DSC250: Advanced Data Mining

Topic Models

Zhiting Hu

Lecture 7, Jan 28, 2025

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

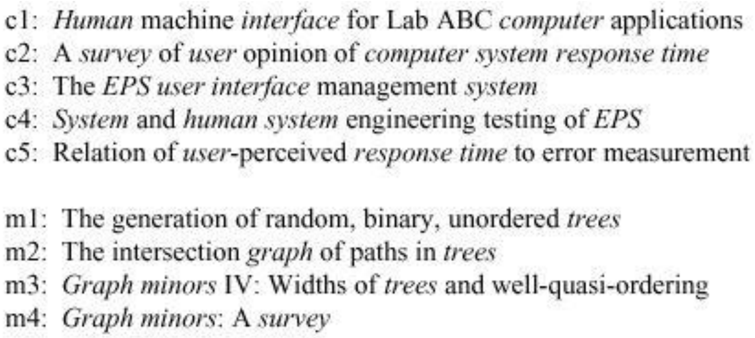
- Topic models: v1, v2, v3

- 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**

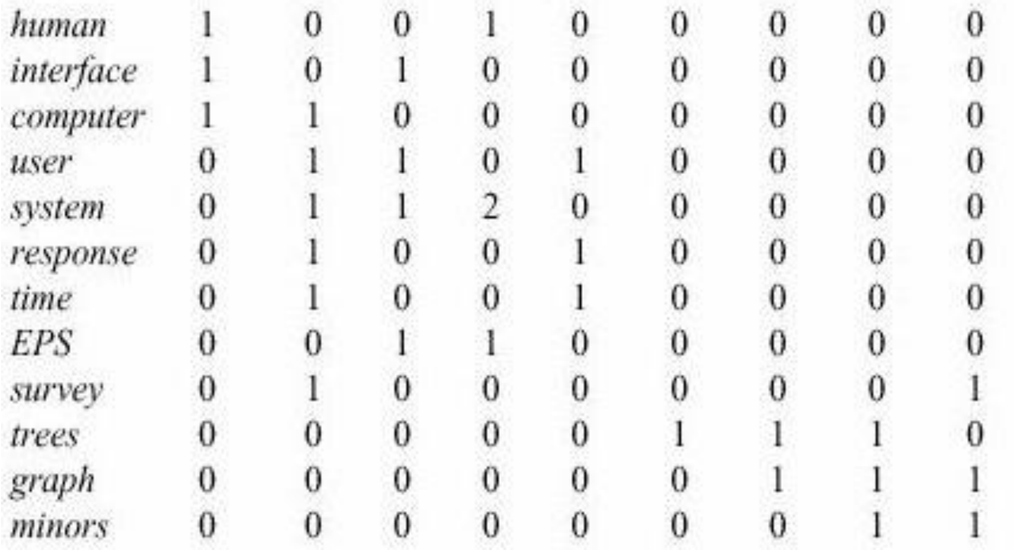
Recap: Represent a Document

- 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*

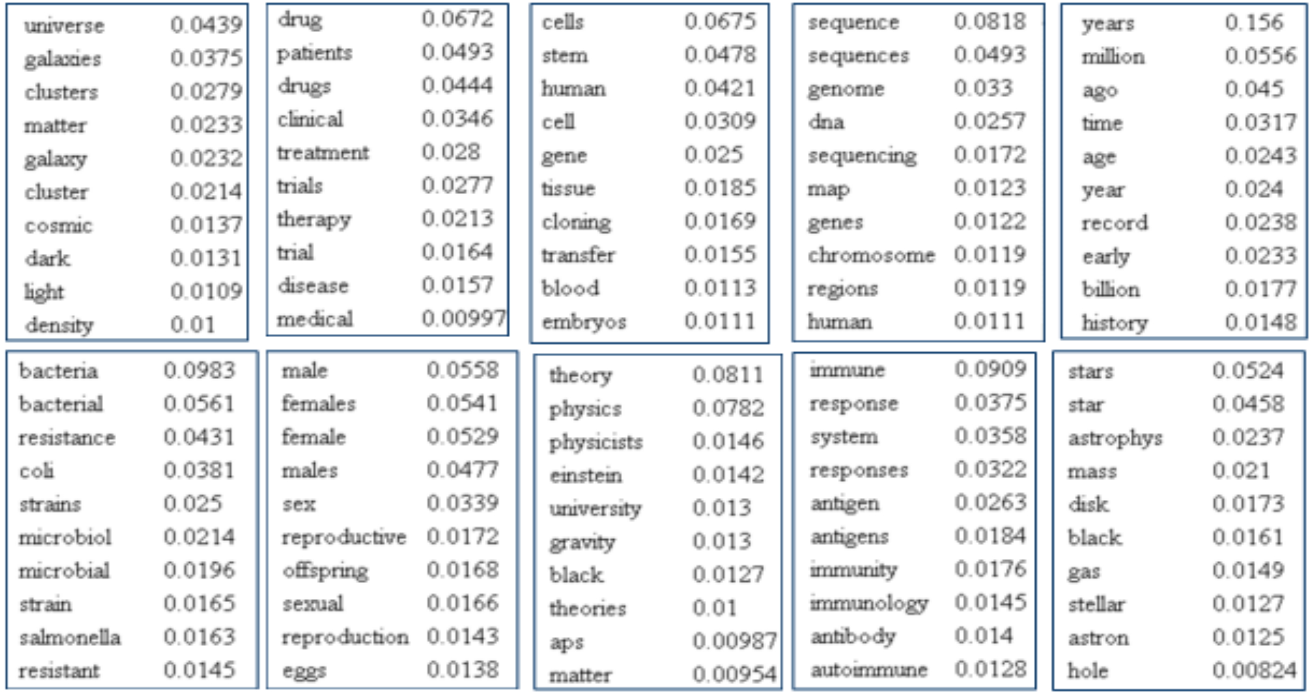


	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors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

Vector space model

Recap: Represent a Topic

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



universe	0.0439
galaxies	0.0375
clusters	0.0279
matter	0.0233
galaxy	0.0232
cluster	0.0214
cosmic	0.0137
dark	0.0131
light	0.0109
density	0.01

drug	0.0672
patients	0.0493
drugs	0.0444
clinical	0.0346
treatment	0.028
trials	0.0277
therapy	0.0213
trial	0.0164
disease	0.0157
medical	0.00997

cells	0.0675
stem	0.0478
human	0.0421
cell	0.0309
gene	0.025
tissue	0.0185
cloning	0.0169
transfer	0.0155
blood	0.0113
embryos	0.0111

sequence	0.0818
sequences	0.0493
genome	0.033
dna	0.0257
sequencing	0.0172
map	0.0123
genes	0.0122
chromosome	0.0119
regions	0.0119
human	0.0111

years	0.156
million	0.0556
ago	0.045
time	0.0317
age	0.0243
year	0.024
record	0.0238
early	0.0233
billion	0.0177
history	0.0148

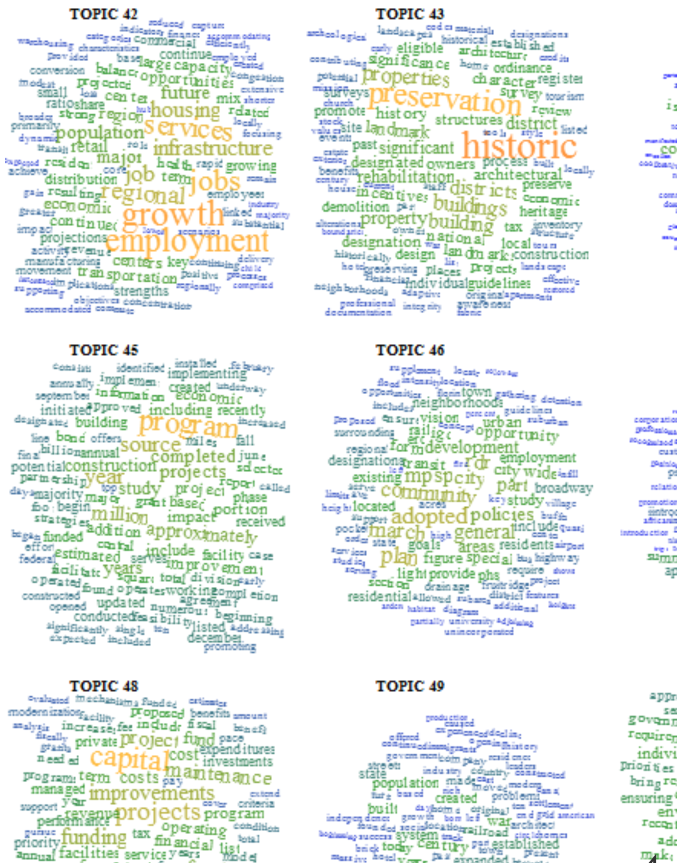
bacteria	0.0983
bacterial	0.0561
resistance	0.0431
coli	0.0381
strains	0.025
microbiol	0.0214
microbial	0.0196
strain	0.0165
salmonella	0.0163
resistant	0.0145

male	0.0558
females	0.0541
female	0.0529
males	0.0477
sex	0.0339
reproductive	0.0172
offspring	0.0168
sexual	0.0166
reproduction	0.0143
eggs	0.0138

theory	0.0811
physics	0.0782
physicists	0.0146
einstein	0.0142
university	0.013
gravity	0.013
black	0.0127
theories	0.01
aps	0.00987
matter	0.00954

immune	0.0909
response	0.0375
system	0.0358
responses	0.0322
antigen	0.0263
antigens	0.0184
immunity	0.0176
immunology	0.0145
antibody	0.014
autoimmune	0.0128

stars	0.0524
star	0.0458
astrophys	0.0237
mass	0.021
disk	0.0173
black	0.0161
gas	0.0149
stellar	0.0127
astron	0.0125
hole	0.00824

