\mathbf{w}_{nj} = [w_{nj1}, \ldots, w_{njV}]^\top$$

Only one element is 1

K by 1 vector
Topic mixture

$$\boldsymbol{\theta}_n = [\theta_{n1}, \ldots, \theta_{nK}]^\top$$

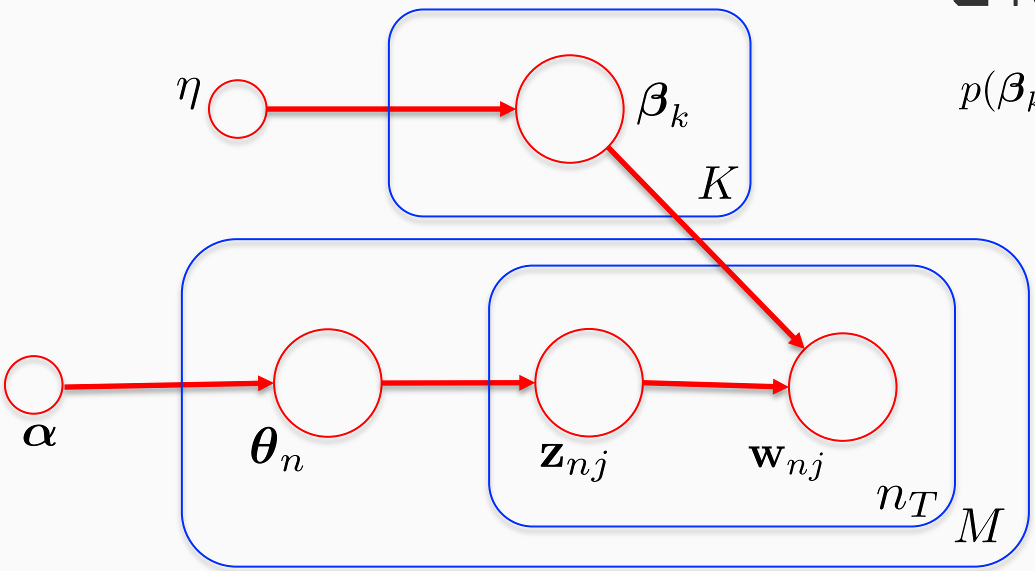
Multinomial parameters

K by 1 one-hot vector
Topic index

$$\mathbf{z}_{nj} = [z_{nj1}, \ldots, z_{njK}]^\top$$

Only one element is 1

LDA: sampling procedure



- For each topic

$$p(\boldsymbol{\beta}_k|\eta) = \mathrm{Dir}(\boldsymbol{\beta}_k|\eta) = \frac{\Gamma(V\eta)}{\prod_{v=1}^{V}\Gamma(\eta)} \prod_{v=1}^{V} \beta_{kv}^{\eta-1}$$

- For each document

$$p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}_n|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{nk}^{\alpha_k-1}$$

