DSC250: Advanced Data Mining

Topic Models

Zhiting Hu Lecture 7, Jan 28, 2025



HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Topic models: v1, v2, v3
- Paper Presentations:
 - (1) Liyuan Jin, Riqian Hu: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism
 - (2) Victoria Jin, Wenqi Li: Large Language Models Are Human-Level Prompt Engineers

Recap: Represent a Document

• Most common way: Bag-of-Words

- Ignore the order of words
- keep the count
- c1: Human machine interface for Lab ABC computer applications
- c2: A survey of user opinion of computer system response time
- c3: The EPS user interface management system
- c4: System and human system engineering testing of EPS
- c5: Relation of user-perceived response time to error measurement
- m1: The generation of random, binary, unordered trees
- m2: The intersection graph of paths in trees
- m3: Graph minors IV: Widths of trees and well-quasi-ordering
- m4: Graph minors: A survey

		CT	C2	63	C4	C5	mı	m2	m3	m4
	human	1	0	0	1	0	0	0	0	0
	interface	1	0	1	0	0	0	0	0	0
	computer	1	1	0	0	0	0	0	0	0
	user	0	1	1	0	1	0	0	0	0
	system	0	1	1	2	0	0	0	0	0
7	response	0	1	0	0	1	0	0	0	0
	time	0	1	0	0	1	0	0	0	0
	EPS	0	0	1	1	0	0	0	0	0
	survey	0	1	0	0	0	0	0	0	1
	trees	0	0	0	0	0	1	1	1	0
	graph	0	0	0	0	0	0	1	1	1
	minors	0	0	0	0	0	0	0	1	1

c1

Vector space model

 m^{2}

Recap: Represent a Topic

• A topic is represented by a word distribution

• Relate to an issue

0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824
	0.0375 0.0279 0.0233 0.0232 0.0214 0.0137 0.0131 0.0109 0.01 0.0983 0.0561 0.0431 0.0381 0.025 0.0214 0.0196 0.0165 0.0163	0.0375 patients 0.0279 drugs 0.0233 clinical 0.0232 treatment 0.0214 trials 0.0137 therapy 0.0137 therapy 0.0137 therapy 0.0131 trial 0.0109 disease 0.01 medical 0.0561 females 0.0431 females 0.025 sex 0.025 sex 0.0196 offspring 0.0165 sexual	0.0375 patients 0.0493 0.0279 drugs 0.0444 0.0233 clinical 0.0346 0.0232 treatment 0.028 0.0214 trials 0.0277 0.0137 therapy 0.0213 0.0137 therapy 0.0213 0.0137 therapy 0.0213 0.0137 therapy 0.0213 0.0131 trial 0.0164 0.0109 disease 0.0157 0.01 medical 0.00997 0.0983 male 0.0558 0.0561 females 0.0541 0.0431 females 0.0477 0.025 sex 0.0339 0.0214 reproductive 0.0172 0.0196 offspring 0.0168 0.0165 sexual 0.0166 0.0163 reproduction 0.0143	0.0439 patients 0.0493 stem 0.0279 drugs 0.0444 human 0.0233 clinical 0.0346 cell 0.0233 clinical 0.0346 cell 0.0232 treatment 0.028 gene 0.0214 trials 0.0277 tissue 0.0137 therapy 0.0213 cloning 0.0137 therapy 0.0213 cloning 0.0131 trial 0.0164 transfer 0.0109 disease 0.0157 blood 0.01 medical 0.00997 embryos 0.0561 females 0.0541 physics 0.0431 female 0.0529 physicists 0.0381 males 0.0477 einstein 0.025 sex 0.0339 university 0.025 sex 0.0339 university 0.0164 reproductive 0.0172 gravity 0.0165 sexual 0.	0.0435 patients 0.0493 stem 0.0478 0.0279 drugs 0.0444 human 0.0421 0.0233 clinical 0.0346 cell 0.0309 0.0232 treatment 0.028 gene 0.025 0.0214 trials 0.0277 tissue 0.0185 0.0137 therapy 0.0213 cloning 0.0169 0.0137 therapy 0.0213 cloning 0.0155 0.0137 therapy 0.0213 cloning 0.0169 0.0131 trial 0.0164 transfer 0.0155 0.0109 disease 0.0157 blood 0.0113 0.01 medical 0.00997 embryos 0.0111 0.0983 male 0.0529 physicists 0.0146 0.0381 males 0.0477 einstein 0.0142 0.025 sex 0.0339 university 0.013 0.0214 reproductive 0.0172	0.0375patients 0.0493 stem 0.0478 sequences 0.0279 drugs 0.0444 human 0.0421 genome 0.0233 clinical 0.0346 cell 0.0309 dna 0.0232 treatment 0.028 gene 0.025 sequencing 0.0214 trials 0.0277 tissue 0.0185 map 0.0137 therapy 0.0213 cloning 0.0169 genes 0.0131 trial 0.0164 transfer 0.0155 chromosome 0.0109 disease 0.0157 blood 0.0111 regions 0.0983 male 0.0558 theory 0.0811 immune 0.0431 females 0.0477 theory 0.0811 immune 0.025 sex 0.0375 university 0.0142 system 0.025 sex 0.0375 university 0.013 antigen 0.025 sex 0.0477 gravity 0.013 antigen 0.025 sex 0.0339 university 0.013 antigen 0.025 sex 0.0379 university 0.013 antigen 0.0165 sexual 0.0166 theories 0.01 immunology 0.0163 reproduction 0.0143 aps 0.00987 antibody	0.0375patients 0.0493 stem 0.0478 sequences 0.0493 0.0279 drugs 0.0444 human 0.0421 genome 0.033 0.0233 clinical 0.0346 cell 0.0309 dna 0.0257 0.0232 treatment 0.028 gene 0.025 sequencing 0.0172 0.0214 trials 0.0277 tissue 0.0185 map 0.0123 0.0137 therapy 0.0213 cloning 0.0169 genes 0.0122 0.0131 trial 0.0164 transfer 0.0155 chromosome 0.0119 0.0109 disease 0.0157 blood 0.0113 regions 0.0119 0.0983 male 0.0558 theory 0.0811 immune 0.0909 0.0431 females 0.0477 physics 0.0782 system 0.0358 0.025 sex 0.0339 university 0.013 antigen 0.0263 0.025 sex 0.0379 gravity 0.013 antigen 0.0263 0.025 sex 0.0339 gravity 0.013 antigen 0.0263 0.025 sexual 0.0168 black 0.0127 immunity 0.0145 0.0165 sexual 0.0166 theories 0.01 immunity 0.0145 0.0163 reproduction 0.0143 aps 0.00987 antibody 0.014	0.0375 0.0279patients0.0493 0.0444stem0.0478 0.0444sequences0.0493 genomemillion0.0233clinical0.0346 0.0232cell0.0309 dnadna0.0257ago0.0232treatment0.028 treatmentgene0.025 tissuesequencing0.0172 genesage0.0137therapy0.0213 therapycloning0.0169 transfergenes0.0122 chromosomerecord0.0131trial0.0164 transfertransfer0.0155 thoryschromosome0.0119 tregionsearly billion0.0109disease0.0157 medical0.0558 0.0561theory0.0811 physicsimmune0.0909 responsestars system0.025sex0.0339 thoradtheory0.0142 transferimmune0.0909 transferstars system0.025sex0.0377 theradetheory0.0811 physicsimmune0.0909 transferstars system0.025sex0.0339 theoryuniversity0.013 transferantigen0.0263 disk0.025sex0.0339 theoriesuniversity0.013 theoriesantigens0.0184 theories0.025sex0.0166 theories0.0127 theoriesantigens0.0184 theoriestheory0.0163reproductive0.0172 theoriesgasantigens0.0184 theoriestheory

