

CS247: ADVANCED DATA MINING


05: Text Data: Topic Models

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

January 31, 2024

Text Data: Topic Models

- Text Data and Topic Models 
- Multinomial Mixture Model
- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA)
- Summary

Text Data

- Word/term
- Document
 - A sequence of words
- Corpus
 - A collection of documents



Represent a Document

- Most common way: Bag-of-Words
 - Ignore the order of words
 - keep the count

	c1	c2	c3	c4	c5	m1	m2	m3	m4
<i>human</i>	1	0	0	1	0	0	0	0	0
<i>interface</i>	1	0	1	0	0	0	0	0	0
<i>computer</i>	1	1	0	0	0	0	0	0	0
<i>user</i>	0	1	1	0	1	0	0	0	0
<i>system</i>	0	1	1	2	0	0	0	0	0
<i>response</i>	0	1	0	0	1	0	0	0	0
<i>time</i>	0	1	0	0	1	0	0	0	0
<i>EPS</i>	0	0	1	1	0	0	0	0	0
<i>survey</i>	0	1	0	0	0	0	0	0	1
<i>trees</i>	0	0	0	0	0	1	1	1	0
<i>graph</i>	0	0	0	0	0	0	1	1	1
<i>minors</i>	0	0	0	0	0	0	0	1	1

Vector space model

More Details

- Represent the doc as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it)
 - Number of words is huge
 - Select and use a smaller set of words that are of interest
 - E.g. uninteresting words: 'and', 'the', 'at', 'is', etc. These are called stop-words
 - Stemming: remove endings. E.g. 'learn', 'learning', 'learnable', 'learned' could be substituted by the single stem 'learn'
 - Other simplifications can also be invented and used
 - The set of different remaining words is called dictionary or vocabulary. Fix an ordering of the terms in the dictionary so that you can operate them by their index.
 - Can be extended to bi-gram, tri-gram, or so

Limitations of Vector Space Model

- Dimensionality
 - High dimensionality
- Sparseness
 - Most of the entries are zero
- Shallow representation
 - The vector representation does not capture semantic relations between words

D1: I love romantic movies.

D2: Kate Winslet is my favorite actress.

Topics

- Topic
- A topic is represented by a word distribution
- Relate to an issue



universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
stran	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824


Topic Models

- Topic modeling
 - Get topics automatically from a corpus
 - Assign documents to topics automatically
- Most frequently used topic models
 - pLSA
 - LDA

“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.

Text Data: Topic Models

- Text Data and Topic Models
- Multinomial Mixture Model 
- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA)
- Summary

Recap of Multinomial Distribution

- Select n data points from K categories, each with probability p_k
 - n trials of independent categorical distribution
 - E.g., get 1-6 from a dice with $1/6$
 - Let x_k be the number of times value k has been observed, note $\sum_k x_k = n$
 - $P(X_1 = x_1, X_2 = x_2, \dots, X_K = x_K) = \frac{n!}{x_1!x_2!\dots x_K!} \prod_k p_k^{x_k}$
- When $K=2$, binomial distribution
 - n trials of independent Bernoulli distribution
 - E.g., flip a coin to get heads or tails



Multinomial Mixture Model

- For documents with bag-of-words representation
 - $\mathbf{x}_d = (x_{d1}, x_{d2}, \dots, x_{dN})$, x_{dn} is the number of words for nth word in the vocabulary
- Generative model
 - For each document
 - Sample its cluster label $z \sim \text{Categorical}(\boldsymbol{\pi})$
 - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of jth cluster
 - $p(z = k) = \pi_k$
 - Sample its word vector $\mathbf{x}_d \sim \text{multinomial}(\boldsymbol{\beta}_z)$
 - $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$, β_{zn} is the parameter associate with nth word in the vocabulary
 - $p(\mathbf{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$

Likelihood Function

- For a set of M documents

$$\begin{aligned} L &= \prod_d p(\mathbf{x}_d) = \prod_d \sum_k p(\mathbf{x}_d, z = k) \\ &= \prod_d \sum_k p(\mathbf{x}_d | z = k) p(z = k) \\ &= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}} \end{aligned}$$

Mixture of Unigrams

- For documents represented by a sequence of words
 - $\mathbf{w}_d = (w_{d1}, w_{d2}, \dots, w_{dN_d})$, N_d is the length of document d , w_{di} is the word at the i th position of the document
- Generative model
 - For each document
 - Sample its cluster label $z \sim \text{Categorical}(\boldsymbol{\pi})$
 - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of k th cluster
 - $p(z = k) = \pi_k$
 - For each word in the sequence
 - Sample the word $w_{di} \sim \text{Categorical}(\boldsymbol{\beta}_z)$
 - $p(w_{di} | z = k) = \beta_{kw_{di}}$

Likelihood Function

- For a set of M documents

$$\begin{aligned} L &= \prod_d p(\mathbf{w}_d) = \prod_d \sum_k p(\mathbf{w}_d, z = k) \\ &= \prod_d \sum_k p(\mathbf{w}_d | z = k) p(z = k) \\ &= \prod_d \sum_k p(z = k) \prod_i \beta_{kw_{di}} \end{aligned}$$


Question

- Are multinomial mixture model and mixture of unigrams model equivalent? Why?