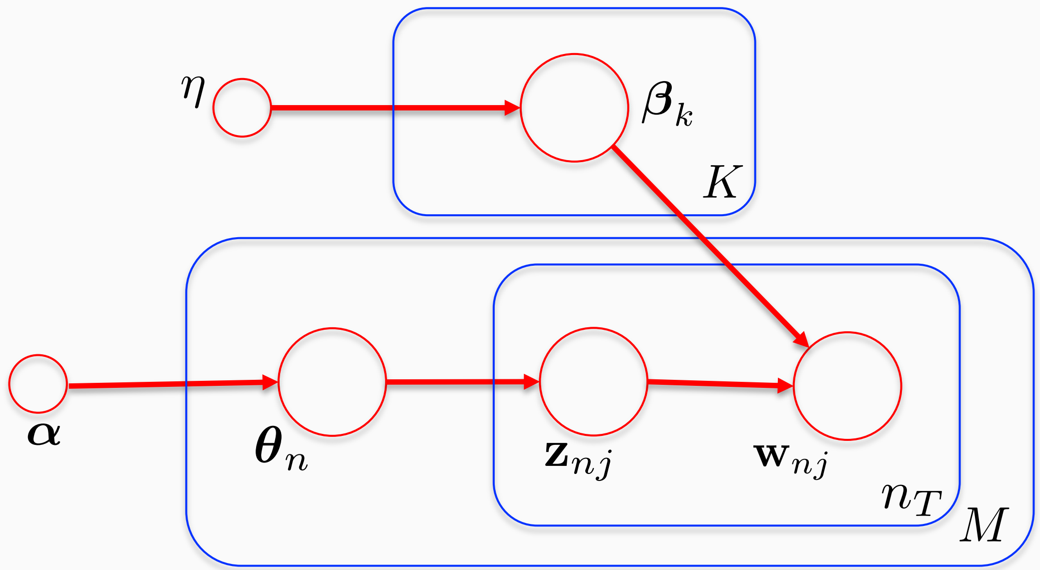
- For each word

$$p(\mathbf{z}_{nj}|\boldsymbol{\theta}_n) = \mathrm{Mul}(\mathbf{z}_{nj}|\boldsymbol{\theta}_n) = \prod_{k=1}^{K} \theta_{nk}^{z_{njk}}$$

$$p(\mathbf{w}_{nj}|z_{njk}=1, \boldsymbol{\beta}) = \mathrm{Mul}(\mathbf{w}_{nj}|\boldsymbol{\beta}_k) = \prod_{v=1}^{V} \beta_{kv}^{w_{njv}}$$

$$p(\mathbf{w}_{nj}|\mathbf{z}_{nj}, \boldsymbol{\beta}) = \prod_{k=1}^{K}\prod_{v=1}^{V} \beta_{kv}^{z_{njk}w_{njv}}$$

LDA: joint probability



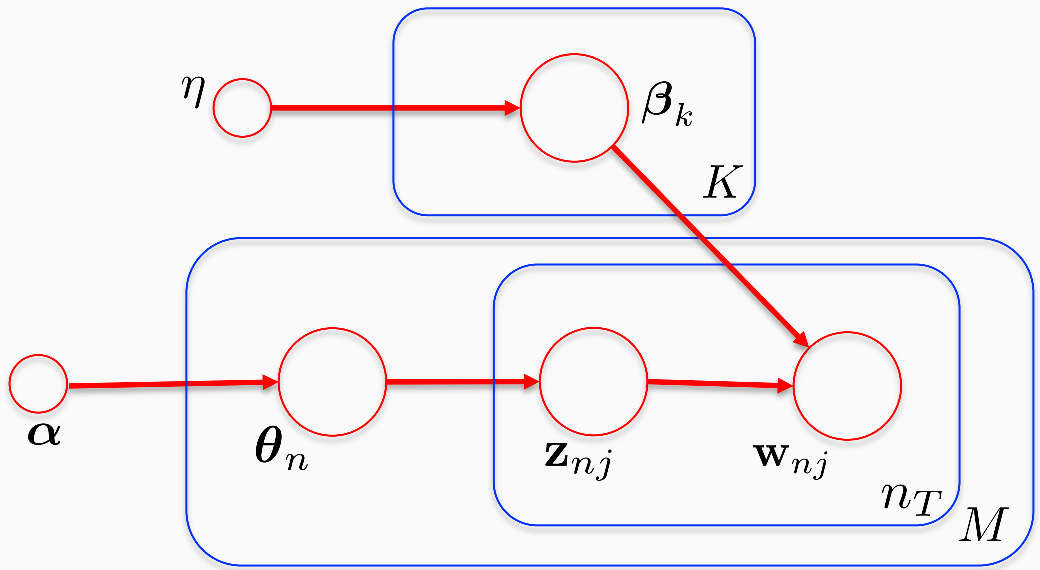
$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{W} | \eta, \boldsymbol{\alpha})$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_k | \eta) \prod_{n=1}^{M} \mathrm{Dir}(\boldsymbol{\theta}_n | \boldsymbol{\alpha}) \left\{ \prod_{j=1}^{n_T} \mathrm{Mul}(\mathbf{z}_{nj} | \boldsymbol{\theta}_n) \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\}$$

Outline

- Latent Dirichlet Allocation
- Variational inference

LDA: inference



How can we compute the posterior of

$\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ Topic words: critical for numerous tasks

$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ Topic mixture: low-rank representation of docs

True posterior is intractable to compute

LDA: variational inference

- We use variational EM algorithm (empirical Bayes)