TOPIC 42

TOPIC 45

andicatory finance appendence categories COMMATCIAI chiese version balancropportunities los center. future mix shorter region housing rdated population^S CrVices and a structure main retail and infrastructure reside major heath spit growing distribution JOb tem Obs CTVIC CS focusing resident continue con coverse and a second se ives concentration

TOPIC 45 consists identified implementing annually implement created underway september in firmation ECOn Office

septembei "namman en even unic linated "proved including recently inated building program" ine band offerscource "bills" fall

ine bad offersource maise international and international and a second a

inded "out a include facility case astimated services in provement facilitate years improvement facilitate years total divisionerly protected

ili tylisted address

operated found operates work ingompletion

uctedia cibi

ovahused Machanitma Junda critinatar modernizationgkeilija. Proposed benefiti emeunt malyni interastritet indud fixal intera gradu, private Project fund pace need at Capita cost typenditures

ogram: term costs an ten an ce

mprovements

revenuprojects program inding tax financial list

TOPIC 48

TOPIC 43

IUric -s landgea pei antheologies antheol potential Droperties characterize ister mathematical properties characterize isteriation in the state of the second seco mattreevent of the second seco Hathelishndividualguide inter effettive Scheberhoods adaptive originalspectre entry profinitional integrity Ware next

i,

CO

dow

afasian Vanisod Cust Painia Pi

sintro:

u ction

sumo

appr

TOPIC 46

supplement locate soloway Dead listensitylocanea appendation family of pathology detension include Jestimbor for the substance papered ensurvision provide fightige (popular unity conserved to the substance of the pepase districting up an original and a second and a second and a second a munity keystudy village heigh/located adopted policies but and residents aime 11 figure SDecial buy highway g hi provide pha órain age

TOPIC 49

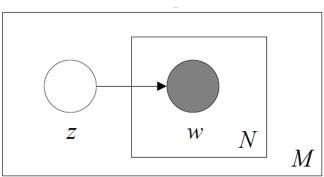
govenn requirer indivi prioriti es bringro ensuring recent ade \mathbf{Z}^{k}

Notations

- Word, document, topic
 - $\circ w, d, z$
- Word count in document:
 - $\circ c(w, d)$: number of times word w occurs in document d
 - \circ or x_{dn} : number of times the *n*th word in the vocabulary occurs in document d
- Word distribution for each topic (β_z)
 - $\beta_{zw}: p(w|z)$

Recap: Topic Model v1: Multinomial Mixture Model

Graphical Model



- Plates indicate replicated variables.
- Shaded nodes are observed; unshaded nodes are hidden.

- Generative model
 - For each document
 - Sample its cluster label $z \sim Categorical(\pi)$

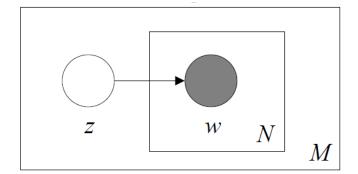
• $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K), \pi_k$ is the proportion of jth cluster

• $p(z=k) = \pi_k$

- Sample its word vector $\mathbf{x}_d \sim multinomial(\boldsymbol{\beta}_z)$
 - $\beta_z = (\beta_{z1}, \beta_{z2}, ..., \beta_{zN}), \beta_{zn}$ is the parameter associate with nth word in the vocabulary

•
$$p(\mathbf{x}_d|z=k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$$

Recap: Likelihood Function



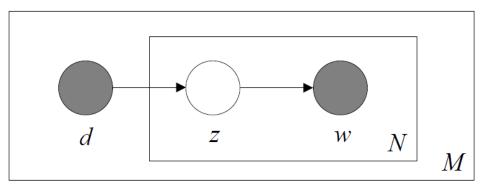
$$L = \prod_{d} p(\mathbf{x}_{d}) = \prod_{d} \sum_{k} p(\mathbf{x}_{d}, z = k)$$
$$= \prod_{d} \sum_{k} p(\mathbf{x}_{d} | z = k) p(z = k)$$
$$= \prod_{d} \frac{(\sum_{n} x_{dn})!}{\prod_{n} x_{dn}!} \sum_{k} p(z = k) \prod_{n} \beta_{kn}^{x_{dn}}$$

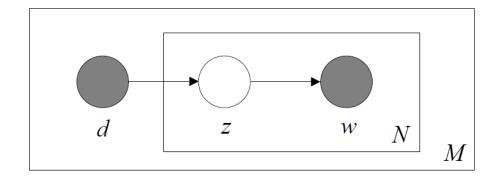
Recap: Topic Model v2: pLSA

- For each position in d, $n = 1, ..., N_d$
 - Generate the topic for the position as $z_n | d \sim Categorical(\theta_d), i.e., p(z_n = k | d) = \theta_{dk}$ (Note, 1 trial multinomial)
 - Generate the word for the position as