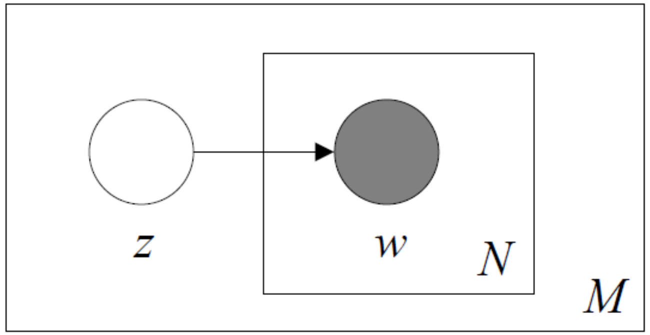


Notations

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

Recap: Topic Model v1: Multinomial Mixture Model

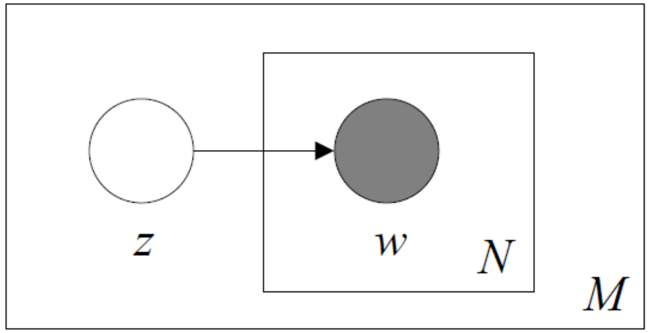
Graphical Model



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

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

Recap: Likelihood Function



$$L = \prod_d p(\boldsymbol{x}_d) \;= \prod_d \sum_k p(\boldsymbol{x}_d, z = k)$$

$$= \prod_d \sum_k p(\boldsymbol{x}_d | z = k) p(z = k)$$

$$= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}}$$

Recap: Topic Model v2: pLSA

- For each position in d, $n = 1, \dots, N_d$
  - Generate the topic for the position as

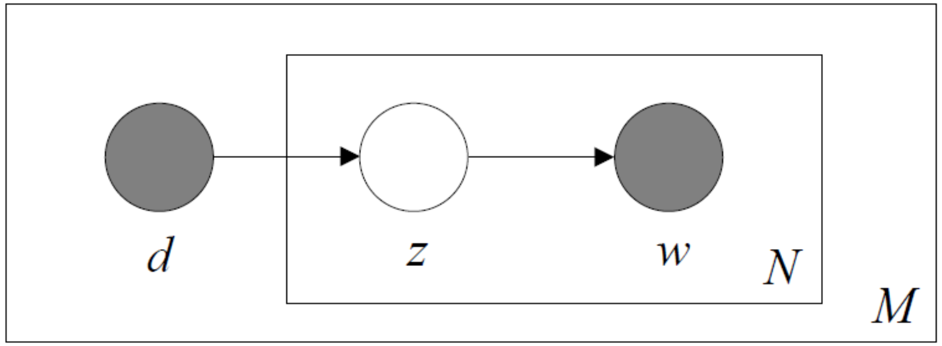
$$z_n|d \sim Categorical(\boldsymbol{\theta}_d), i.e., p(z_n = k|d) = \theta_{dk}$$

(Note, 1 trial multinomial)

  - Generate the word for the position as

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

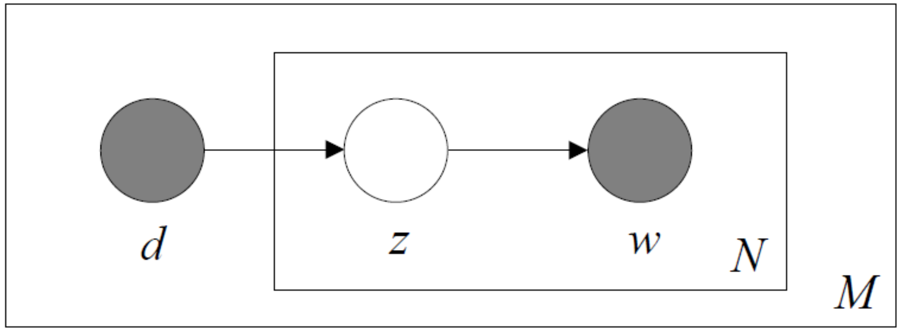
Graphical Model

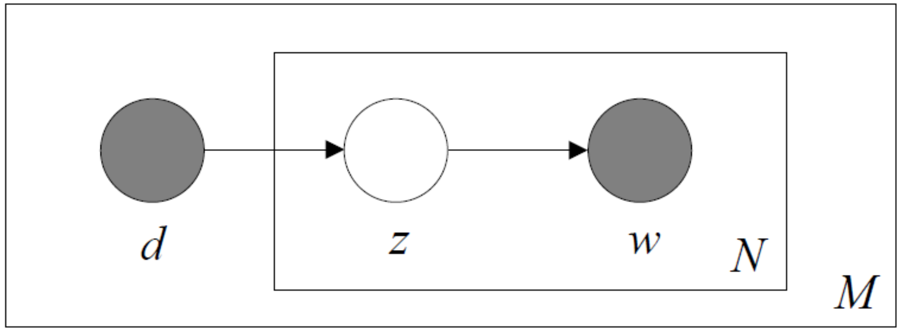


Likelihood Function

- Probability of a word w

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



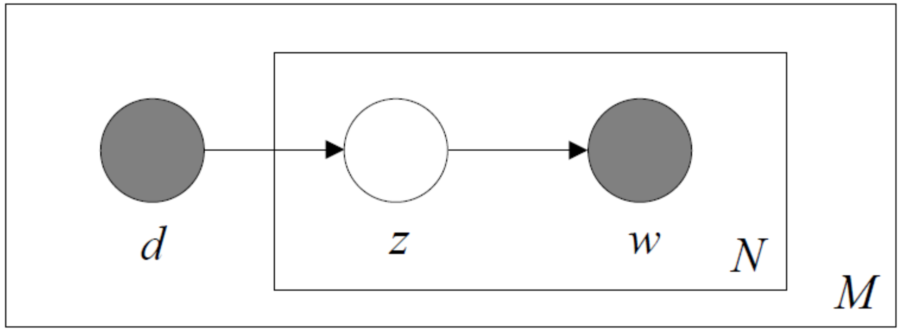


Likelihood Function

- 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

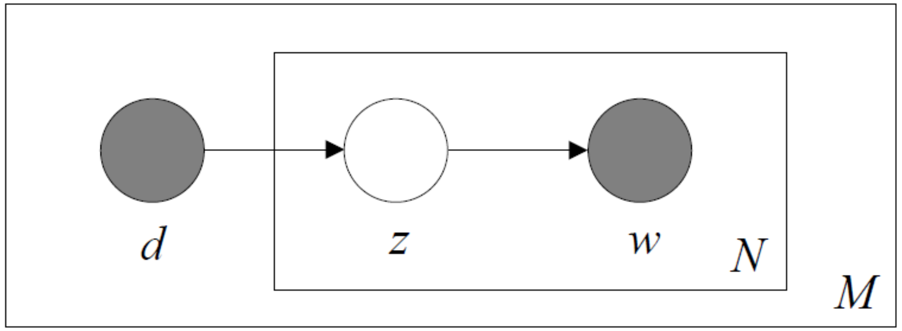
- Probability of a word w

$$p(w|d,\theta,\beta) = \sum_k p(w, z = k|d,\theta,\beta)$$

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

- 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

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

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

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

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

Re-arrange the Likelihood Function

- Group the same word from different positions together

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

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

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

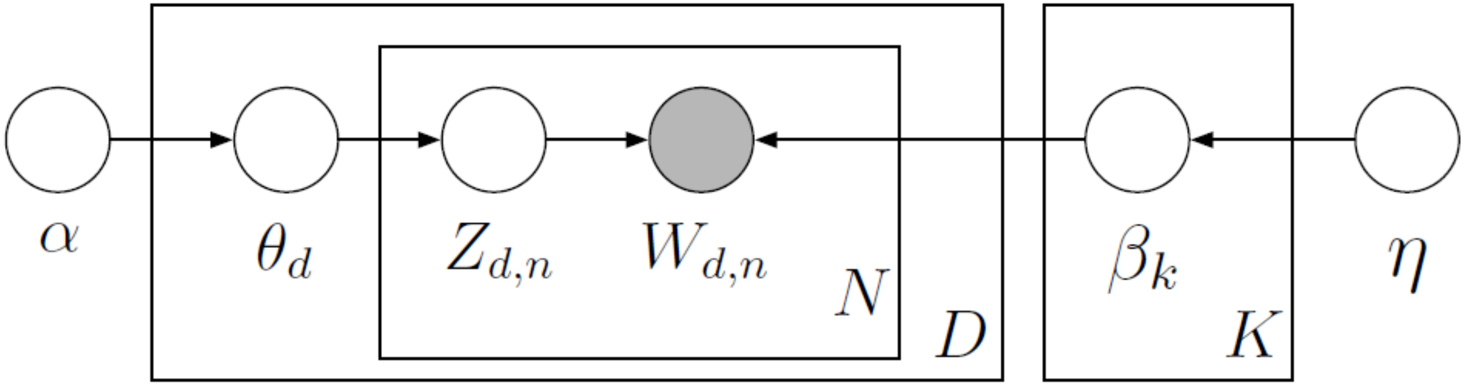
- Not a proper generative model
  - $\boldsymbol{\theta}_d$ is treated as a parameter
  - Cannot model new documents

