Natural Language Processing (4)

Topic Models

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1)
8. Parsing (2)
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1)
13. Guest Talk (2)
14. Project: Survey or Programming
15. Project Presentation

Review: Distributional Hypothesis

- Linguistic items with similar distributions have similar meanings
- To obtain such distributions, we typically count co-occurring words in the context of the target word
  - dog = (eat:48, bite:31, bark:63, lick:23, ...)
  - cat = (eat:29, bite:13, bark:9, lick:47, ...)

Topic models

- Word probabilities vary according to topics:
  - domains, themes, meanings
  - time, regions
  - documents, sections, paragraphs
  - styles, authors, languages
- Estimate latent topics hidden in corpora

Example: frequency of "said"

	Frequency / 10^6 words
• Department of Energy Abst.	41
• Groliers Encyclopedia	64
• Federalist Papers	287
• Hansard	1072
• Harper & Row Books	1632
• Brown Corpus	1645
• Wall Street Journal	5600
• Associated Press 1990	10040

[Church and Gale, 95]

Example: topic model

"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

[Blei+ 2003]

Example: topic model

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.

[Blei+ 2003]

Naïve Bayes

- A simple generative model
- For example, classify whether an email is a spam or a non-spam
  - spam: $k=1$
  - non-spam: $k=0$

$$\arg\max_k p(k \mid d) = \arg\max_k p(k)p(d \mid k) = \arg\max_k p(k)\prod_{w \in d} p(w \mid k)$$

Example

$$
D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array}
\begin{array}{c}
\begin{array}{ccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array} \\
\begin{pmatrix}
1 & 2 & 1 & & 1 & & \\
 & 2 & & & 1 & 1 & 1 \\
1 & & 1 & 1 & & 2 &
\end{pmatrix}
\end{array}
$$

non-spam (k=0)

spam (k=1)

$p(k{=}0){=}?$, $k(k{=}1){=}?$

$p(w_1|k{=}0){=}?$, $p(w_2|k{=}0){=}?$, ...

$p(w_1|k{=}1){=}?$, $p(w_2|k{=}1){=}?$, ...

Example

- $d_{new} = \{w_1,\ w_6\}$: spam or non-spam?

$$\arg\max_k p(k \mid d) = \arg\max_k p(k) p(d \mid k) = \arg\max_k p(k) \prod_{w \in d} p(w \mid k)$$

Prior and posterior

- posterior ∝ prior × likelihood

$$p(k \mid d) \propto p(k)\,p(d \mid k)$$

Unigram Mixtures (UM)

[Nigam+ 2000]

- We usually do not have a corpus with category (or topic) assignments

- Naïve Bayes:

$$p(d,k) = p(k)\prod_{w\in d} p(w \mid k)$$

- Unigram mixtures:

$$p(d) = \sum_{k} p(k)\prod_{w\in d} p(w \mid k)$$

  – k: a latent variable

UM: parameter estimation

- EM algorithm
  1. Expectation: estimate $p(k|d)$
  2. Maximization: estimate $p(k)$ and $p(w|k)$

UM: example

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array} \\ \begin{pmatrix} 1 & 2 & 1 & & 1 & & \\ & 2 & & & 1 & 1 & 1 \\ 1 & & 1 & 1 & & 2 & \end{pmatrix} \end{array}$$

1. $p(k|d_1) = [0.4, 0.6]$, $p(k|d_2) = [0.6, 0.4]$, $p(k|d_3) = [0.4, 0.6]$
2. estimate $p(k)$

$$p(k) = \frac{\sum_d p(k \mid d)}{\sum_k \sum_d p(k \mid d)}$$

UM: example

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array} \\ \begin{pmatrix} 1 & 2 & 1 & & 1 & & \\ & 2 & & & 1 & 1 & 1 \\ 1 & & 1 & 1 & & 2 & \end{pmatrix} \end{array}$$

1. $p(k|d_1) = [0.4, 0.6]$, $p(k|d_2) = [0.6, 0.4]$, $p(k|d_3) = [0.4, 0.6]$
2. estimate $p(k)$ and $p(w|k)$

$$p(w \mid k) = \frac{\sum_d p(k \mid d) n(d, w)}{\sum_w \sum_d p(k \mid d) n(d, w)}$$

UM: example

$$
\begin{array}{cc}
 & \begin{array}{ccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array} \\
D = \begin{array}{c} d_1 \\ d_2 \\ d_3 \end{array} & \begin{pmatrix} 1 & 2 & 1 & & 1 & & \\ & 2 & & & 1 & 1 & 1 \\ 1 & & 1 & 1 & & 2 & \end{pmatrix}
\end{array}
$$