Question

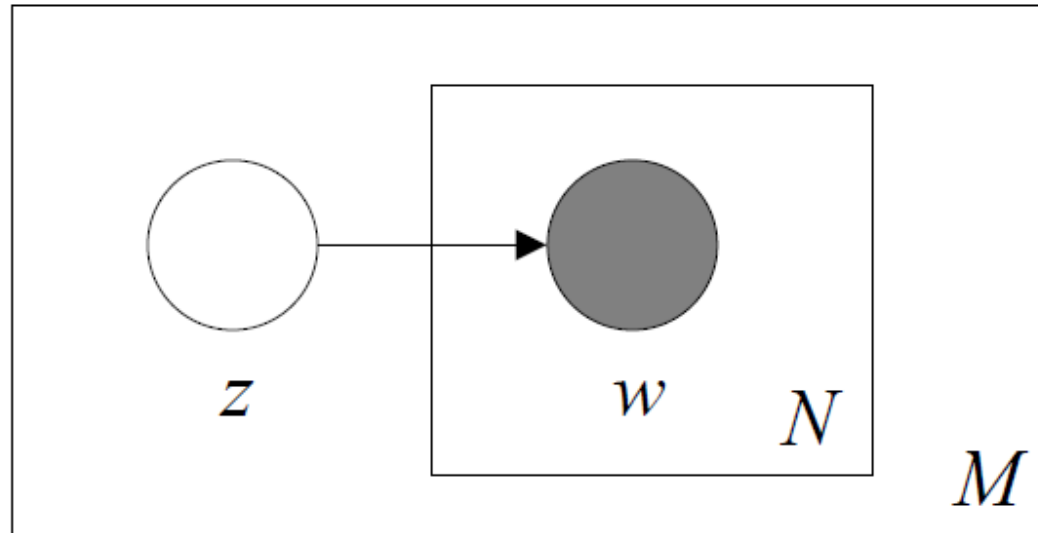
- How to conduct EM algorithm on mixture of unigrams?

Text Data: Topic Models

- Text Data and Topic Models
- Multinomial Mixture Model
- Probabilistic Latent Semantic Analysis (pLSA) 
- Latent Dirichlet Allocation (LDA)
- Summary

Issues of Mixture of Unigrams

- All the words in the same documents are sampled from the same topic



- In practice, people switch topics during their writing

Illustration of pLSA

“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.

Notations

- Word, document, topic
 - w, d, z
- Word count in document
 - $c(w, d)$
- Word distribution for each topic (β_z)
 - $\beta_{zw}: p(w|z)$
- Topic distribution for each document (θ_d)
 - $\theta_{dz}: p(z|d)$ (Yes, soft clustering)

Generative Model for pLSA

- Describe how a document d is generated probabilistically

- For each position in d , $n = 1, \dots, N_d$

- Generate the topic for the position as

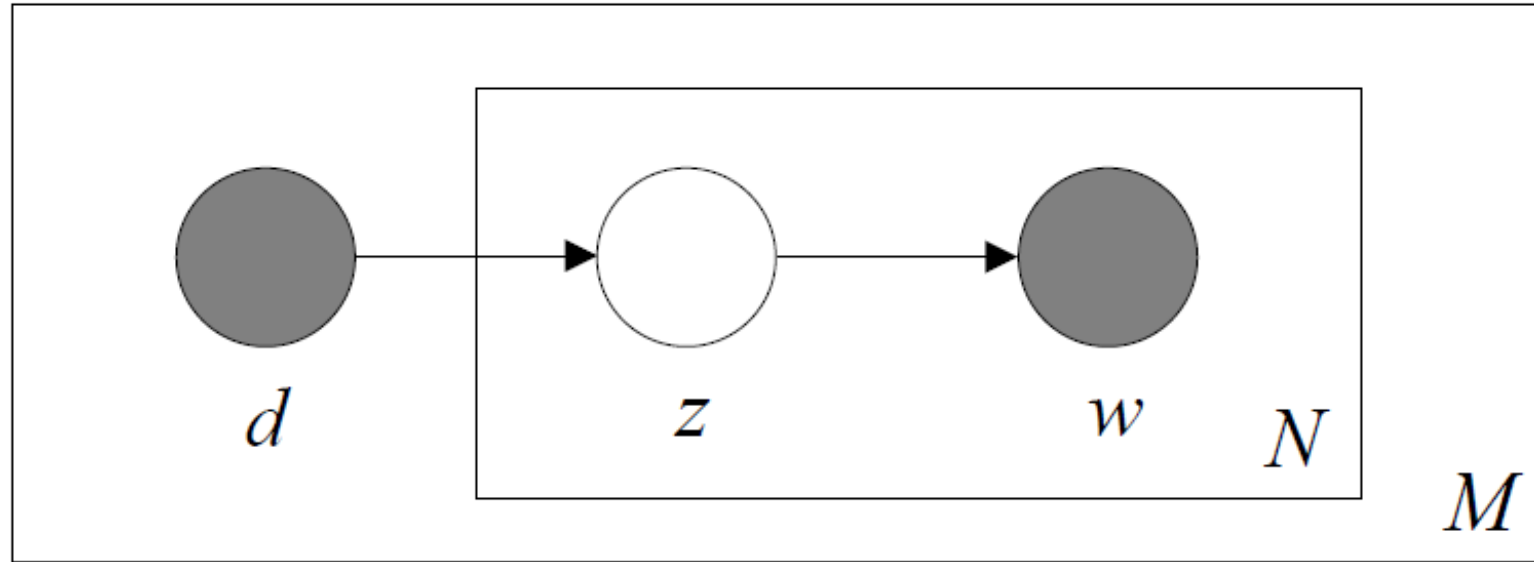
$$z_n | d \sim \text{Categorical}(\boldsymbol{\theta}_d), \text{ i. e., } p(z_n = k | d) = \theta_{dk}$$

(Note, 1 trial multinomial)

- Generate the word for the position as

$$w_n | z_n \sim \text{Categorical}(\boldsymbol{\beta}_{z_n}), \text{ i. e., } p(w_n = w | z_n) = \beta_{z_n w}$$

Graphical Model



Note: Sometimes, people add parameters such as θ and β into the graphical model

The Likelihood Function for a Corpus

- Probability of a word w

$$p(w|d, \theta, \beta) = \sum_k p(w, z = k|d, \theta, \beta) = \sum_k p(w|z = k, d, \theta, \beta)p(z = k|d, \theta, \beta) = \sum_k \beta_{kw}\theta_{dk}$$