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



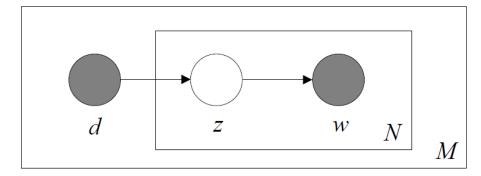




Likelihood FunctionProbability of a word w

 $p(w|d,\theta,\beta)$

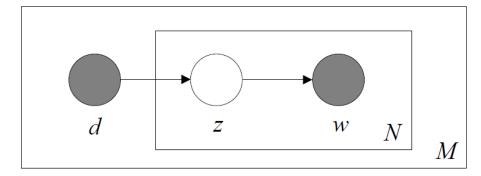
Likelihood Function



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 Function

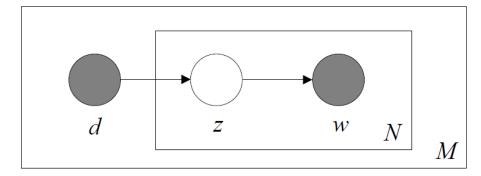


Probability of a word w

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

• Likelihood of a corpus

Likelihood Function



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

$$\prod_{d=1}^{n} P(w_1, \cdots, w_{N_d}, d | \theta, \beta, \pi)$$

$$= \prod_{d=1}^{n} P(d) \left\{ \prod_{n=1}^{N_d} \left(\sum_k P(z_n = k | d, \theta_d) P(w_n | \beta_k) \right) \right\}$$

$$= \prod_{d=1}^{n} \pi_d \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk} \beta_{kw_n} \right) \right\}$$

$$\pi_d \text{ is usually considered as uniform, i.e., 1/M}$$

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. $\sum_{z} \theta_{dz} = 1$ and $\sum_{w} \beta_{zw} = 1$

Limitations of pLSA

- Not a proper generative model
 - $\boldsymbol{\theta}_d$ is treated as a parameter
 - Cannot model new documents

• Solution:

• Make it a proper generative model by adding priors to θ and β

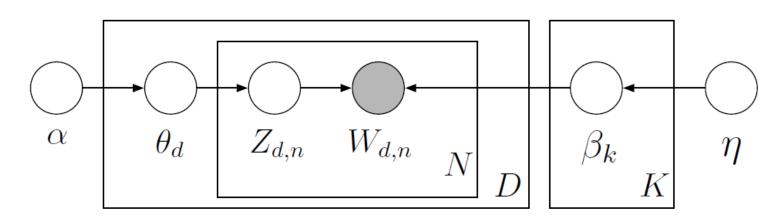
Limitations of pLSA

- Not a proper generative model
 - $\boldsymbol{\theta}_d$ is treated as a parameter
 - Cannot model new documents

- Solution:
 - Make it a proper generative model by adding priors to θ and β

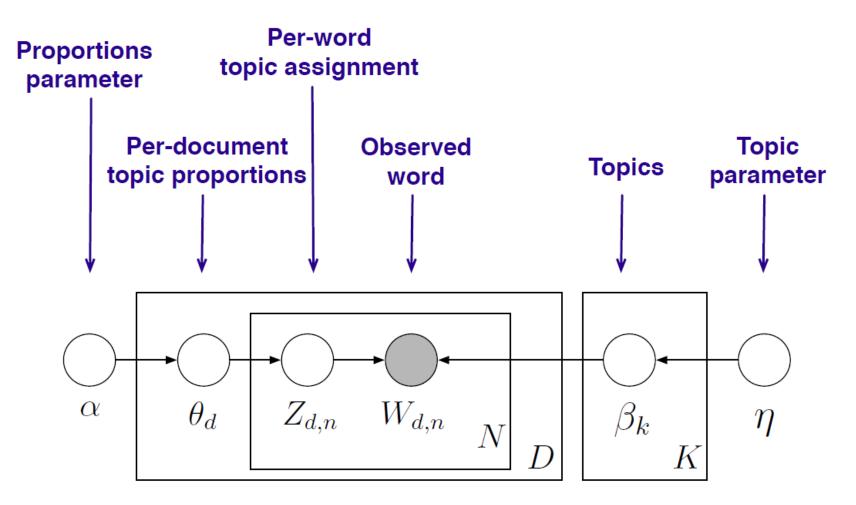
Topic Model v3: Latent Dirichlet Allocation (LDA)

Topic Model v3: Latent Dirichlet Allocation (LDA)



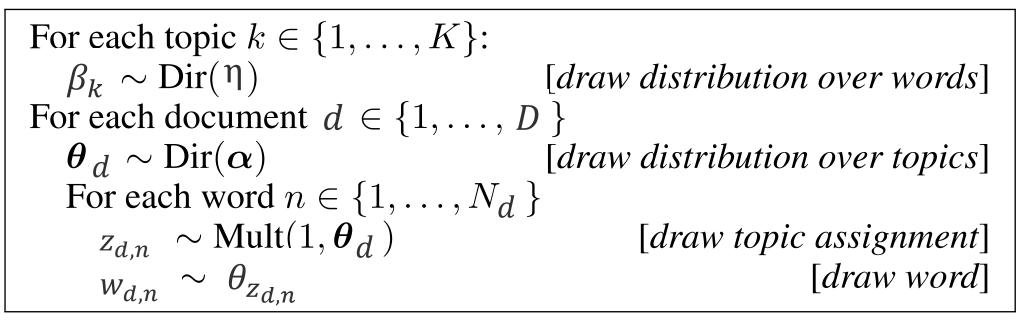
 $\theta_d \sim Dirichlet(\alpha)$: address topic distribution for unseen documents $\beta_k \sim Dirichlet(\eta)$: smoothing over words

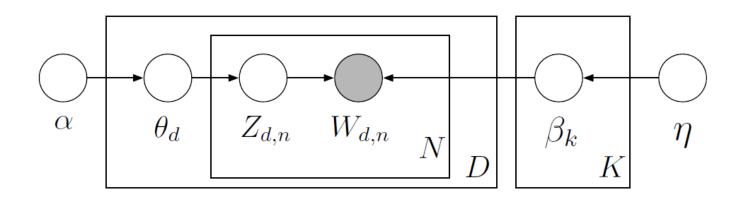
Topic Model v3: Latent Dirichlet Allocation (LDA)