1. $p(k|d_1) = [0.4, 0.6]$, $p(k|d_2) = [0.6, 0.4]$, $p(k|d_3) = [0.4, 0.6]$
2. estimate $p(k)$ and $p(w|k)$
3. update $p(k|d)$

$$
p(k \mid d) = \frac{p(k)\prod_{w \in d} p(w \mid k)}{\sum_k p(k)\prod_{w \in d} p(w \mid k)}
$$

UM: results

- Topical words according to $p(w|k)$ in a newspaper corpus

Topic 5
の,を,に,する,細胞,
など,船,レーザー,
ブロック,し,こと,
靴,から,や,な,型,
銀河,融合,核,足,
研究,状,宇宙,評価,
が,方法,サイズ,不審,
物質,高速,なる,ず,
意見,建造,グループ,星

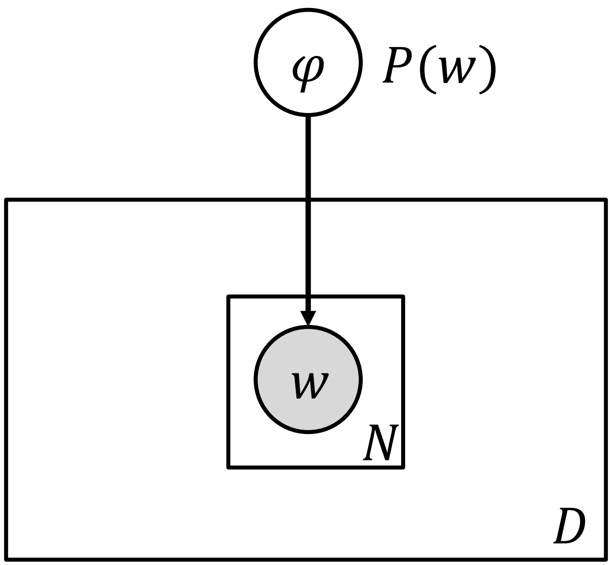
Topic 10
た,し,に,と,が,て,
者,い,処分,こと,
は,生徒,れ,人,さ,
を,教委,問題,府,
女子,男性,被害,
保護,県,生活,保険,
など,ない,浪人,
よる,あっ,教職員

Topic 100
た,さん,て,で,容疑,
い,調べ,ごろ,と,捜査,
署,市,れ,時,者,事件,
が,いる,し,逮捕,午後,
男,み,県,県警,分,
男性,本部,殺人,いう,
から,午前,町,車,
同署,人,員,死亡,疑い,
乗用車,女性,府警

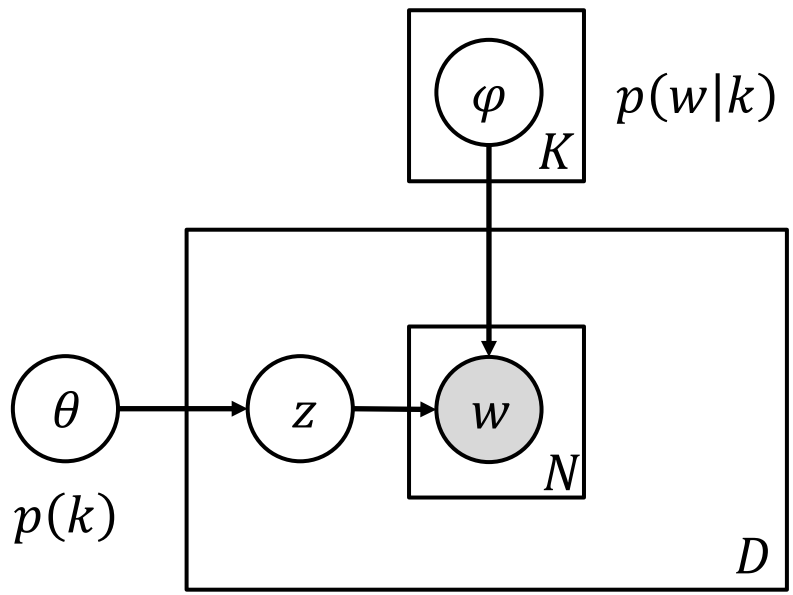
[Mochihashi 2012]