- Likelihood of a corpus

$$\begin{aligned} & \prod_{d=1} P(w_1, \dots, w_{N_d}, d|\theta, \beta, \pi) \\ &= \prod_{d=1} P(d) \left\{ \prod_{n=1}^{N_d} \left(\sum_k P(z_n = k|d, \theta_d)P(w_n|\beta_k) \right) \right\} \\ &= \prod_{d=1} \pi_d \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk}\beta_{kw_n} \right) \right\} \end{aligned}$$

π_d is usually considered as uniform, i.e., $1/M$

Re-arrange the Likelihood Function

- Group the same word from different positions together

$$\max \log L = \sum_{dw} c(w, d) \log \sum_z \theta_{dz} \beta_{zw}$$

$$s. t. \sum_z \theta_{dz} = 1 \text{ and } \sum_w \beta_{zw} = 1$$

Optimization: EM Algorithm

- Repeat until converge
 - E-step: for each word in each document, calculate its conditional probability belonging to each topic

$$p(z|w, d) \propto p(w|z, d)p(z|d) = \beta_{zw}\theta_{dz} \text{ (i. e., } p(z|w, d) = \frac{\beta_{zw}\theta_{dz}}{\sum_{z'} \beta_{z'w}\theta_{dz'}})$$

- M-step: given the conditional distribution, find the parameters that can maximize the expected complete log-likelihood

$$\sum_{dw} c(w, d) \sum_z p(z|w, d) \log(\theta_{dz} \beta_{zw})$$

$$\beta_{zw} \propto \sum_d p(z|w, d) c(w, d) \text{ (i. e., } \beta_{zw} = \frac{\sum_d p(z|w, d) c(w, d)}{\sum_{w', d} p(z|w', d) c(w', d)})$$

$$\theta_{dz} \propto \sum_w p(z|w, d) c(w, d) \text{ (i. e., } \theta_{dz} = \frac{\sum_w p(z|w, d) c(w, d)}{N_d})$$

Comparison to Naïve Bayes

- Naïve Bayes

- $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}}$

- $\sum_{d:y_d=j} x_{dn}$: total count of word n in class j

- $\sum_{d:y_d=j} \sum_{n'} x_{dn'}$: total count of words in class j

- pLSA (M-step)

- $\beta_{zw} = \frac{\sum_d p(z|w, d)c(w, d)}{\sum_{w', d} p(z|w', d)c(w', d)}$

Example

- Two documents, two topics

- Vocabulary: {1: data, 2: mining, 3: frequent, 4: pattern, 5: web, 6: information, 7: retrieval}

- A

word (w)	word count in Document 1 ($c(w, d_1)$)	$p(z = 1 w, d_1)$
data	5	0.8
mining	4	0.8
frequent	3	0.6
pattern	2	0.8
web	2	0.5
information	1	0.2

word (w)	word count in Document 2 ($c(w, d_2)$)	$p(z = 1 w, d_2)$
information	5	0.2
retrieval	4	0.2
web	3	0.1
mining	3	0.5
frequent	2	0.6
data	2	0.5

Example (Continued)

- M-step

$$\beta_{11} = \frac{0.8 * 5 + 0.5 * 2}{11.8 + 5.8} = 5/17.6$$

$$\beta_{12} = \frac{0.8 * 4 + 0.5 * 3}{11.8 + 5.8} = 4.7/17.6$$

$$\beta_{13} = 3/17.6$$

$$\beta_{14} = 1.6/17.6$$

$$\beta_{15} = 1.3/17.6$$

$$\beta_{16} = 1.2/17.6$$

$$\beta_{17} = 0.8/17.6$$

$$\theta_{11} = \frac{11.8}{17}$$


$$\theta_{12} = \frac{5.2}{17}$$

Also, $\beta_{21}, \beta_{22}, \dots, \beta_{27}; \theta_{21}, \theta_{22}$

Question

- For the same word in different positions in a document, do they have the same conditional probability $p(z|w, d)$?

Text Data: Topic Models

- Text Data and Topic Models
- Multinomial Mixture Model
- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA) 
- Summary

Limitations of pLSA

- Not a proper generative model
 - θ_d is treated as a parameter
 - Cannot model new documents
- Solution:
 - Make it a proper generative model by adding priors to θ and β

Review of Conjugate Prior

- Model:
 - $p(x|\theta)$
- Prior:
 - $p(\theta|\alpha)$
- Posterior:
 - $p(\theta|x, \alpha) \propto p(\theta, x|\alpha) = p(x|\theta)p(\theta|\alpha)$
- Conjugate prior:
 - If $p(\theta|\alpha)$ and $p(\theta|x, \alpha)$ belong to the same distribution family (with different parameters), $p(\theta|\alpha)$ is called a conjugate prior for the likelihood function

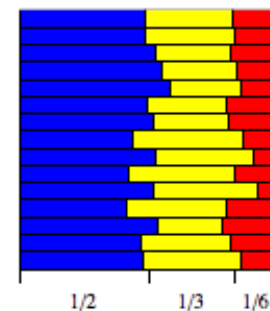
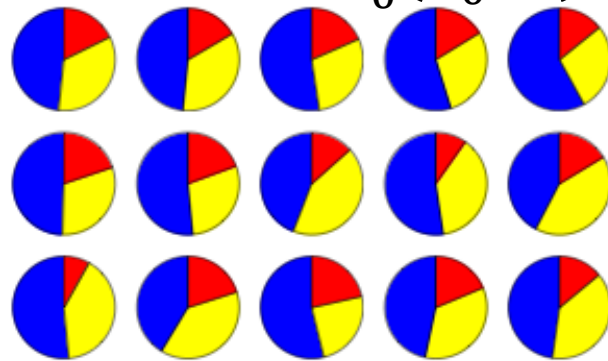
Dirichlet Distribution