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

⇩

Topic Model v3: Latent Dirichlet Allocation (LDA)

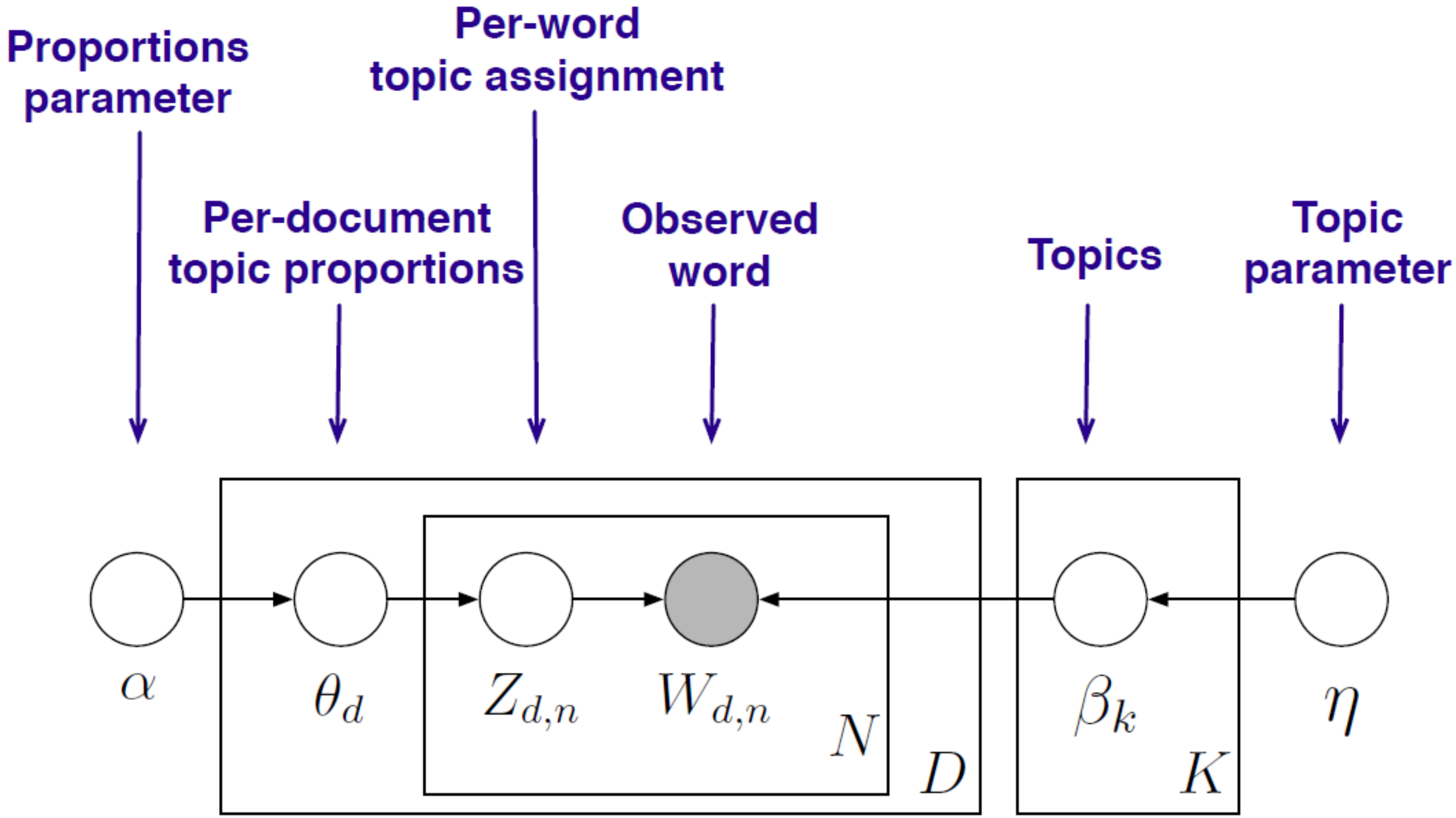
Topic Model v3: Latent Dirichlet Allocation (LDA)



$\boldsymbol{\theta_d \sim Dirichlet(\alpha)}$: **address topic distribution for unseen documents**

$\boldsymbol{\beta_k \sim Dirichlet(\eta)}$: **smoothing over words**

Topic Model v3: Latent Dirichlet Allocation (LDA)

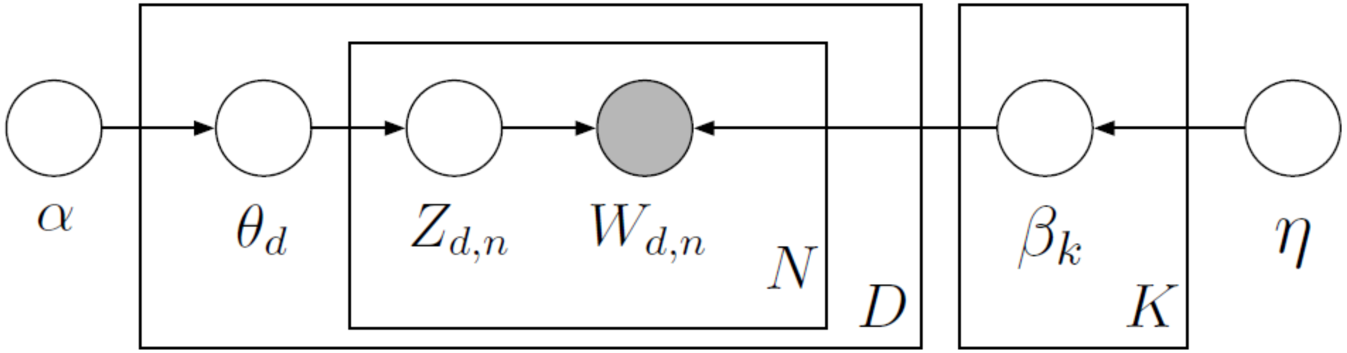


$\theta_d \sim Dirichlet(\alpha)$: address topic distribution for unseen documents

$\beta_k \sim Dirichlet(\eta)$: smoothing over words

Generative Model for LDA

```
For each topic $k \in \{1, \ldots, K\}$:
    $\beta_k \sim \mathrm{Dir}(\eta)$                              [*draw distribution over words*]
For each document $d \in \{1, \ldots, D\}$
    $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$                              [*draw distribution over topics*]
    For each word $n \in \{1, \ldots, N_d\}$
        $z_{d,n} \sim \mathrm{Mult}(1, \boldsymbol{\theta}_d)$                   [*draw topic assignment*]
        $w_{d,n} \sim \theta_{z_{d,n}}$                              [*draw word*]
```



Review: Dirichlet Distribution

- Dirichlet distribution: $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
  - $i.e., p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, where $\alpha_k > 0$
    - $\Gamma(\cdot)$ *is gamma function*: $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$
      - $\Gamma(z+1) = \mathrm{z}\Gamma(z)$

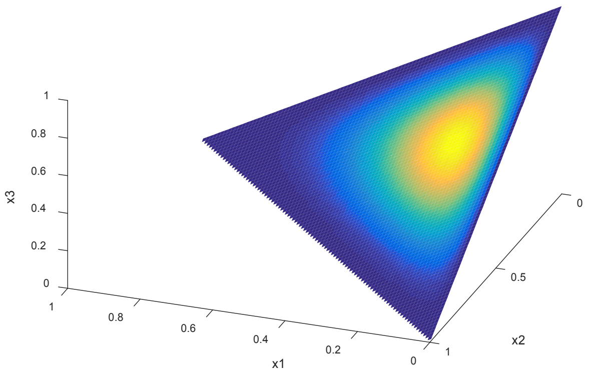
Review: Dirichlet Distribution

- Dirichlet distribution: $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
  - $i.e., p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, 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:

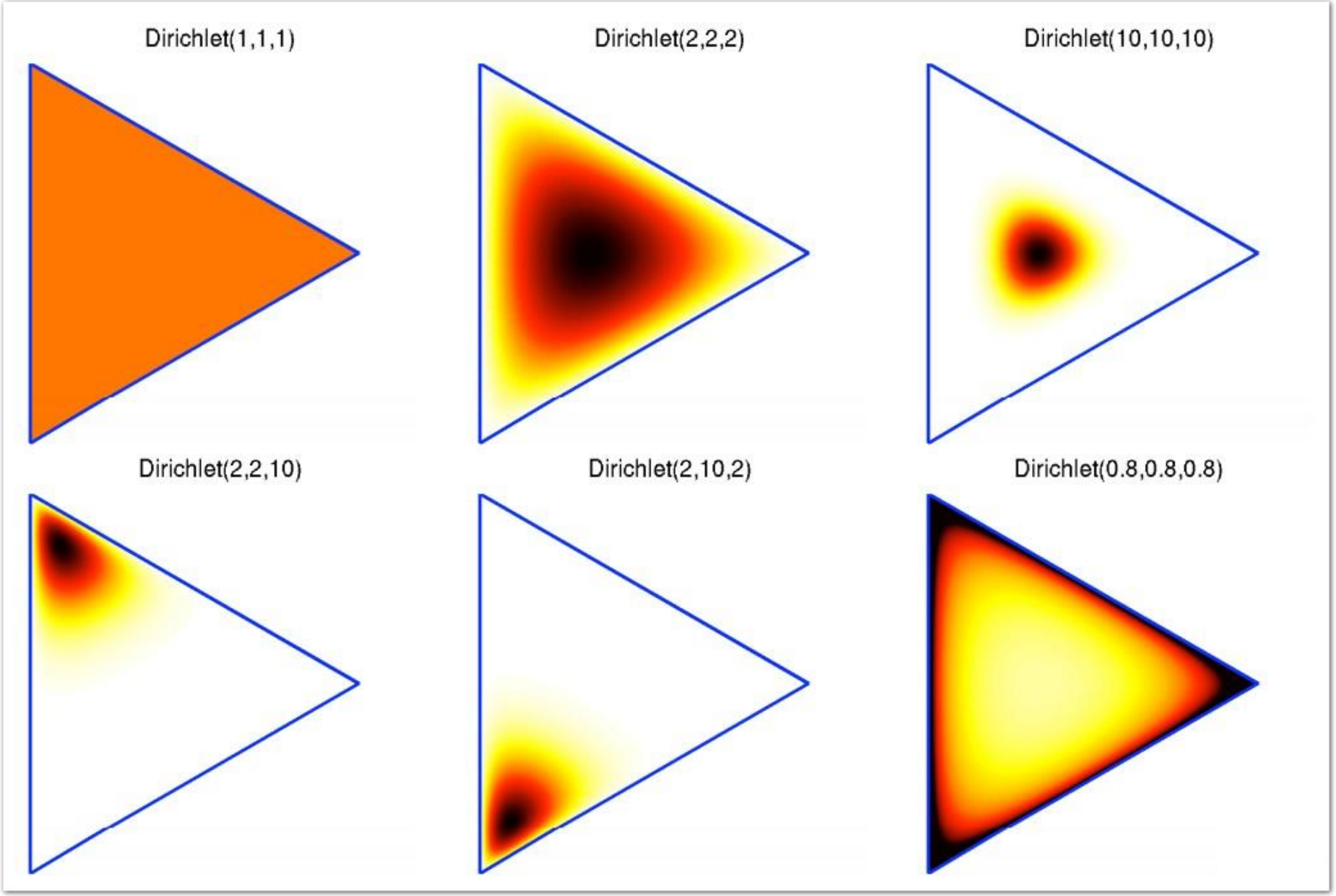
$$\theta = \theta_1 (1,0,0) + \theta_2 (0,1,0) + \theta_3 (0,0,1)$$

- Where $0 \le \theta_1, \theta_2, \theta_3 \le 1$ *and* $\theta_1 + \theta_2 + \theta_3 = 1$



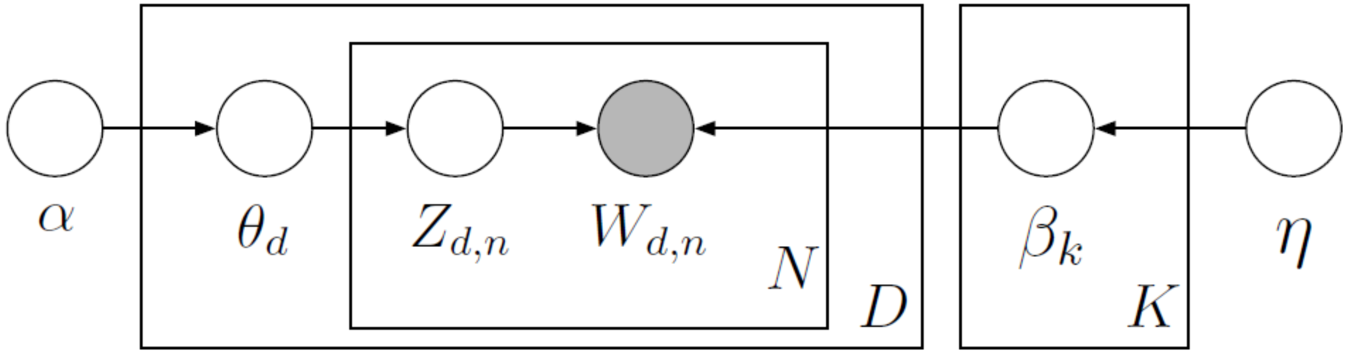
$\theta \,|\alpha \sim Dir(\alpha), \alpha = (2,3,4)$

More Examples in the Simplex View



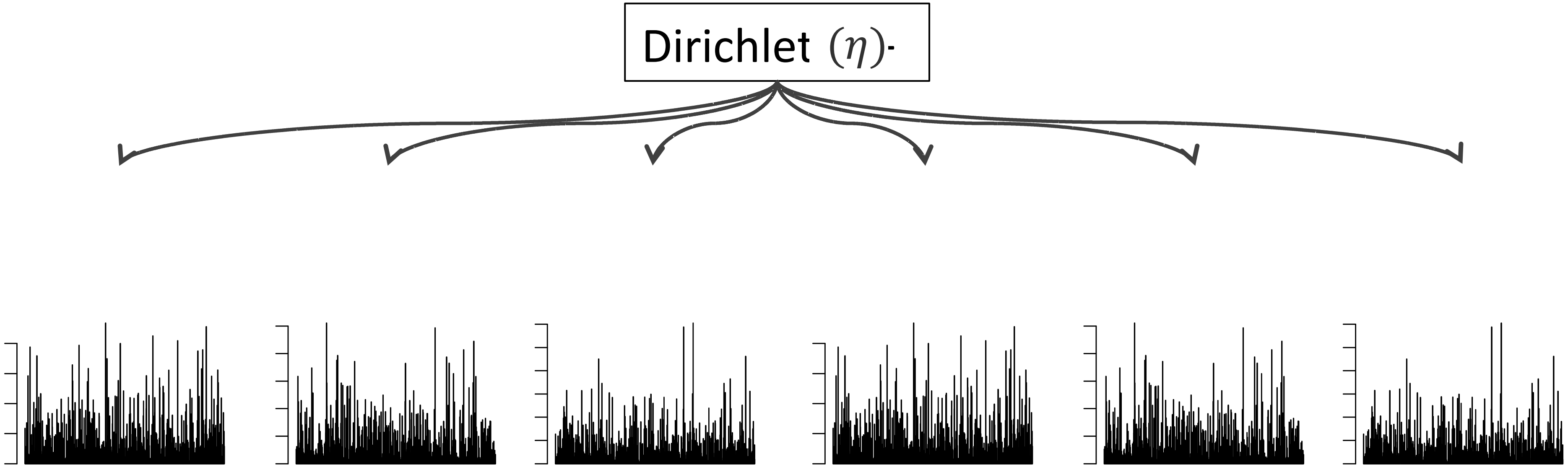
Generative Model for LDA

```
For each topic $k \in \{1, \ldots, K\}$:
    $\beta_k \sim \mathrm{Dir}(\eta)$                                [draw distribution over words]
For each document $d \in \{1, \ldots, D\}$
    $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$                               [draw distribution over topics]
    For each word $n \in \{1, \ldots, N_d\}$
        $z_{d,n} \sim \mathrm{Mult}(1, \boldsymbol{\theta}_d)$                   [draw topic assignment]
        $w_{d,n} \sim \theta_{z_{d,n}}$                                 [draw word]
```



(Blei,'Ng,'&'Jordan,'2003)"

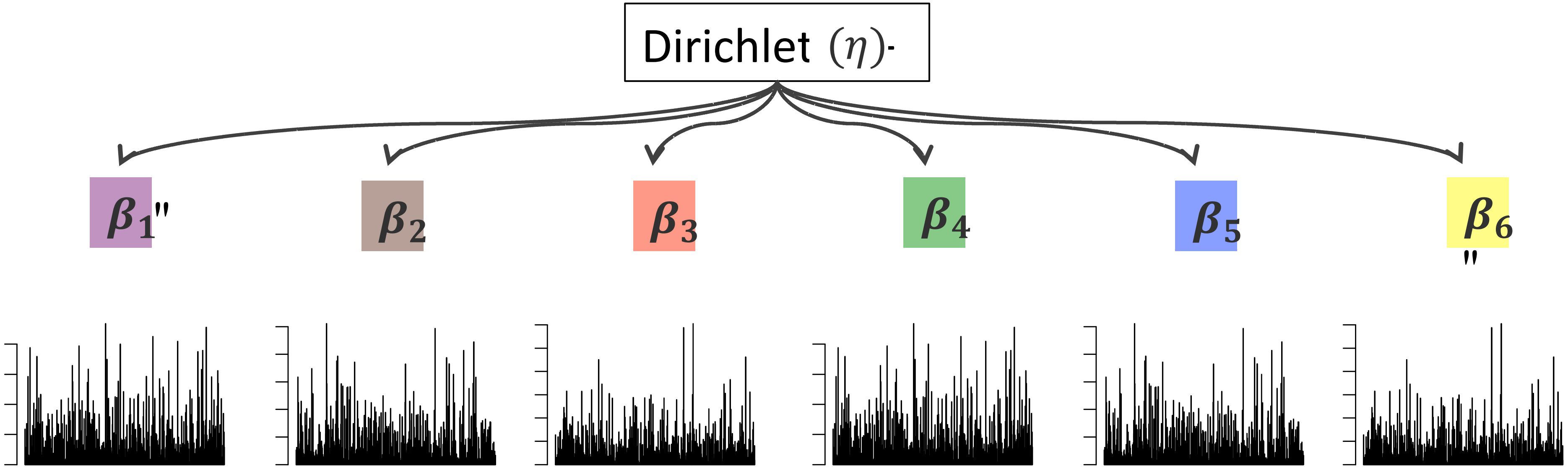
LDA"for"Topic"Modeling"