Graphical models

- Unigram



- Unigram mixtures



UM: summary

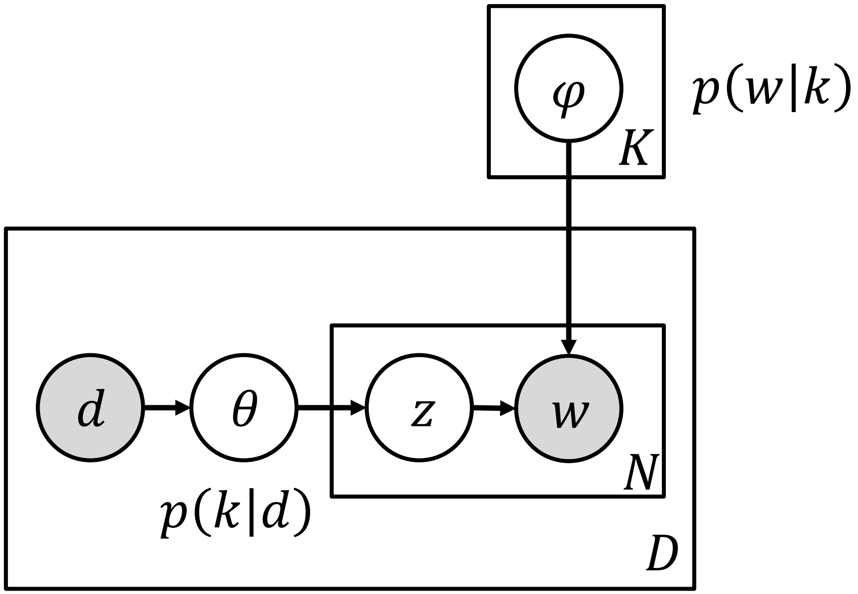
- The simplest topic model
- One latent topic for each document
- Parameters: $p(k), p(w|k)$

Better topic models

- Probabilistic Latent Semantic Indexing (PLSI)
- Latent Dirichlet Allocation (LDA)

PLSI

- Topic k is assigned to each word
- Topic distribution $p(k|d)$ is defined for each document d
  - Topic k is generated for each word from $p(k|d)$
  - Word w is generated from topic k



PLSI

- Unigram mixtures

$$p(d) = \sum_k p(k) \prod_{w \in d} p(w \mid k) \qquad p(d) = \sum_k p(k) \prod_{w \in V} p(w \mid k)^{n(d,w)}$$

- PLSI

$$p(d,w) = p(d) p(w \mid d)$$
$$= p(d) \sum_k p(w \mid k) p(k \mid d)$$

$$p(d,w) = p(d) \sum_k p(w \mid k) \frac{p(d \mid k) p(k)}{p(d)}$$
$$= \sum_k p(k) p(d \mid k) p(w \mid k)$$

PLSI: parameter estimation

- EM algorithm
  - Initialization: initialize $p(k)$, $p(d|k)$, $p(w|k)$
  - E-step: estimate topic distribution $p(k|d, w)$

$$p(k \mid d, w) \propto p(k) p(d \mid k) p(w \mid k)$$

  - M-step: update $p(k)$, $p(d|k)$, $p(w|k)$

PLSI: results

- Topical words according to $p(w|k)$ in a newspaper corpus

Topic 1
先,後,#,歩,銀,四,
五,六,同,二,飛,
八,成,玉,七,三,
金,九,桂,角,と,
谷川,が,た,手,は,
丸山,一,香,の,で,
局,図,戦,段

Topic 2
の,号,事故,機,が,
た,に,安全,#,部分,
を,原発,原因,は,
基,水,運転,装置,
爆発,器,原子力,
炉,作業,し,燃料,
で,漏れ,発生,と,
配管,原子,ガス

Topic 3
#,勝,敗,戦,
イチロー,日,回,
リーグ,大リーグ,
マリナーズ,新庄,
試合,安打,点,ス,
で,手,共同,メッツ,
外野,は,大,投手,
第,米,の,打席,
ソックス,
ヤンキース,記録,
バックス,打率,
ニューヨーク

Topic 4
研究,細胞,
遺伝子,移植,
の,治療,物質,
教授,を,患者,
科学,脳,医療,
病院,ローン,
ヒト,実験,薬,
グループ,遺伝,
が,臓器,体,ク,
病,する,に,学会,
さ,DNA,開発,
臨床,人間,神経

[Mochihashi, 12]

PLSI: summary

- By considering latent topics for each word, we can obtain a better model than unigram mixtures
  - Unigram mixtures:
    1. Generate a topic *k* from a mixing ratio
    2. Generate a document (words) from topic *k*
  - PLSI
    1. Generate a mixing ratio for each document
    2. Select a topic *k* from the mixing ratio
    3. Generate a word from topic *k*