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

Generative Model for LDA





Review: Dirichlet Distribution

• Dirichlet distribution: $\theta \sim Dirichlet(\alpha)$

• *i.e.*,
$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k} \alpha_{k})}{\prod_{k} \Gamma(\alpha_{k})} \prod_{k} \theta_{k}^{\alpha_{k}-1}$$
, where $\alpha_{k} > 0$
• $\Gamma(\cdot)$ is gamma function: $\Gamma(z) = \int_{0}^{\infty} e^{-t} t^{z-1} dt$
• $\Gamma(z+1) = z\Gamma(z)$

Review: Dirichlet Distribution

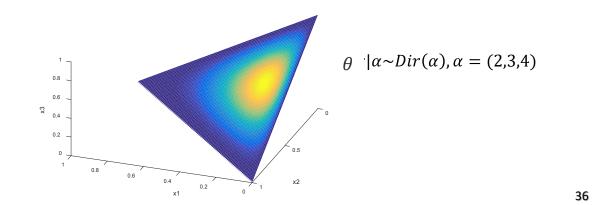
• Dirichlet distribution: $\theta \sim Dirichlet(\alpha)$

• *i.e.*,
$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k} \alpha_{k})}{\prod_{k} \Gamma(\alpha_{k})} \prod_{k} \theta_{k}^{\alpha_{k}-1}$$
, where $\alpha_{k} > 0$
• $\Gamma(\cdot)$ is gamma function: $\Gamma(z) = \int_{0}^{\infty} e^{-t} t^{z-1} dt$
• $\Gamma(z+1) = z\Gamma(z)$

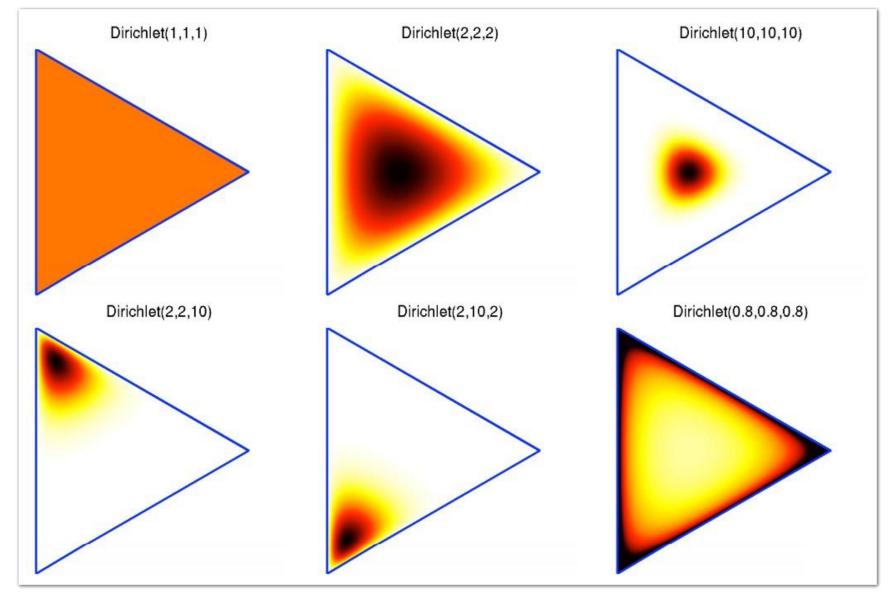
Simplex view:

 $\theta = \theta_1(1,0,0) + \theta_2(0,1,0) + \theta_3(0,0,1)$

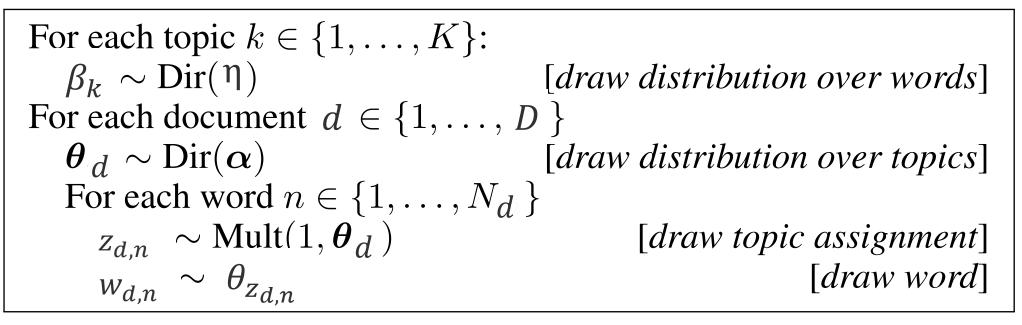
• Where $0 \leq \theta_1, \theta_2, \theta_3 \leq 1$ and $\theta_1 + \theta_2 + \theta_3 = 1$