- The"**generative&story&**begins"with"only"a"**Dirichlet& prior&**over"the"topics."
- Each"**topic&**is"defind "as"a"**Multinomial&distribution**" over"the"vocabulary,"parameterized"by' $\boldsymbol{\beta_k}$ '

(Blei, Ng, & Jordan, 2003)

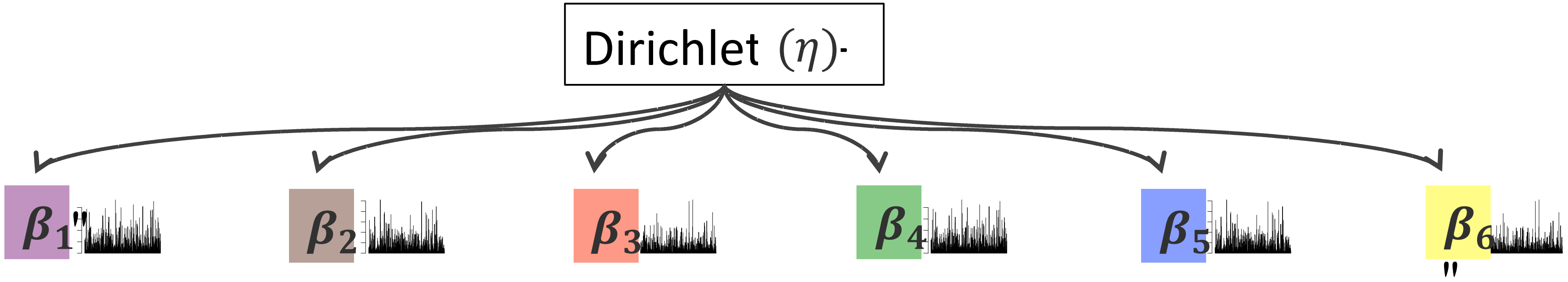
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defind as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\beta}_k$

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"

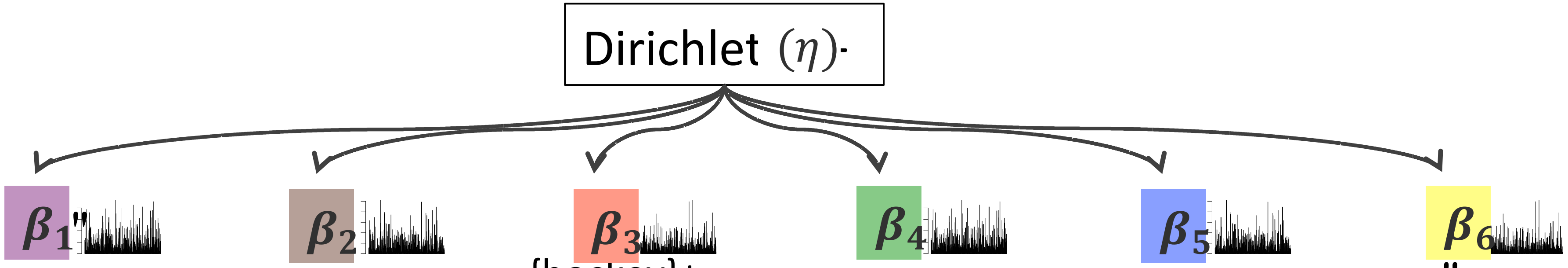


team,+season,+
hockey,+player,+
penguins,+ice,++
canadiens,+
puck,+montreal,+
stanley,+cup+

- A'topic'is'visualized'as'its'**high&probability& words.'"**

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"



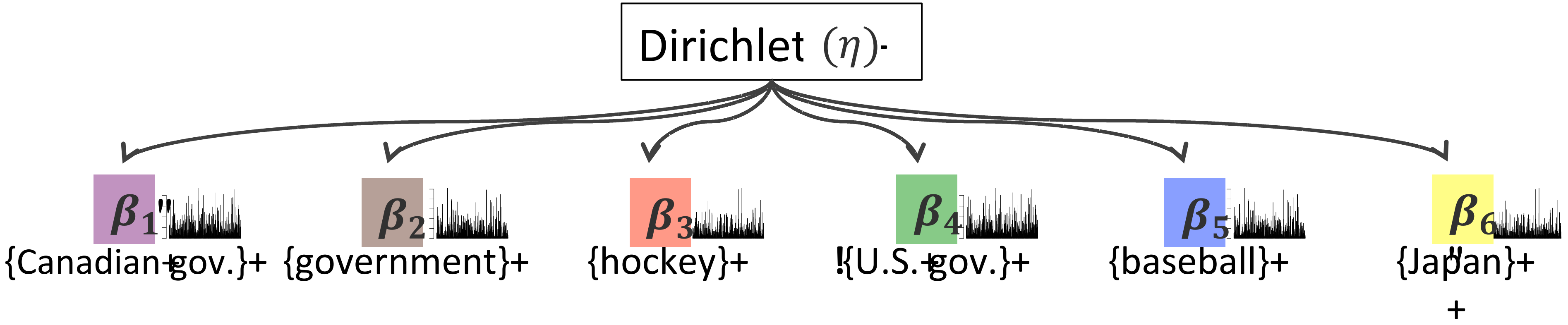
{hockey}+

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

- A'topic'is'visualized'as'its'**high&probability& words.**'"
- A'pedagogical'**label&**is'used'to'identify'the'topic."

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"



Dirichlet (η)

β_1 {Canadian gov.}

β_2 {government}

β_3 {hockey}

β_4 {U.S. gov.}

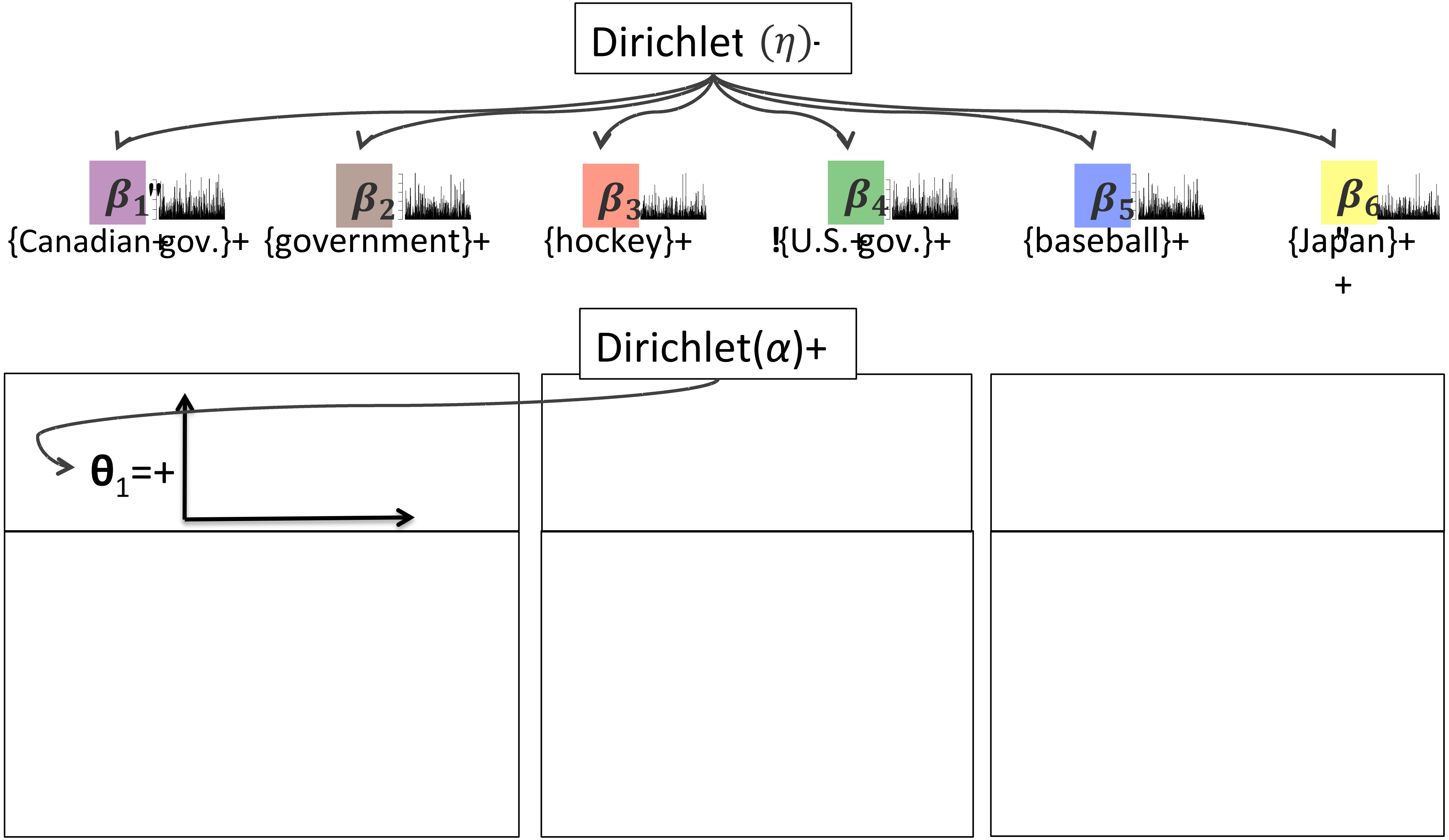
β_5 {baseball}

β_6 {Japan}

- A'topic'is'visualized'as'its'high'probability" words.'"
- A'pedagogical'**label**'is'used'to'identify'the'topic."