PLSI: problems

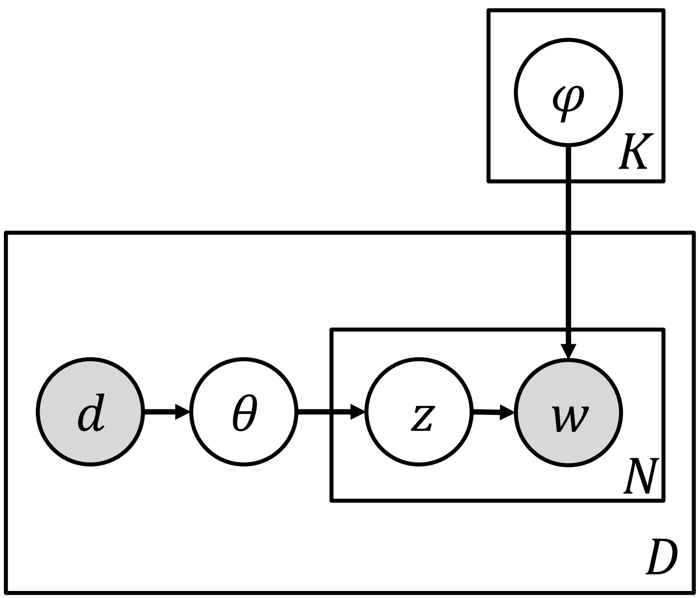
- $p(k|d_{new})$ is undefined
  - $p(k|d)$ is defined only for training data
  - $p(k|d_{new})$ can be approximately calculated (ad hoc)
  - $p(k|d)$ should be probabilistically generated (→ LDA)
- Parameters of PLSI: $p(k)$, $\boldsymbol{p(d|k)}$, $p(w|k)$
  - The number of parameters is large
  - PLSI is likely to overfit to training data

LDA

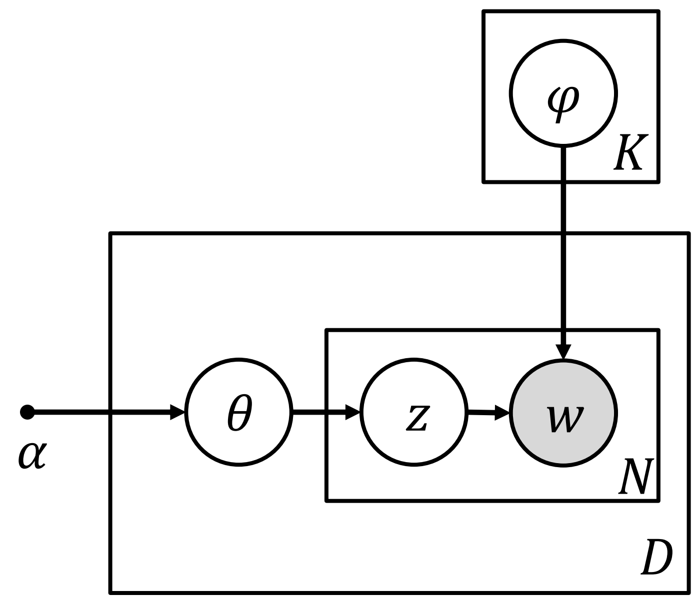
- PLSI
  1. Generate a mixing ratio $\theta = p(k|d)$ (fixed) for each document
  2. Generate a topic *k* from the mixing ratio $\theta = p(k|d)$
  3. Generate a word from topic *k*
- LDA
  1. Generate a topic distribution $\theta \sim p(\theta|\alpha)$
  2. Generate a topic k from θ
  3. Generate a word from topic *k*