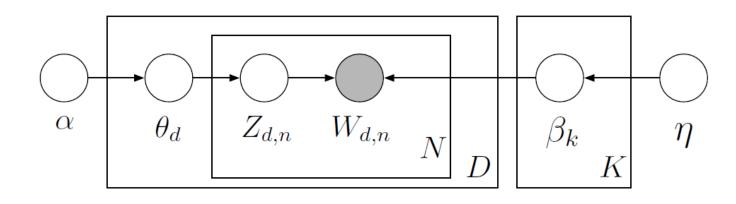


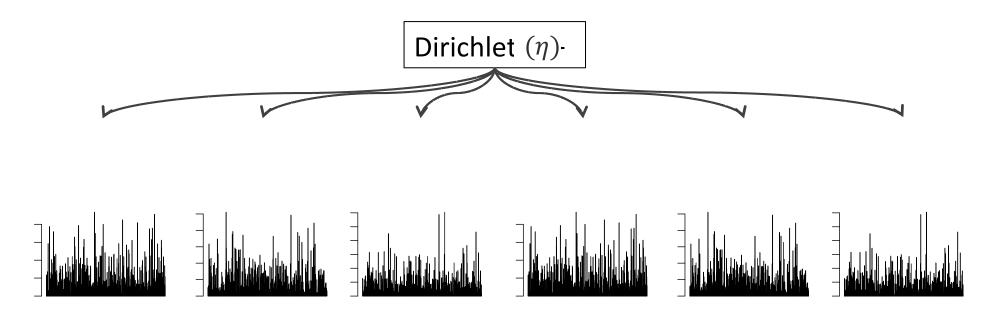
More Examples in the Simplex View



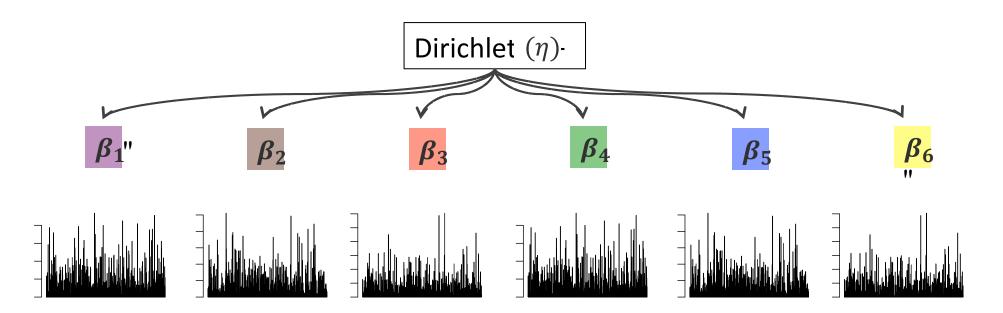
Generative Model for LDA





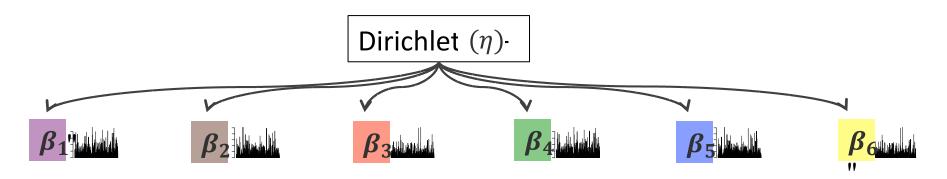


- The'generative&tory&egins'with'bnly'a'Dirichlet& prior&ver'the'topics."
- Each'topic&s'defind 'as'a'Multinomial&distribution'' over'the'vocabulary,'parameterized'by' β_k '



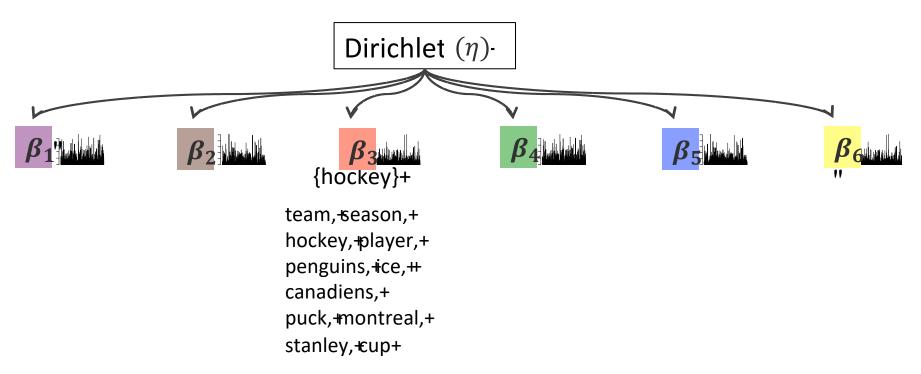
- The'generative&tory&egins'with'bnly'a'Dirichlet& prior&ver'the'topics."
- Each'topic&'defind 'as'a'Multinomial&istribution'' over'the'vocabulary,'parameterized'by' β_k '

LDA'for'Topic'Modeling"

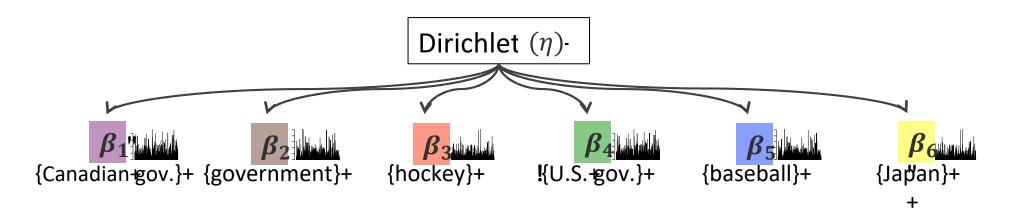


team, season, + hockey, player, + penguins, ice, ++ canadiens, + puck, +montreal, + stanley, +cup+

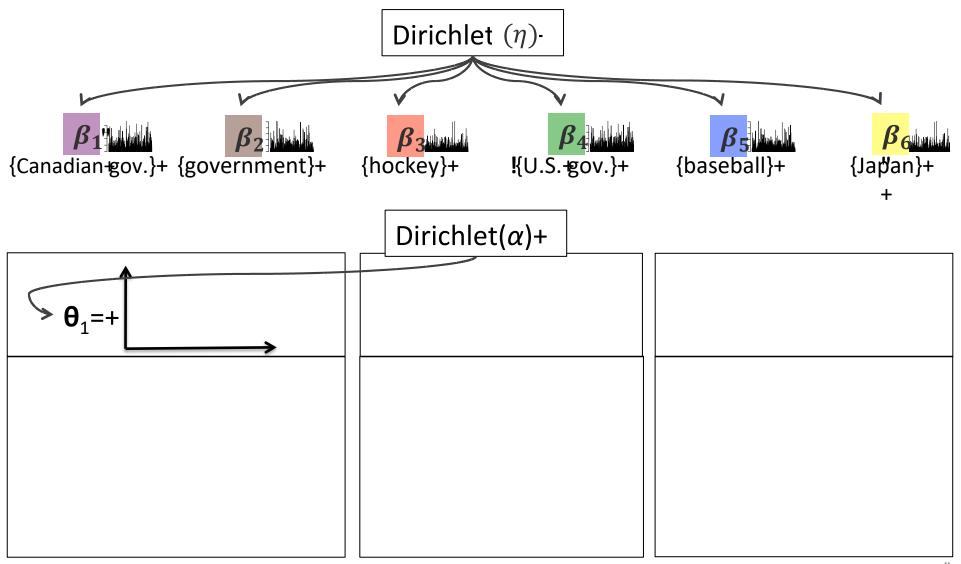
 A'topic'ls'visualized'as'lts'high&robability& words.'''