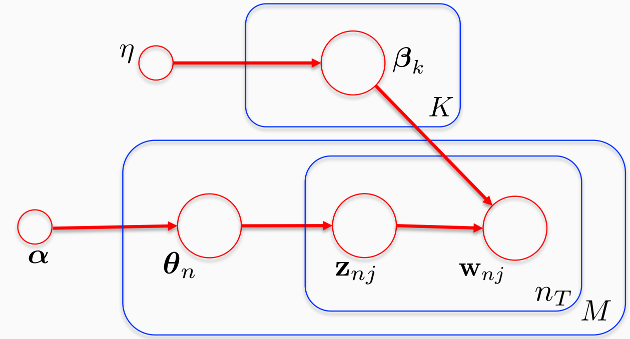
  ❑ E step: mean-field update

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}) = \prod_{k=1}^{K} q(\boldsymbol{\beta}_k) \prod_{n=1}^{M} \left[q(\boldsymbol{\theta}_n) \prod_{j=1}^{n_T} q(\mathbf{z}_{nj}) \right]$$

  ❑ M step

  Maximize variational lower bound of the model evidence w.r.t $\boldsymbol{\alpha}, \eta$

LDA: variational E update



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{W} | \eta, \boldsymbol{\alpha})$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_k|\eta) \prod_{n=1}^{M} \mathrm{Dir}(\boldsymbol{\theta}_n|\boldsymbol{\alpha}) \left\{ \prod_{j=1}^{n_T} \mathrm{Mul}(\mathbf{z}_{nj}|\boldsymbol{\theta}_n) \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\}$$

Update each $q(\mathbf{z}_{nj})$

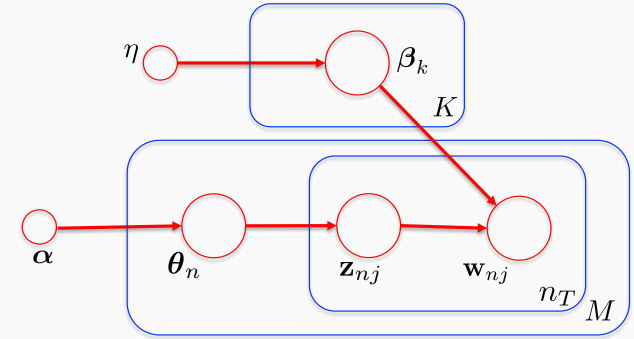
$$q(\mathbf{z}_{nj}) \propto \exp\left(\mathbb{E}_q \log \left[\mathrm{Mul}(\mathbf{z}_{nj}|\boldsymbol{\theta}_n) \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right)$$

$$\exp\left(\sum_{k=1}^{K} z_{njk} \left[\mathbb{E}_q[\log \theta_{nk}] + \sum_{v=1}^{V} w_{njv} \mathbb{E}_q[\log \beta_{kv}] \right] \right)$$

$$q(\mathbf{z}_{nj}) = \mathrm{Mul}(\mathbf{z}_{nj}|\boldsymbol{\phi}_{nj})$$

$$\phi_{njk} \propto \exp\left(\mathbb{E}_q \log[\theta_{nk}] + \sum_{v=1}^{V} w_{njv} \mathbb{E}_q \log \beta_{kv} \right)$$

LDA: variational E update



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{W} | \eta, \boldsymbol{\alpha})$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_k | \eta) \prod_{n=1}^{M} \mathrm{Dir}(\boldsymbol{\theta}_n | \boldsymbol{\alpha}) \left\{ \prod_{j=1}^{n_T} \mathrm{Mul}(\mathbf{z}_{nj} | \boldsymbol{\theta}_n) \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\}$$

Update each $q(\boldsymbol{\theta}_n)$

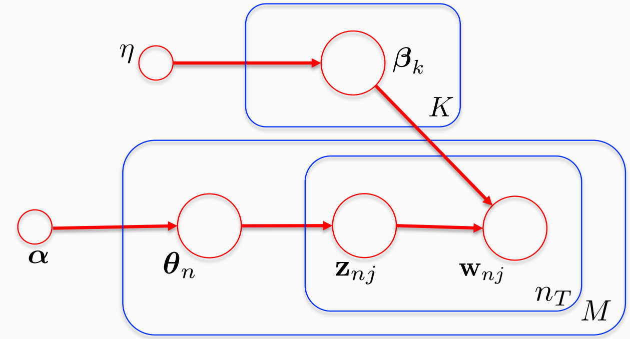
$$q(\boldsymbol{\theta}_n) \propto \exp\left(\mathbb{E}_q \left[\log[\mathrm{Dir}(\boldsymbol{\theta}_n | \boldsymbol{\alpha})] + \sum_{j=1}^{n_T} \log[\mathrm{Mul}(\mathbf{z}_{nj} | \boldsymbol{\theta}_n)] \right] \right)$$

$$\sum_{k=1}^{K} (\alpha_k + \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}] - 1) \log \theta_{nk}$$

$$q(\boldsymbol{\theta}_n) = \mathrm{Dir}(\boldsymbol{\theta}_n | \gamma_n)$$

$$\gamma_{nk} = \alpha_k + \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}]$$