LDA

- PLSI



- LDA



Dirichlet distribution

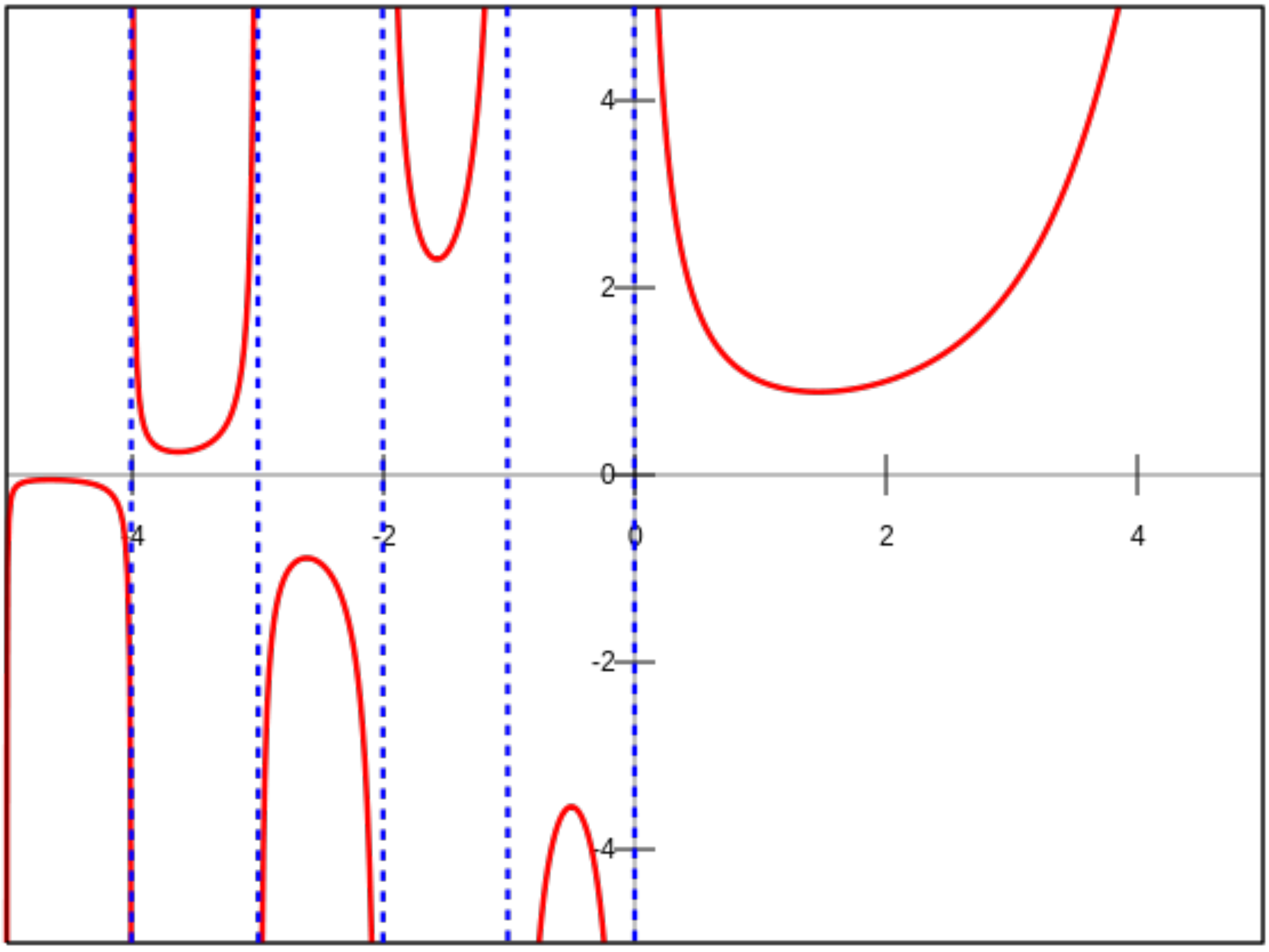
$$\mathrm{Dir}(\theta \mid \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

where $\alpha_k > 0 \quad (k = 1, \cdots, K)$

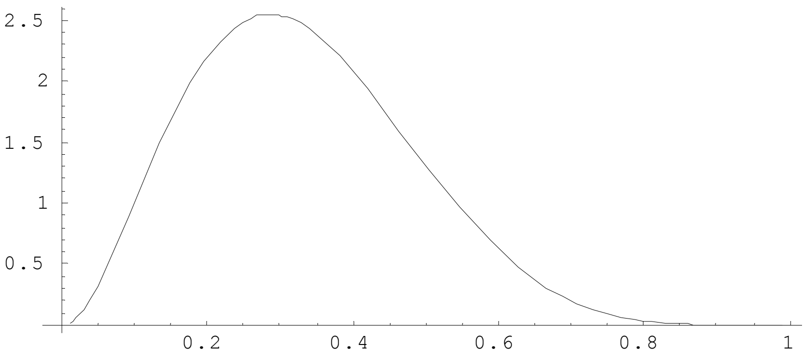
$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x}\, dx$$