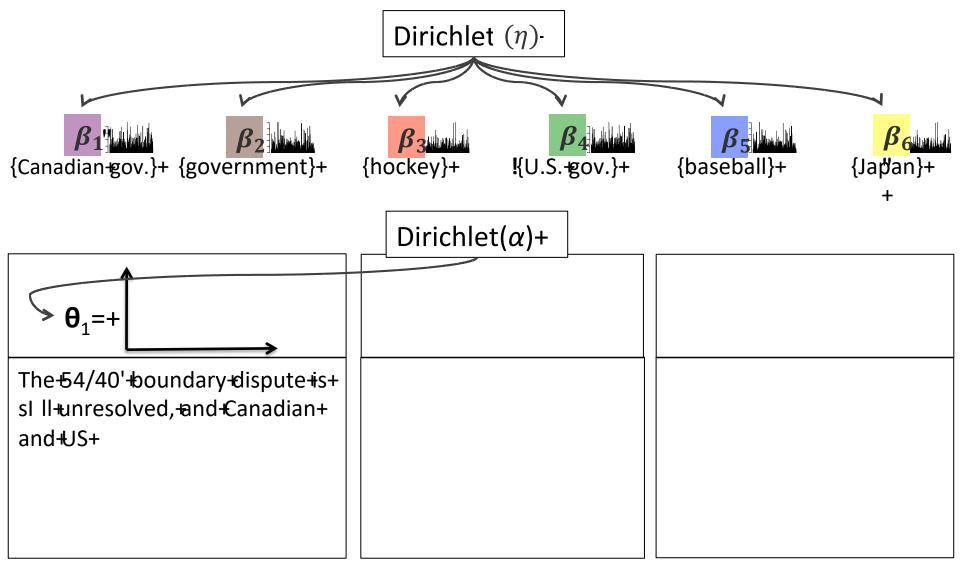


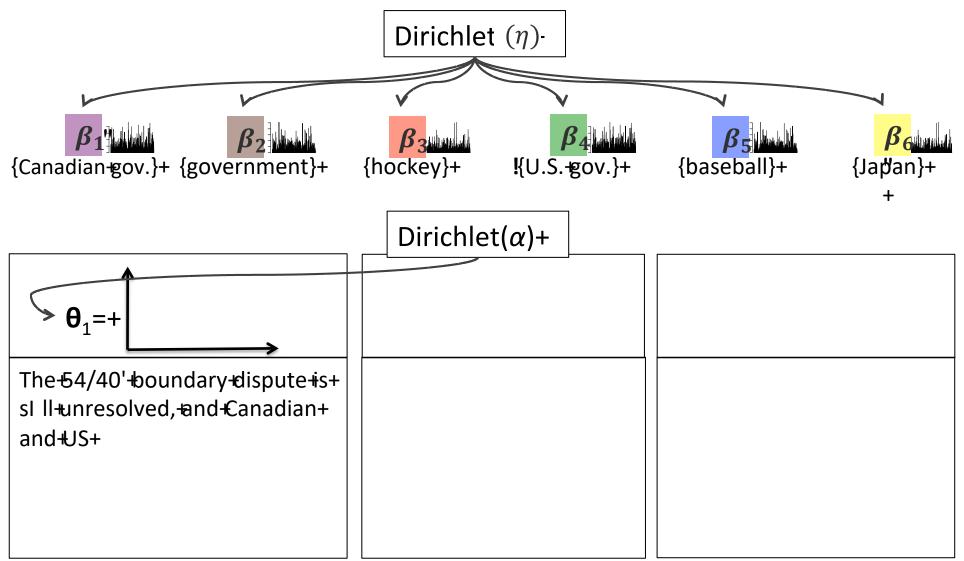
- A'topic'ls'visualized'as'lts'high&robability& words.'''
- A'þedagogical'**label&**s'used'to'ldentify'the'topic."

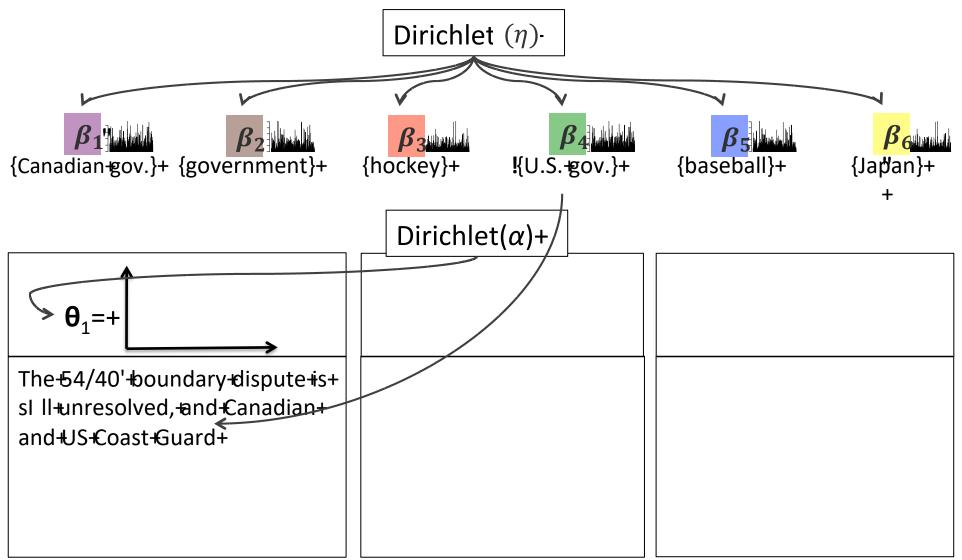


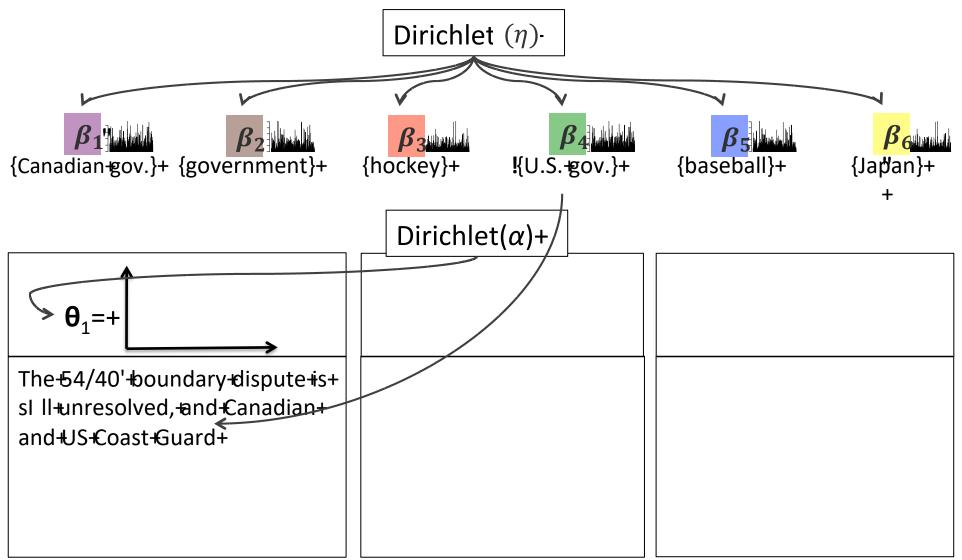
- A'topic'ls'visualized'as'lts'high'probability" words.'''
- A'þedagogical'**label**&s'used'to'ldentify'the'topic."