LDA: variational E update



$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{W} | \eta, \boldsymbol{\alpha})$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_k | \eta) \prod_{n=1}^{M} \mathrm{Dir}(\boldsymbol{\theta}_n | \boldsymbol{\alpha}) \left\{ \prod_{j=1}^{n_T} \mathrm{Mul}(\mathbf{z}_{nj} | \boldsymbol{\theta}_n) \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\}$$

Update each $q(\boldsymbol{\beta}_k)$

$$q(\boldsymbol{\beta}_k) \propto \exp\left(\mathbf{E}_q \left[\log[\mathrm{Dir}(\boldsymbol{\beta}_k | \eta) + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \sum_{v=1}^{V} z_{njk} w_{njv} \beta_{kv}] \right] \right)$$

$$\sum_{v=1}^{V} \log[\beta_{kv}] (\eta + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}] w_{njv} - 1)$$

$$q(\boldsymbol{\beta}_k) = \mathrm{Dir}(\boldsymbol{\beta}_k | \boldsymbol{\psi}_k)$$

$$\psi_{kv} = \eta + \sum_{n=1}^{M} \sum_{j=1}^{n_T} \mathbb{E}_q[z_{njk}] w_{njv}$$

LDA: variational E update

- How to compute the required moments in the update?

$$q(\mathbf{z}_{nj}) = \mathrm{Mul}(\mathbf{z}_{nj}|\boldsymbol{\phi}_{nj})$$

$$\mathbb{E}_q[z_{njk}]$$

$$q(\boldsymbol{\theta}_n) = \mathrm{Dir}(\boldsymbol{\theta}_n|\boldsymbol{\gamma}_n)$$

$$\mathbb{E}_q \log[\theta_{nk}]$$

$$q(\boldsymbol{\beta}_k) = \mathrm{Dir}(\boldsymbol{\beta}_k|\boldsymbol{\psi}_k)$$

$$\mathbb{E}_q \log \beta_{kv}$$

Leave it as your review!

LDA: variational M step

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{W} | \eta, \boldsymbol{\alpha})$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\beta}_k | \eta) \prod_{n=1}^{M} \mathrm{Dir}(\boldsymbol{\theta}_n | \boldsymbol{\alpha}) \left\{ \prod_{j=1}^{n_T} \mathrm{Mul}(\mathbf{z}_{nj} | \boldsymbol{\theta}_n) \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{z_{njk} w_{njv}} \right] \right\}$$

$$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{nk}^{\alpha_k - 1}$$

Variational lower bound

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{M} \log \Gamma(\sum_{k=1}^{K} \alpha_k) - \sum_{n=1}^{M} \sum_{k=1}^{K} \log \Gamma(\alpha_k)$$
$$+ \sum_{n=1}^{M} \sum_{k=1}^{K} (\alpha_k - 1) \mathbb{E}_q[\log \theta_{nk}] + \mathrm{const}$$

LDA: variational M step

Variational lower bound

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{M} \log\Gamma(\sum_{k=1}^{K} \alpha_k) - \sum_{n=1}^{M}\sum_{k=1}^{K} \log\Gamma(\alpha_k)$$
$$+ \sum_{n=1}^{M}\sum_{k=1}^{K}(\alpha_k - 1)\mathbb{E}_q[\log\theta_{nk}] + \text{const}$$

$$\frac{\partial\mathcal{L}}{\partial\alpha_k} = M\Psi(\sum_{k=1}^{K}\alpha_k) - M\Psi(\alpha_k) + \sum_{n=1}^{M}\mathbb{E}_q[\log\theta_{nk}]$$

Digamma function

LDA: variational M step

- Derive $\frac{\partial \mathcal{L}}{\partial \eta}$

Note that η is a scalar

Leave it as your exercise

Use any gradient based algorithm with constraints

$\boldsymbol{\alpha} > \mathbf{0}, \eta > 0$ e.g., LBFGS-B

What you need to do

- Write down LDA sampling procedure and joint probability
- Derive the variational E updates and gradients for M step
- Implement an algorithm for LDA inference and test it on real-world data (see homework assignments).