※ $\Gamma(\mathrm{n}) = (\mathrm{n} - 1)!$

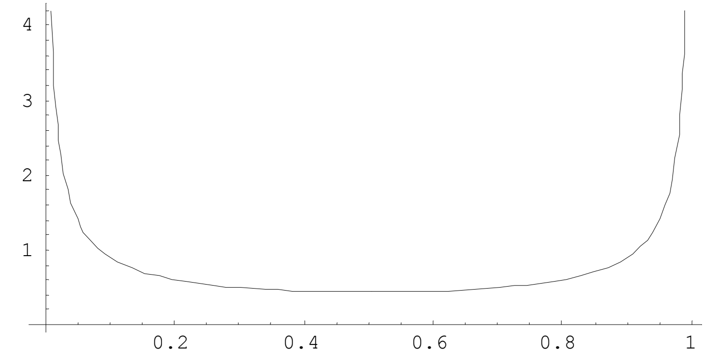
Gamma function



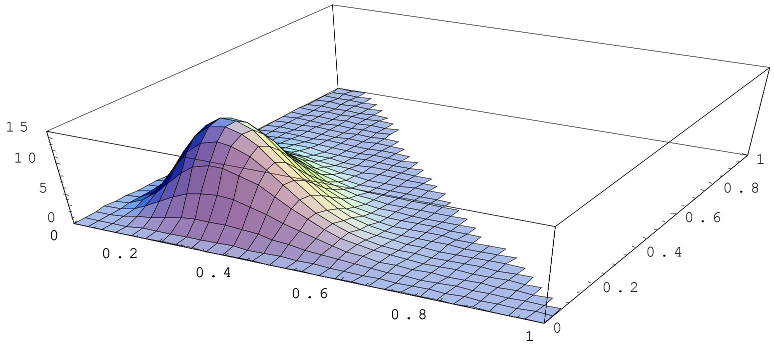
Dirichlet distribution



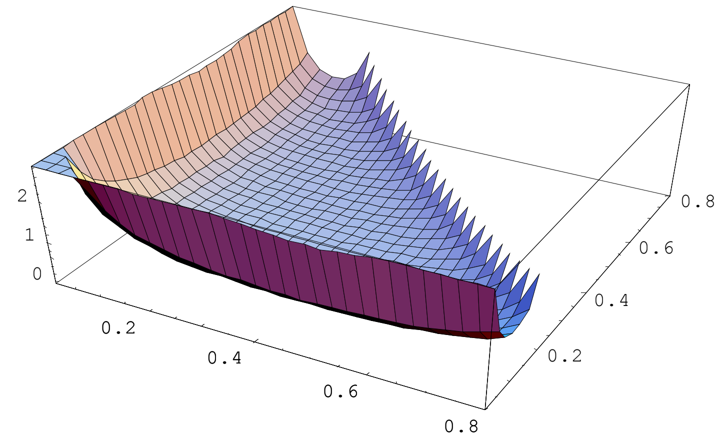
$$\alpha_1 = 3, \alpha_2 = 6$$



$$\alpha_1 = 0.3, \alpha_2 = 0.3$$



$$\alpha_1 = 4, \alpha_2 = 4, \alpha_3 = 8$$



$$\alpha_1 = 0.3, \alpha_2 = 0.3, \alpha_3 = 0.3$$

Dirichlet distribution

- The Dirichlet distribution is widely used as a prior for the multinomial distribution
  - Conjugate distribution
  - posterior ∝ prior × likelihood

$$p(\theta \mid n, \alpha) \propto \mathrm{Dir}(\theta \mid \alpha)\,\mathrm{Multi}(n \mid \theta)$$

Multinomial distribution

$$\mathrm{Multi}(n \mid \theta) = \frac{N!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \theta_k^{n_k}$$

$$= \frac{\Gamma(N+1)}{\prod_{k=1}^{K} \Gamma(n_k + 1)} \prod_{k=1}^{K} \theta_k^{n_k}$$

	w_1	w_2	$\cdots$	w_K
prob	θ_1	θ_2	$\cdots$	θ_K
freq	n_1	n_2	$\cdots$	n_K

where $\Gamma(n) = (n-1)!$

then $p(\theta \mid n, \alpha) \propto \mathrm{Dir}(\theta \mid n + \alpha)$

LDA: parameter estimation

- Variational Bayes [Blei+, 03]
  http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Gibbs Sampling [Griffiths and Steyvers, 04]
  http://psiexp.ss.uci.edu/research/papers/sciencetopics.pdf
- Collapsed Variational Bayes [Teh+, 06]
  https://papers.nips.cc/paper/3113-a-collapsed-variational-bayesian-inference-algorithm-for-latent-dirichlet-allocation

Example: topic model

"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

[Blei+ 2003]

Example: topic model

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.