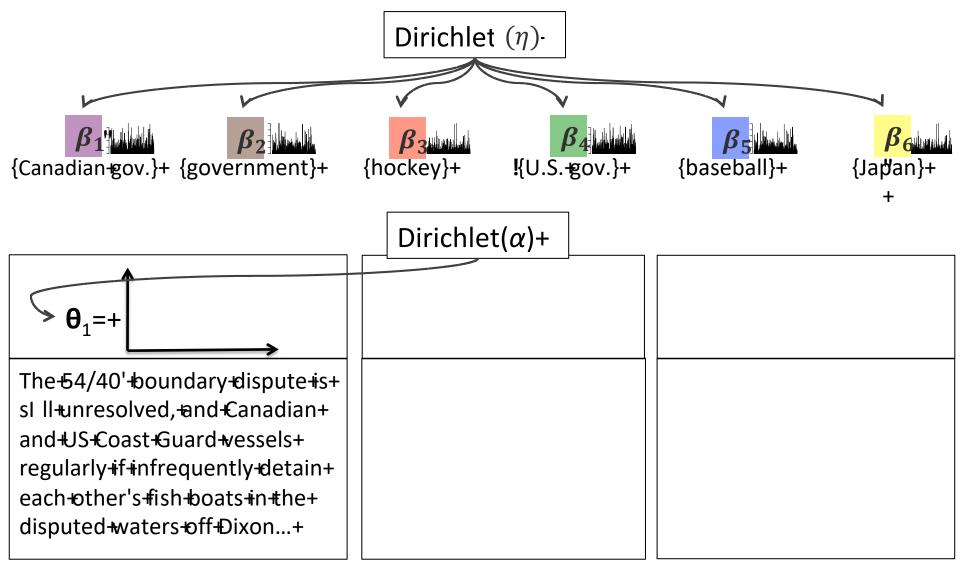


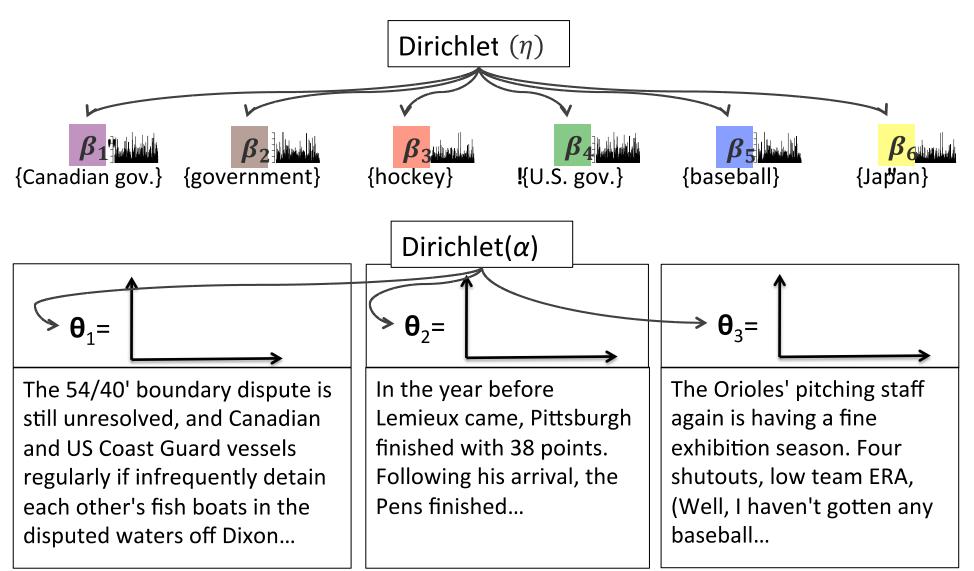


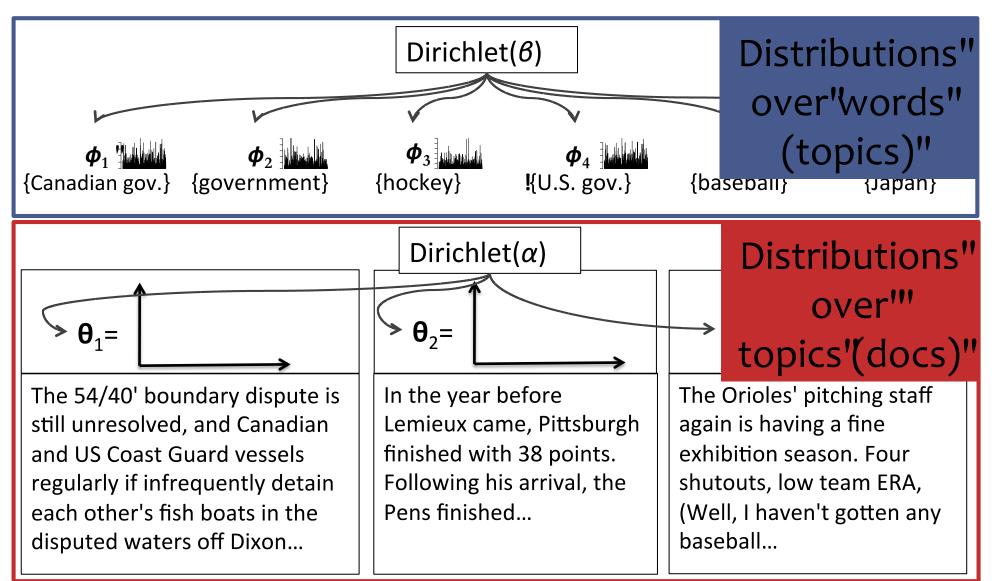




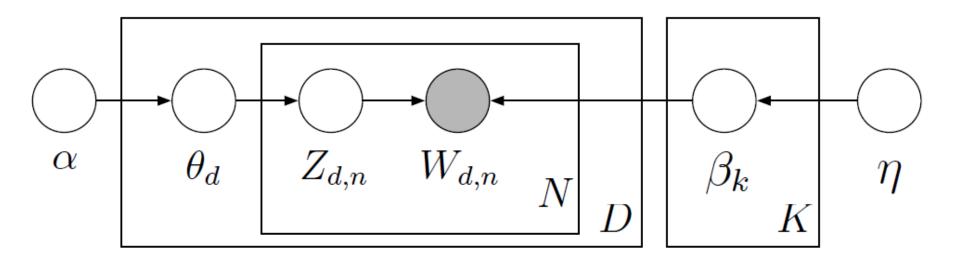








Joint Distribution for LDA

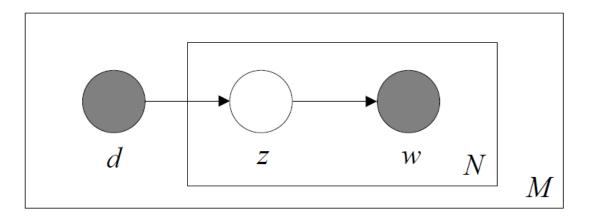


 Joint distribution of latent variables and documents is:

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{w}_{1:D} | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{i=1}^{K} p(\boldsymbol{\beta}_{i} | \boldsymbol{\eta}) \prod_{d=1}^{D} p(\boldsymbol{\theta}_{d} | \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} p(\boldsymbol{z}_{d,n} | \boldsymbol{\theta}_{d}) p(\boldsymbol{w}_{d,n} | \boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{d,n}) \right)$$

Learning of Topic Models

Recap: pLSA Topic Model



• Observed variables:

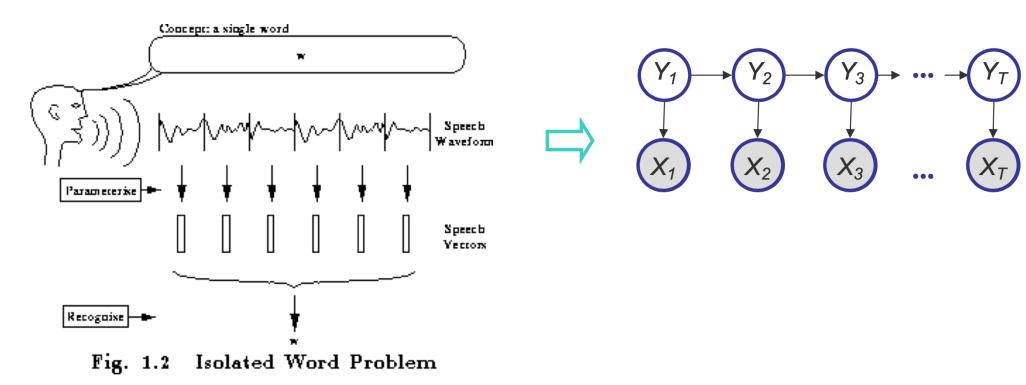
- Latent variables:
- Parameters:

The General Unsupervised Learning Problem

- Each data instance is partitioned into two parts:
 - \circ observed variables x
 - latent (unobserved) variables Z
- Want to learn a model $p_{\theta}(\mathbf{x}, \mathbf{z})$

Latent (unobserved) variables

- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., topic model, speech recognition models, ...



Latent (unobserved) variables

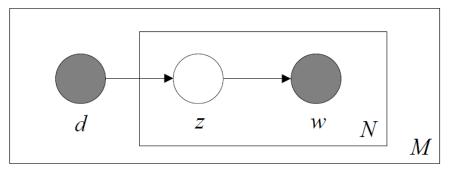
- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., topic model, speech recognition models, ...



Latent (unobserved) variables

- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., topic model, speech recognition models, ...
 - a real-world object (and/or phenomena), but difficult or impossible to measure
 - e.g., the temperature of a star, causes of a disease, evolutionary ancestors ...
 - a real-world object (and/or phenomena), but sometimes wasn't measured, because of faulty sensors, etc.