- Dirichlet distribution: $\theta \sim \text{Dirichlet}(\alpha)$

- *i. e.*, $p(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, where $\alpha_k > 0$

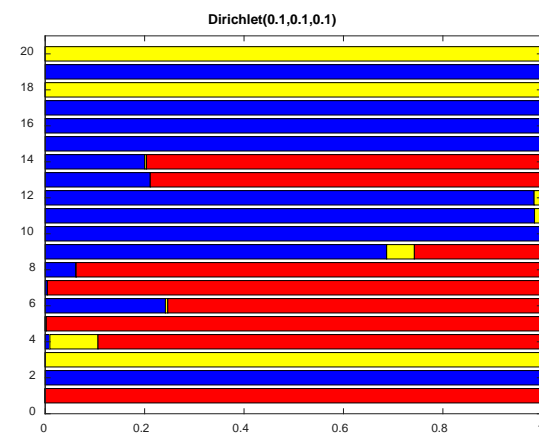
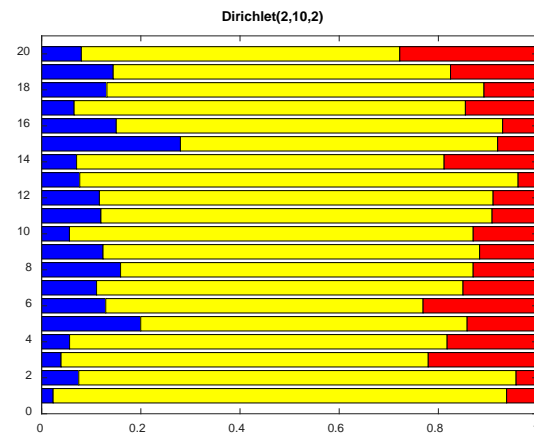
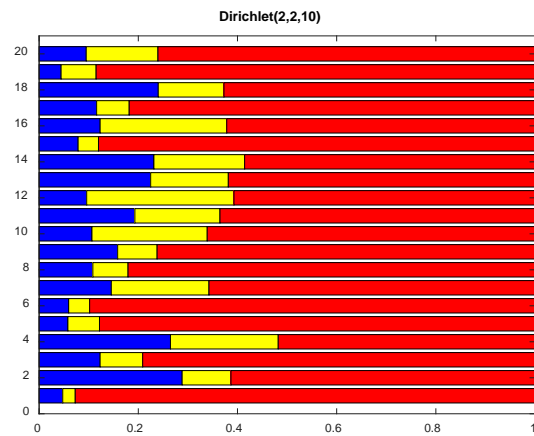
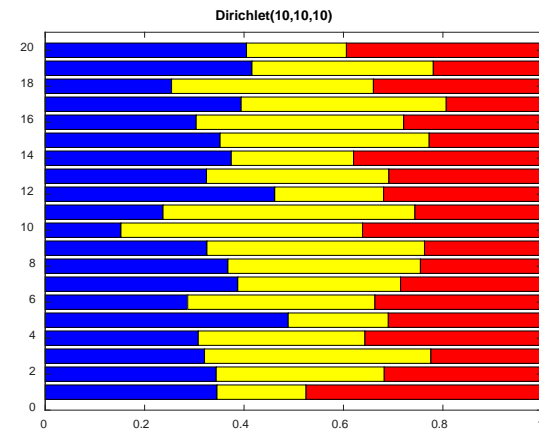
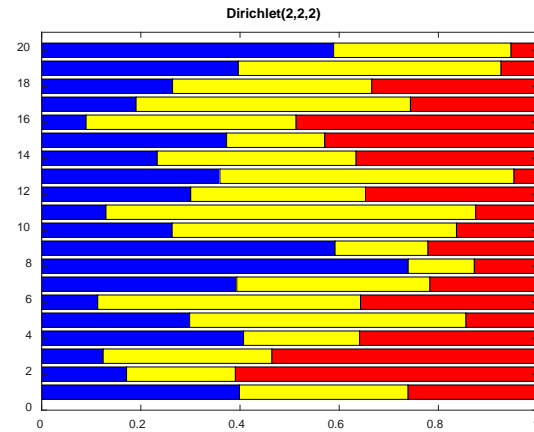
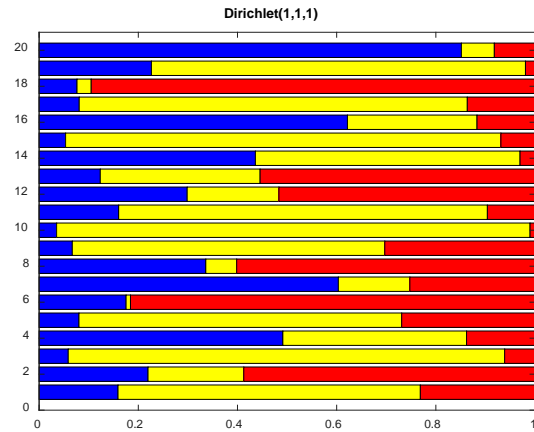
- $\Gamma(\cdot)$ is gamma function: $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$
 - $\Gamma(z + 1) = z\Gamma(z)$

- $E(\theta_k) = \frac{\alpha_k}{\sum_{k'} \alpha_{k'}}$, $Var(\theta_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$, where $\alpha_0 = \sum_k \alpha_k$