(Blei,'Ng,'&"Jordan,'2003)"

LDA'for"Topic'Modeling"



Dirichlet (η)-

β_1" {Canadian+gov.}+

β_2 {government}+

β_3 {hockey}+

β_4 !{U.S.+gov.}+

β_5 {baseball}+

β_6 {Ja"pan}+

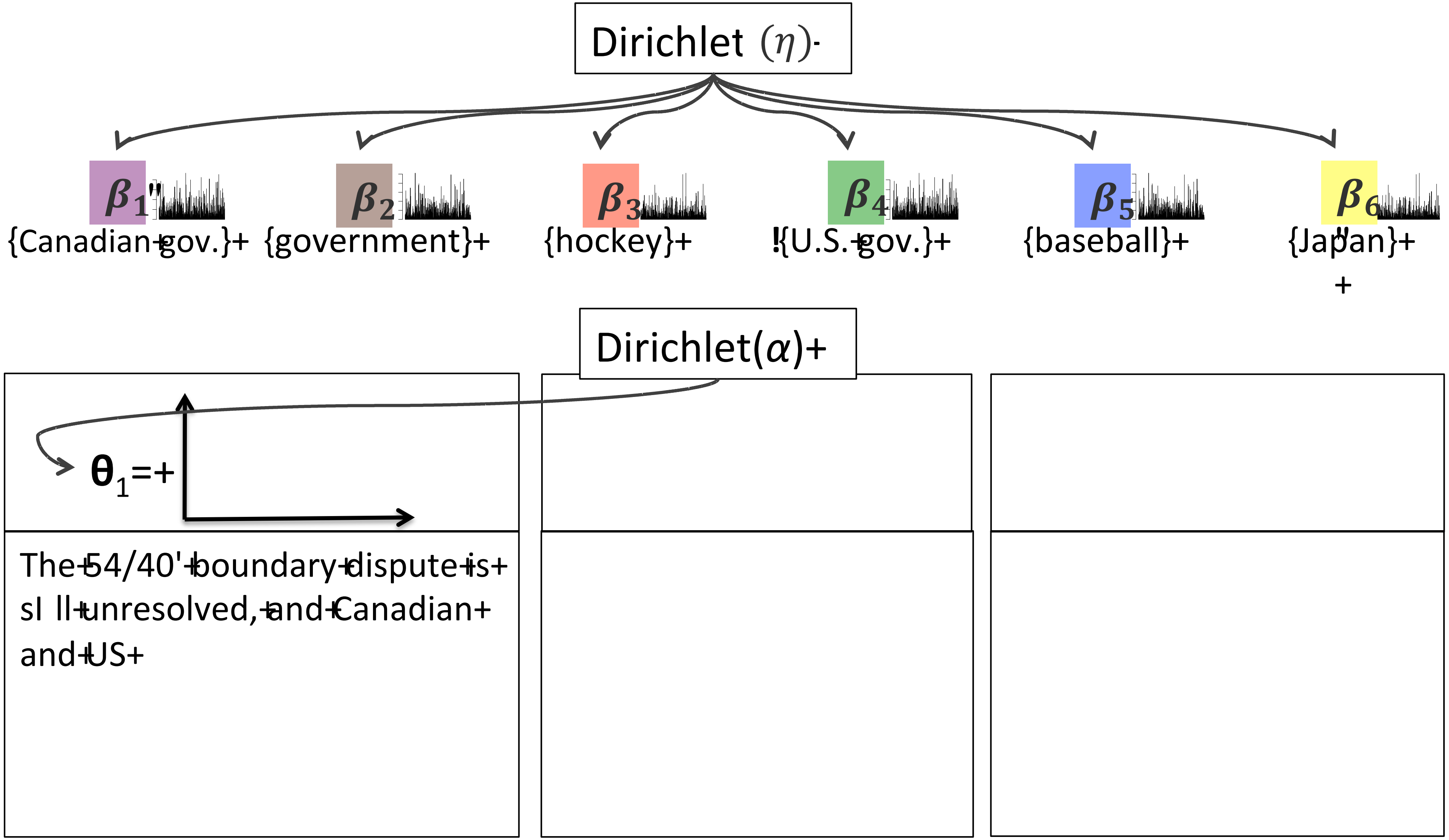
+

Dirichlet(α)+

θ_1=+

(Blei,'Ng,'&"Jordan,'2003)"

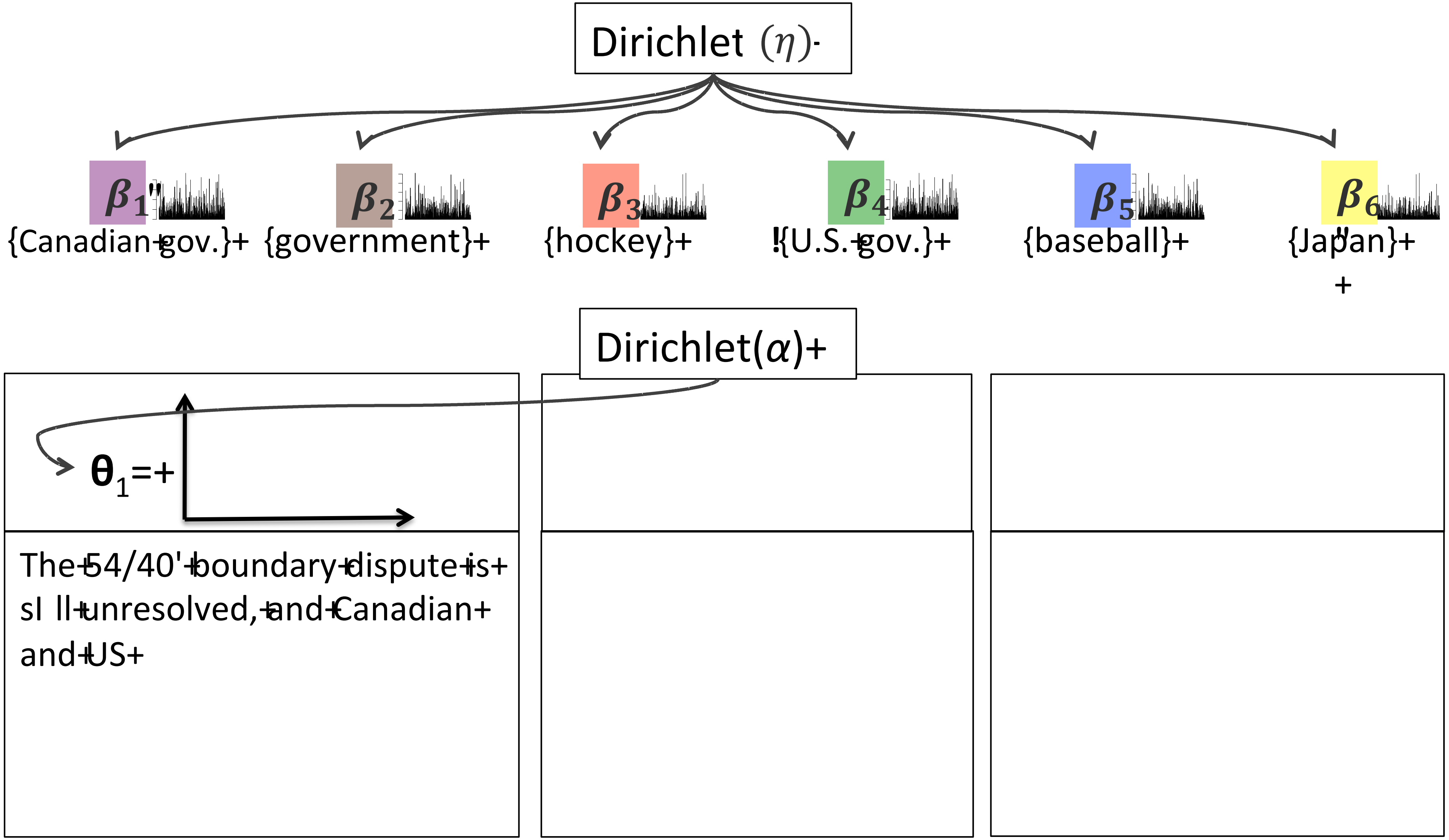
LDA"for"Topic"Modeling"



The+54/40'+boundary+dispute+is+
sI ll+unresolved,+and+Canadian+
and+US+

(Blei,'Ng,'&"Jordan,'2003)"