- Discrete latent variables can be used to partition/cluster data into sub-groups
- Continuous latent variables (factors) can be used for dimensionality reduction (e.g., factor analysis, etc.)

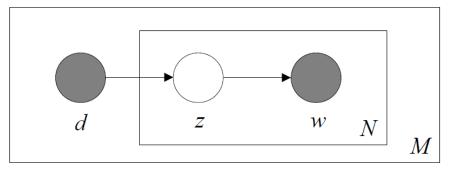
Recap: pLSA Topic Model



• Likelihood function of a word w:

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

Recap: pLSA Topic Model



• Likelihood function of a word w:

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

• Learning by maximizing the log likelihood:

Why is Learning Harder?

Why is Learning Harder?

• Complete log likelihood: if both x and z can be observed, then

 $\ell_c(\theta; \mathbf{x}, \mathbf{z}) = \log p(\mathbf{x}, \mathbf{z} | \theta) = \log p(\mathbf{z} | \theta_z) + \log p(\mathbf{x} | \mathbf{z}, \theta_x)$

- Decomposes into a sum of factors, the parameter for each factor can be estimated separately
- But given that z is not observed, $\ell_c(\theta; x, z)$ is a random quantity, cannot be maximized directly

Why is Learning Harder?

• Complete log likelihood: if both x and z can be observed, then

 $\ell_c(\theta; \mathbf{x}, \mathbf{z}) = \log p(\mathbf{x}, \mathbf{z}|\theta) = \log p(\mathbf{z}|\theta_z) + \log p(\mathbf{x}|\mathbf{z}, \theta_x)$

- Decomposes into a sum of factors, the parameter for each factor can be estimated separately
- But given that z is not observed, $\ell_c(\theta; x, z)$ is a random quantity, cannot be maximized directly
- Incomplete (or marginal) log likelihood: with z unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta) = \log \sum_{z} p(\mathbf{x}, \mathbf{z}|\theta)$$

- All parameters become coupled together
- In other models when z is complex (continuous) variables, marginalization over z is intractable.

Questions?