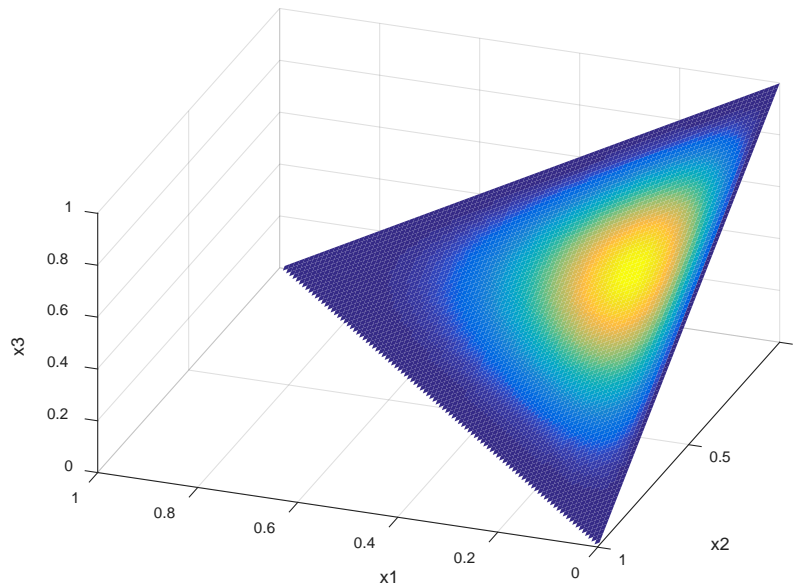
Example: $\theta \sim \text{Dirichlet}(\alpha)$, where $\alpha/\alpha_0 = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$

More Examples



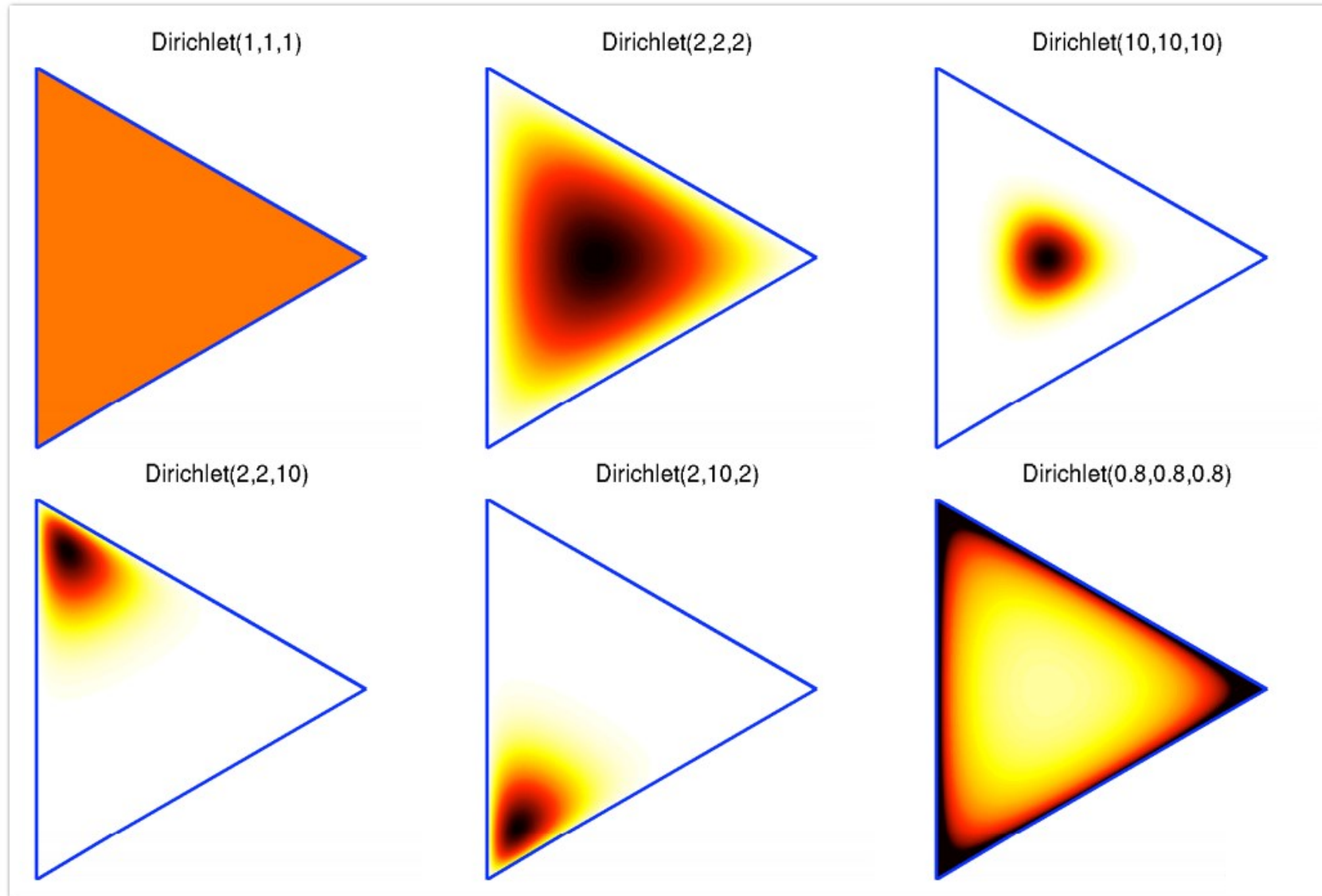
Simplex View

- Simplex:
 - “a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions.”
 - $x = x_1(1,0,0) + x_2(0,1,0) + x_3(0,0,1)$
 - Where $0 \leq x_1, x_2, x_3 \leq 1$ and $x_1 + x_2 + x_3 = 1$