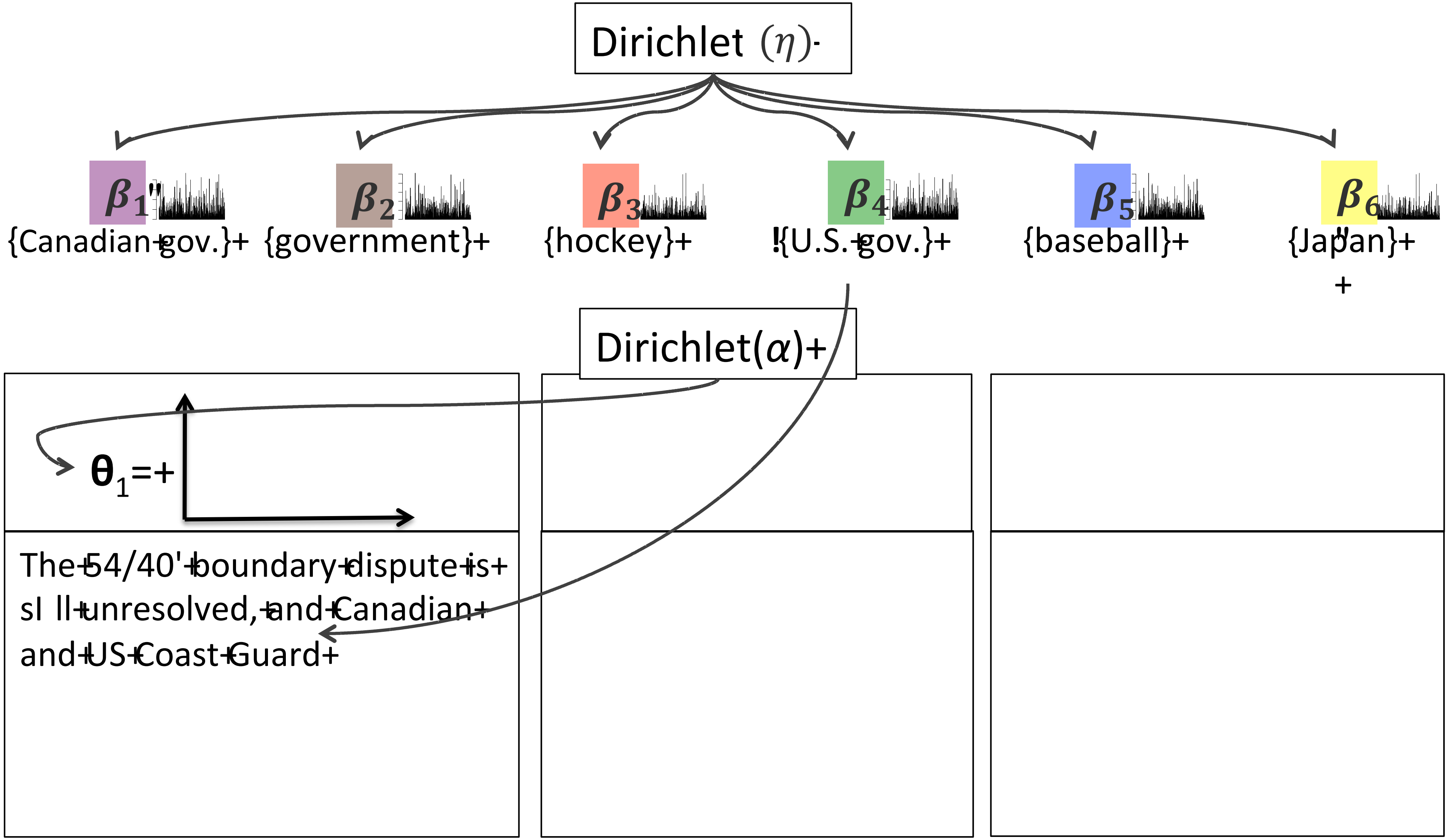
LDA"for"Topic"Modeling"



The+54/40'+boundary+dispute+is+ sI ll+unresolved,+and+Canadian+ and+US+

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"



Dirichlet (η)·

β_1" {Canadian+gov.}+

β_2 {government}+

β_3 {hockey}+

β_4 !{U.S.+gov.}+

β_5 {baseball}+

β_6 {Jap"an}+

+

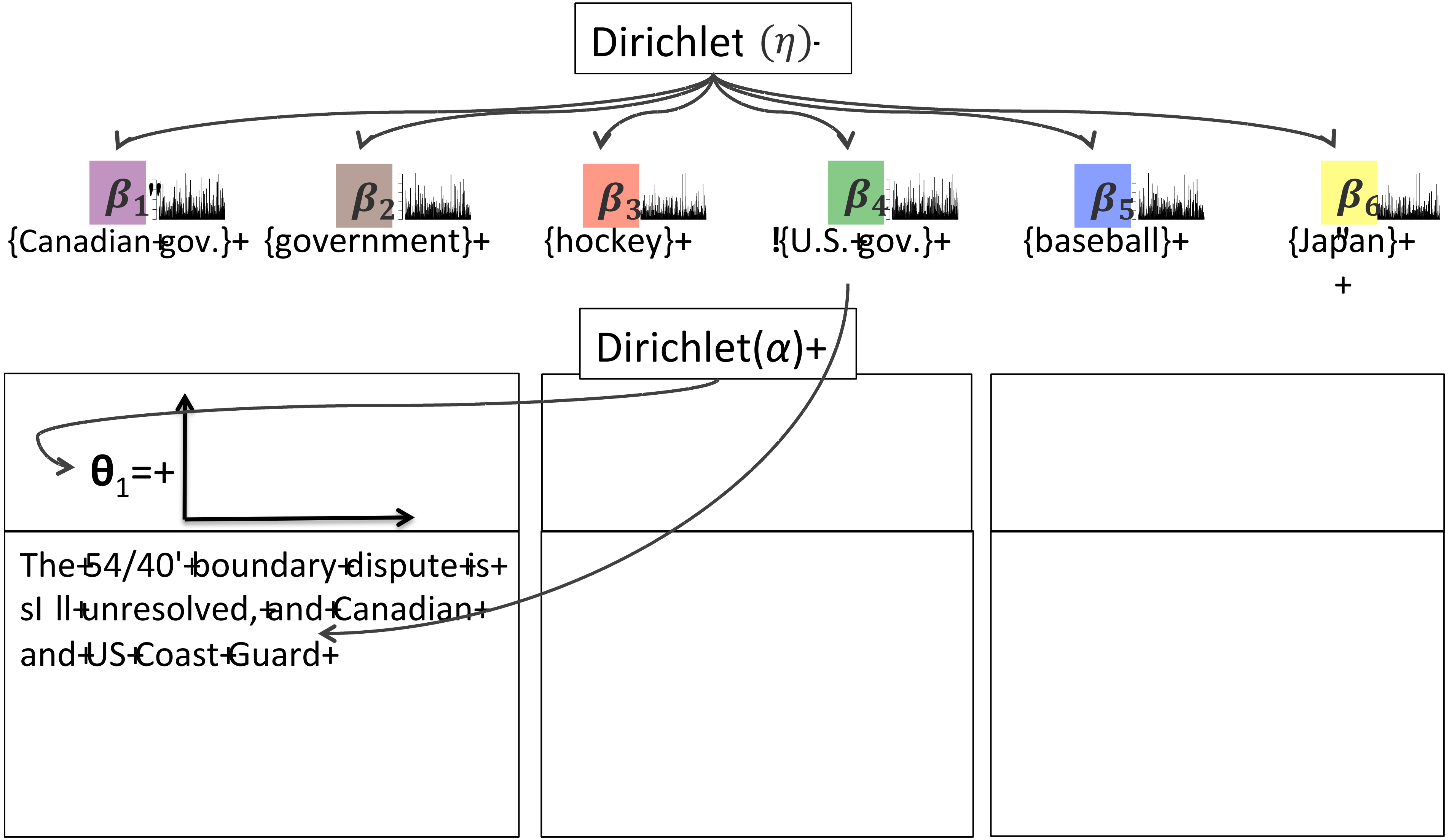
Dirichlet(α)+

θ_1=+

The+54/40'+boundary+dispute+is+ still+unresolved,+and+Canadian+ and+US+Coast+Guard+

(Blei,'Ng,'&'Jordan,'2003)"