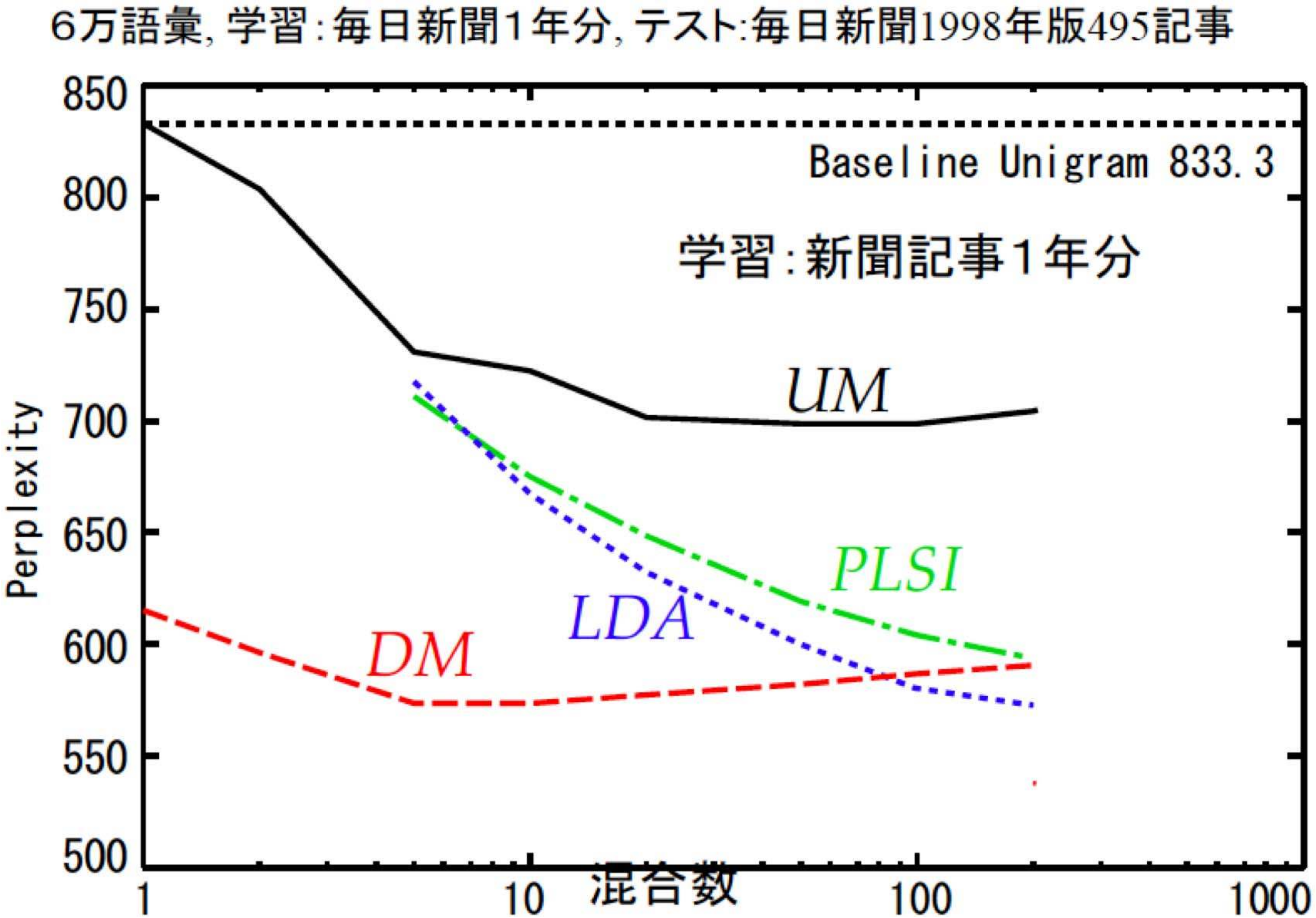
[Blei+ 2003]

Evaluation

- Training and testing a model
  - Estimate parameters θ that maximize the probability of data $p(D|\theta)$
  - A better model gets a higher probability $p(D'|\theta)$ for new data D'
- Evaluation measure: perplexity
  - $p(D'|\theta)$ depends on the number of data N
  - Consider $p(D'|\theta)^{1/N}$ and take its reciprocal

$$\mathrm{PPL} = p(D'|\theta)^{-\frac{1}{N}} = \exp\left(-\frac{1}{N}\log p(D'|\theta)\right)$$