$$x|\alpha \sim \text{Dir}(\alpha), \alpha = (2,3,4)$$

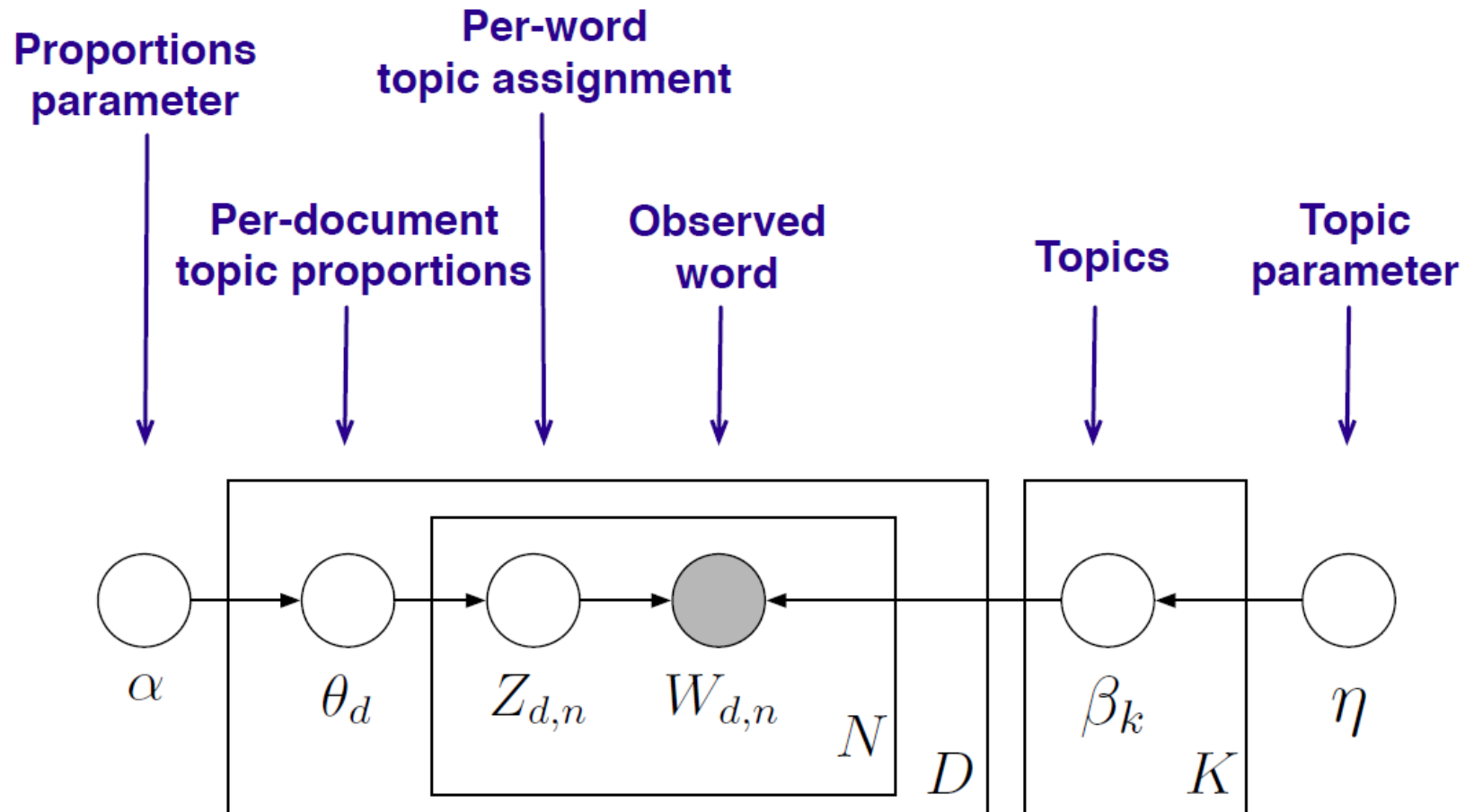
More Examples in the Simplex View



Dirichlet-Multinomial/Categorical Conjugate

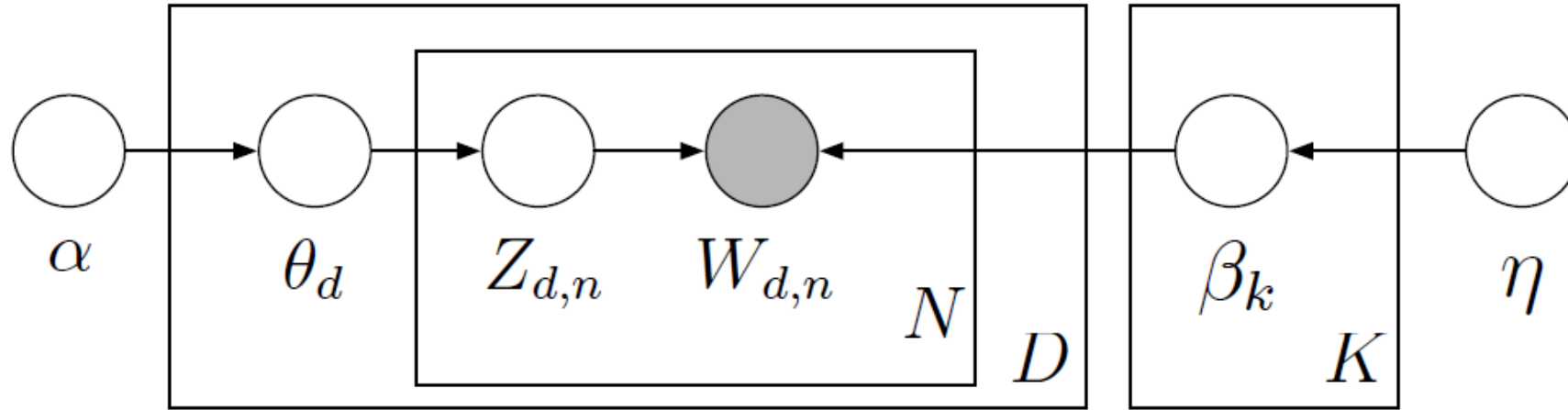
- The model (independent categorical distribution)
 - $p(z_1, \dots, z_n | \boldsymbol{\theta}) = \prod_k \theta_k^{c_k}$
 - where $z_i \in \{1, \dots, K\}$, and c_k is the number of z that takes value of k
- Posterior
 - $p(\boldsymbol{\theta} | \boldsymbol{\alpha}, z_1, z_2, \dots, z_n) \propto p(z_1, \dots, z_n | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\alpha})$
 $\propto \prod_k \theta_k^{\alpha_k + c_k - 1}$
 - Dirichlet distribution again!

The Graphical Model of LDA



$\theta_d \sim \text{Dirichlet}(\alpha)$: address topic distribution for unseen documents
 $\beta_k \sim \text{Dirichlet}(\eta)$: smoothing over words

Joint Distribution for LDA



- Joint distribution of latent variables and documents is:

$$p(\boldsymbol{\beta}_{1:K}, \mathbf{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \mathbf{w}_{1:D} | \alpha, \eta) =$$

$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

Posterior Inference

- Posterior of the latent variables

$$p(\beta, \theta, \mathbf{z} | \mathbf{w})$$
$$= \frac{p(\beta, \theta, \mathbf{z}, \mathbf{w})}{\int_{\beta} \int_{\theta} \sum_{\mathbf{z}} p(\beta, \theta, \mathbf{z}, \mathbf{w})}$$

- Solutions


marginal $p(\mathbf{w})$ is intractable!

- Gibbs sampling [Griffiths et al., Finding Scientific Topics, PNAS (2004)]
- Variational inference [Blei et al., Latent Dirichlet Allocation, JLMR (2003)]

Parameter Estimation

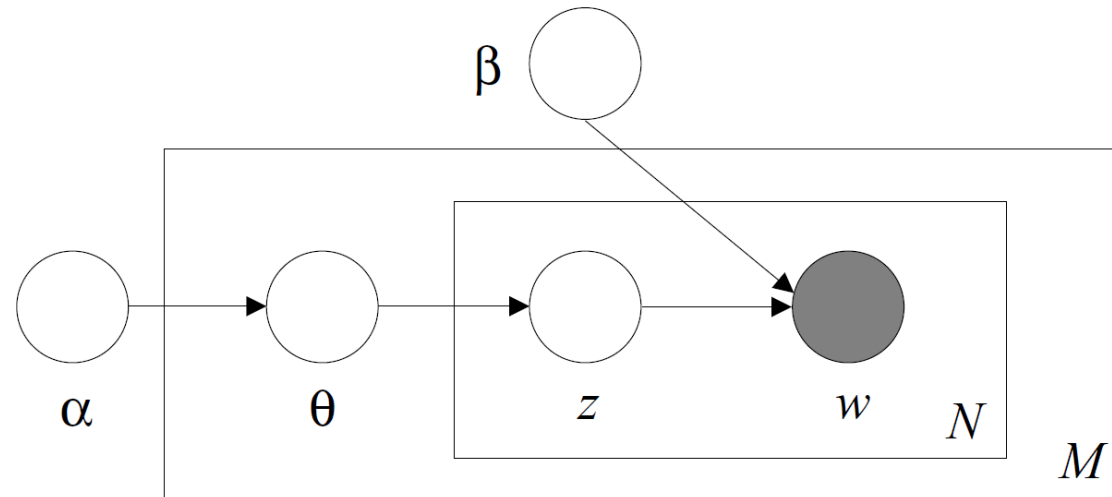
- Estimate α and η based on training dataset
- Solution
 - Variational EM algorithm

Generative Process of A Simplified LDA

- Assuming there is no prior for β
 - I.e., Given α, β
 - For each document \mathbf{w}_d in a corpus
 - 1. Choose document length $N_d \sim \text{Poisson}(\xi)$
 - 2. Choose $\theta_d \sim \text{Dir}(\alpha)$
 - 3. For each word in the document $w_{d,n}$:
 - a) Choose a topic $z_{d,n} \sim \text{Categorical}(\theta_d)$
 - b) Choose a word $w_{d,n}$ from $p(w_{d,n} | z_{d,n}, \beta)$, which is a categorical distribution conditional on topic $z_{d,n}$
- Not essential, and can be ignored*
- 

Graphical Model and Joint Distribution

- Graphical model



- Joint distribution for a single document

- $$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

Marginal Distribution and Likelihood

- Marginal distribution of a document

- $$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta$$

- Question: Why can we push summation of z inside?