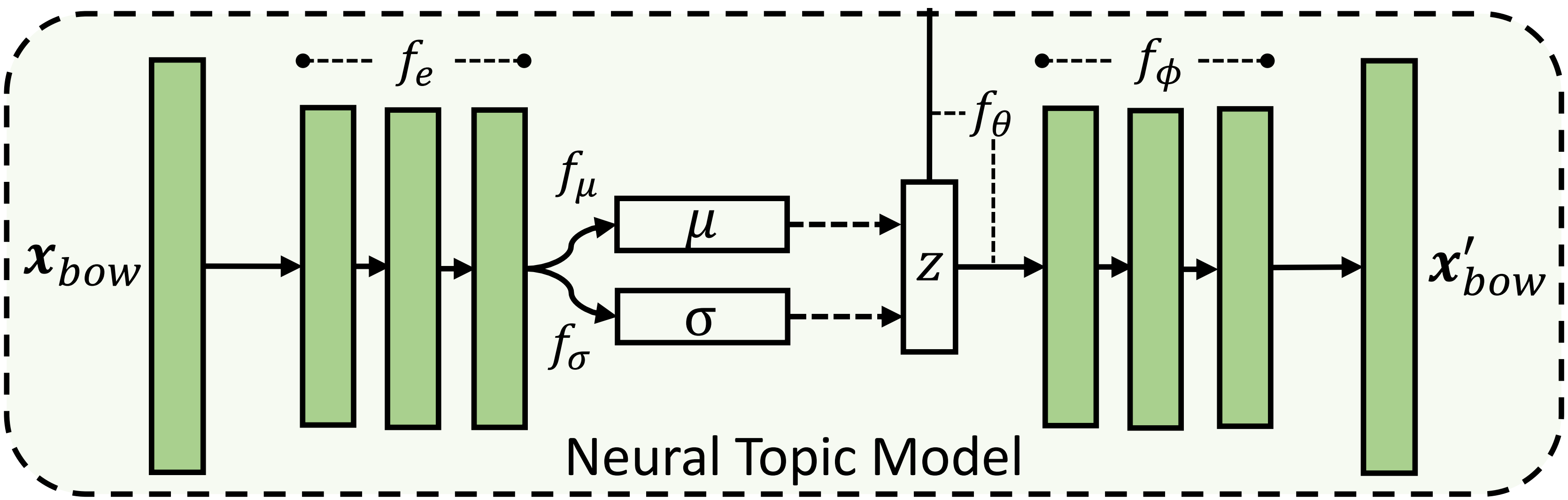
Evaluation



[Yamamoto and Mochihashi 2006]

Neural Topic Models

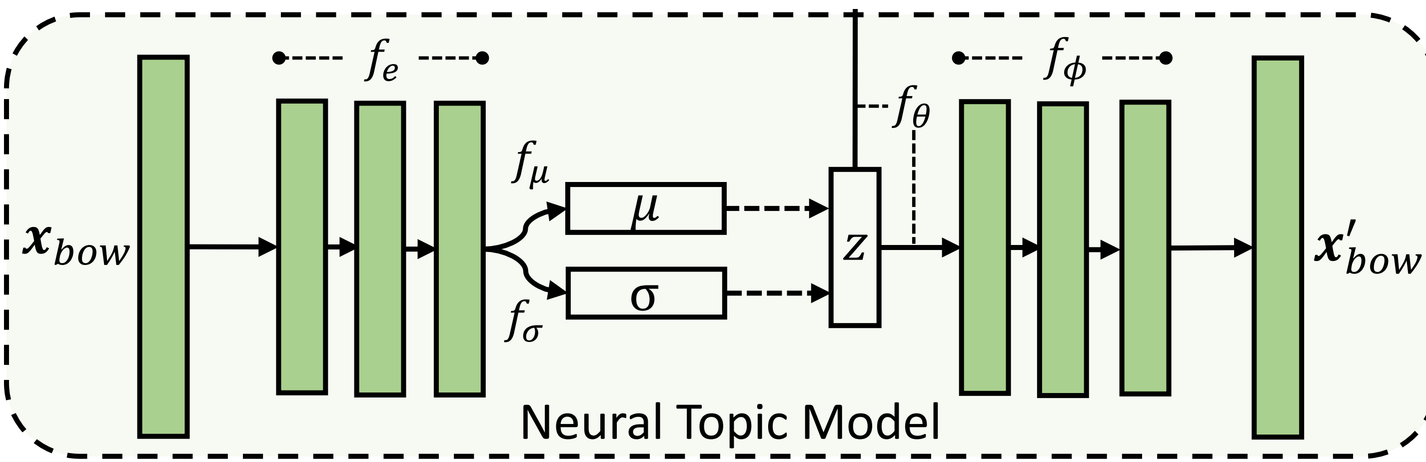
- [Miao+ 2017]
  - Variational autoencoder
  - Gaussian softmax



[Wang+ 2019]

Neural Topic Models

- Encoder
  - $\mu = f_u\big(f_e(\boldsymbol{x}_{bow})\big)$
  - $\log \sigma = f_\sigma\big(f_e(\boldsymbol{x}_{bow})\big)$
- Decoder
  - Latent topic variable $\boldsymbol{z} \sim \mathcal{N}(\mu, \sigma^2)$
  - Topic mixture $\theta = softmax\big(f_\theta(\boldsymbol{z})\big)$
  - $w \in \boldsymbol{x} \sim softmax\Big(f_\phi(\theta)\Big)$



$f_*(x)$: multi-layer perceptron (MLP)

Summary

- Topic models
  - Uni-topic models: unigram mixtures
  - Multi-topic models: PLSI, LDA
  - Neural topic models
- Evaluation measure
  - Perplexity