- Likelihood of a corpus

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d$$

Inference

- Compute the posterior distribution of the hidden variables given a document

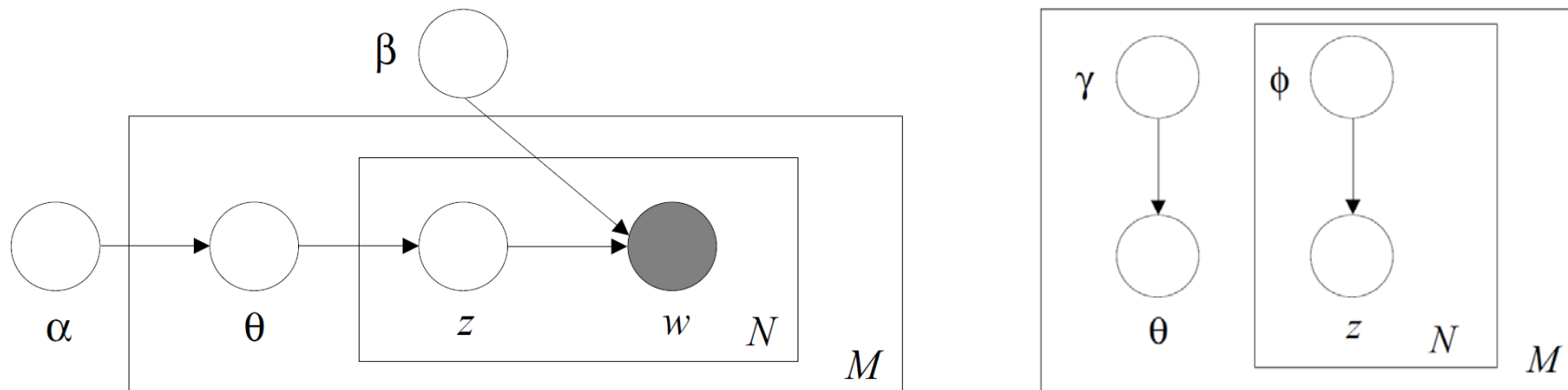
Joint distribution computation: Easy

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}$$

Marginal distribution computation: intractable

Variational Inference

- Find a simple distribution, called **variational distribution**, $q(\theta, \mathbf{z}|\gamma, \phi)$ to approximate $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$
 - By dropping edges among θ, z , and w , which causes the coupling between θ and β



- $q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^N q(z_n|\phi_n)$
 - γ, ϕ are **free** variational parameters

More about the variational distribution

- $q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n)$
- $q(\theta | \gamma)$
 - θ follows Dirichlet distribution with parameter γ
- $q(z_n | \phi_n)$
 - z_n follows categorical distribution with parameter ϕ_n

The Criterion to Choose q

- KL divergence between two distributions

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- Minimize the KL divergence between $q(\theta, \mathbf{z}|\gamma, \phi)$ and $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$
 - $D(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) = E_q(\log q(\theta, \mathbf{z}|\gamma, \phi) - \log p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$

$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, \mathbf{z}|\gamma, \phi) || p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

ELBO: Evidence Lower Bound

- Instead of minimizing KL divergence directly, maximize ELBO
- Can be obtained by applying Jensen's inequality: $f(E(X)) \geq E(f(X))$, if f is concave

$$\begin{aligned}\log p(\mathbf{w} | \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta \\ &= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\ &\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta - \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\ &= \underbrace{E_q[\log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})]}_{\text{ELBO, denoted as } L(\gamma, \phi; \alpha, \beta)}.\end{aligned}$$

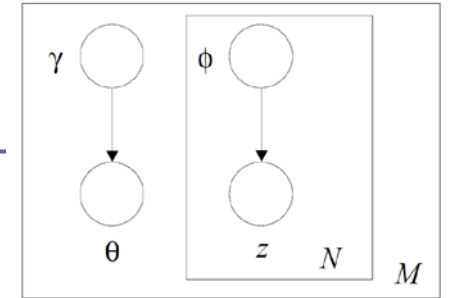
ELBO, denoted as $L(\gamma, \phi; \alpha, \beta)$

Why?

- ELBO + KL-divergence = Constant $L(\gamma, \phi; \alpha, \beta) = \mathbb{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)] - \mathbb{E}_q[\log q(\theta, \mathbf{z})]$.

$$\log p(\mathbf{w} | \alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + D(q(\theta, \mathbf{z} | \gamma, \phi) || p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta))$$

ELBO Optimization



- Coordinate ascent

- Iteratively optimizing each variational parameter holding the others fixed

- (1) initialize $\phi_{ni}^0 := 1/k$ for all i and n
- (2) initialize $\gamma_i := \alpha_i + N/k$ for all i
- (3) **repeat**
- (4) **for** $n = 1$ **to** N
- (5) **for** $i = 1$ **to** k **$\Psi(\mathbf{x})$: first derivative of $\log\Gamma(\mathbf{x})$**
- (6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
- (7) normalize ϕ_n^{t+1} to sum to 1.
- (8) $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi_n^{t+1}$
- (9) **until** convergence

Parameter Estimation

- Objective: Find α and β that can maximize log-likelihood function with constraints

- $$\ell(\alpha, \beta) = \sum_{d=1}^M \log p(\mathbf{w}_d | \alpha, \beta)$$

- Solution: Variational EM algorithm

- E-step: find γ_d^* and ϕ_d^* for each document by maximizing the lower bound of the likelihood function (Done!)
- M-step: maximizing the lower bound obtained from E-step with respect to α and β

$\Psi(\mathbf{x})$: first derivative of $\log\Gamma(\mathbf{x})$


$$\frac{\partial L}{\partial \alpha_i} = M(\Psi(\sum_{j=1}^k \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^M (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}))$$

$$\beta_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j \quad \mathbf{w}_{dn}^j: \mathbf{1}\{\mathbf{w}_{dn} == \mathbf{j}\}$$

Question

- Connection of variational EM and EM algorithm
 - What is the KL divergence between $q(z|x, \theta)$ and the true posterior $p(z|x, \theta)$ in EM algorithm, e.g., in GMM? (Hint: how did we choose $q(z|x, \theta)$?)

Text Data: Topic Models

- Text Data and Topic Models
- Revisit of Multinomial Mixture Model
- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA)
- Summary 

Summary

- Basic Concepts
 - Word/term, document, corpus, topic
 - How to represent a document
- Mixture of unigrams
- pLSA
 - Generative model
 - Likelihood function
 - EM algorithm
- LDA
 - Dirichlet-multinomial conjugate
 - Posterior inference

References

- T. Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.
- David Blei, lecture notes on variational inference:
<https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf>
- David Blei, Andrew Ng, Michael Jordan, “Latent Dirichlet Allocation”, JMLR, 2003.
<http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Thomas L. Griffiths and Mark Steyvers, “Finding Scientific Topics”, PNAS, 2004.
https://www.pnas.org/content/101/suppl_1/5228