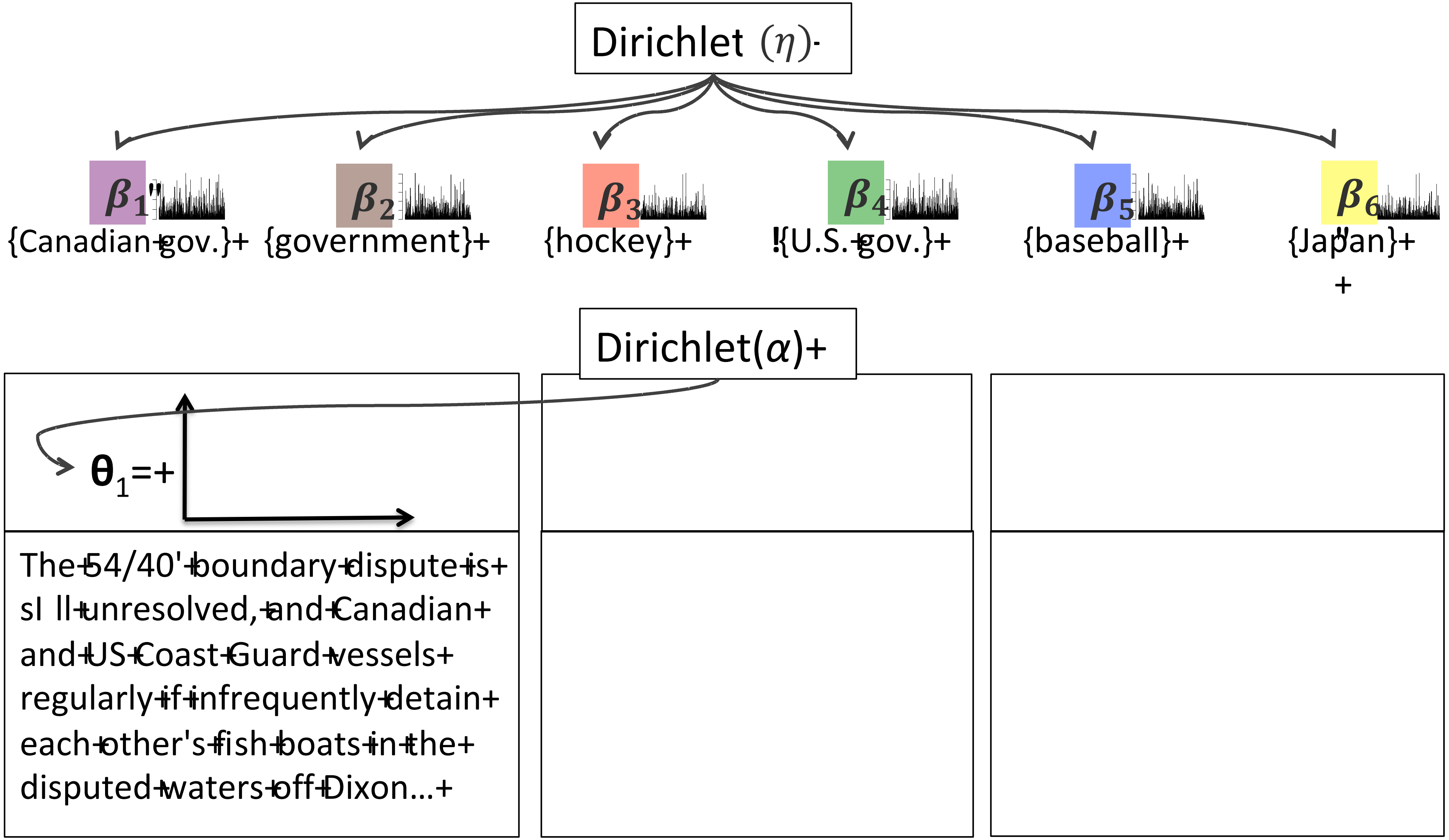
LDA"for"Topic"Modeling"



The+54/40'+boundary+dispute+is+ sl ll+unresolved,+and+Canadian+ and+US+Coast+Guard+

(Blei,'Ng,'&'Jordan,'2003)"

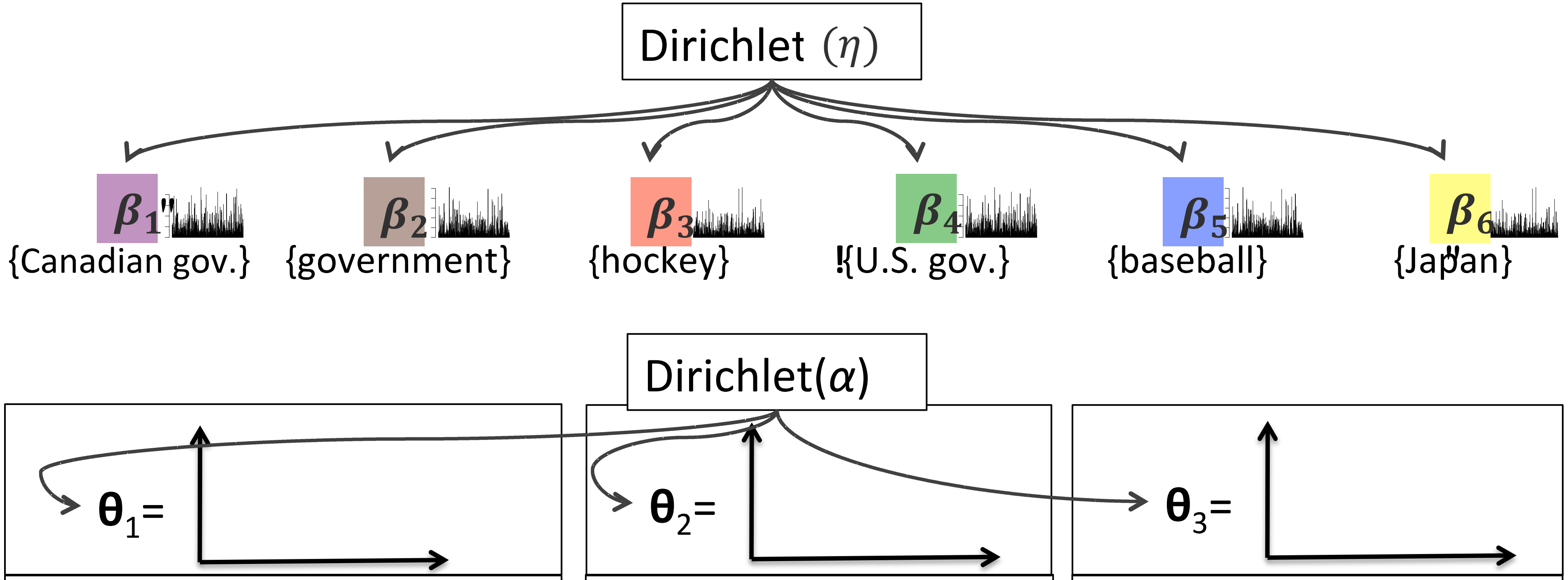
LDA'for'Topic'Modeling"



The+54/40'+boundary+dispute+is+ still+unresolved,+and+Canadian+ and+US+Coast+Guard+vessels+ regularly+if+infrequently+detain+ each+other's+fish+boats+in+the+ disputed+waters+off+Dixon…+

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"



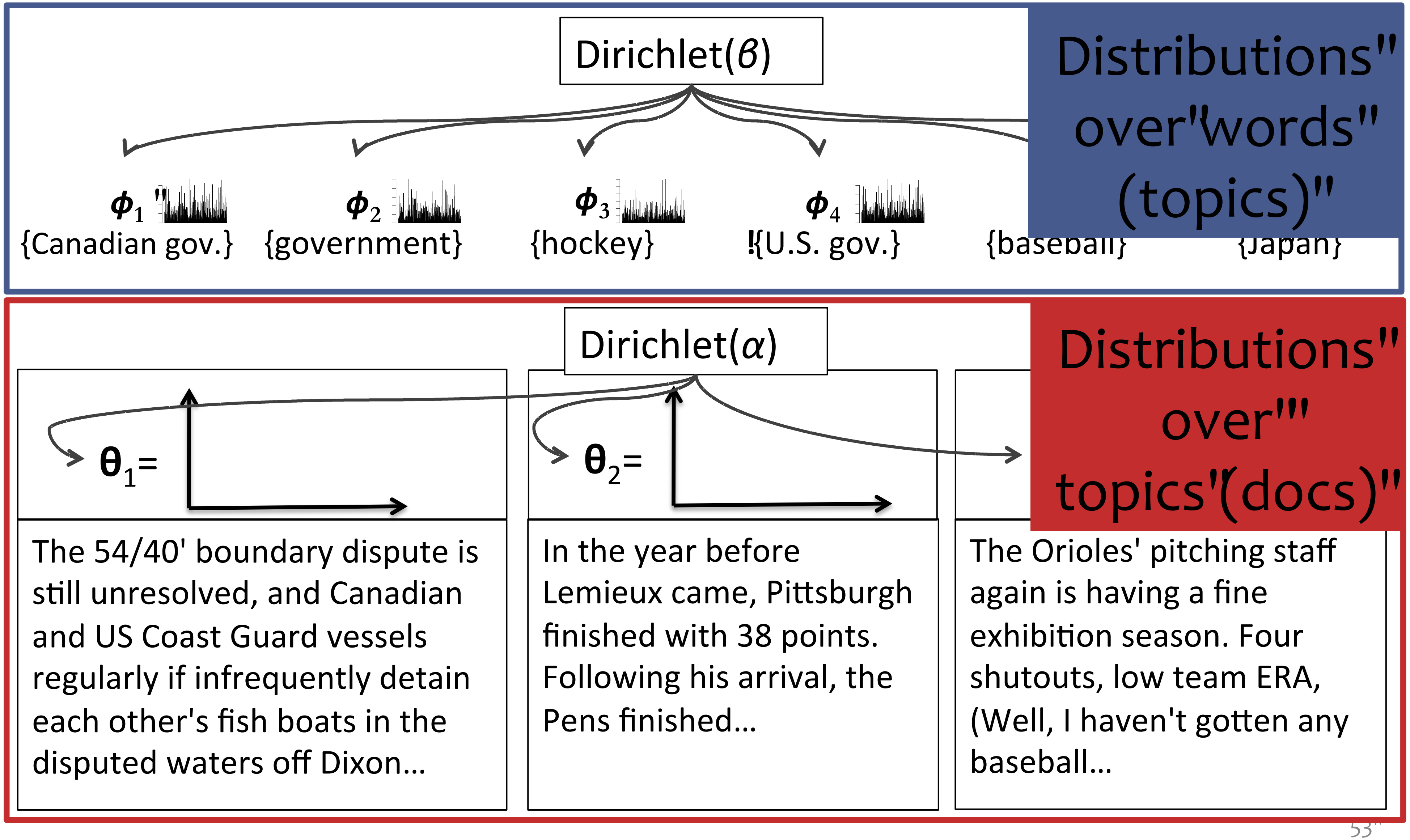
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon…

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished…

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball…

(Blei,'Ng,'&'Jordan,'2003)"

LDA'for'Topic'Modeling"

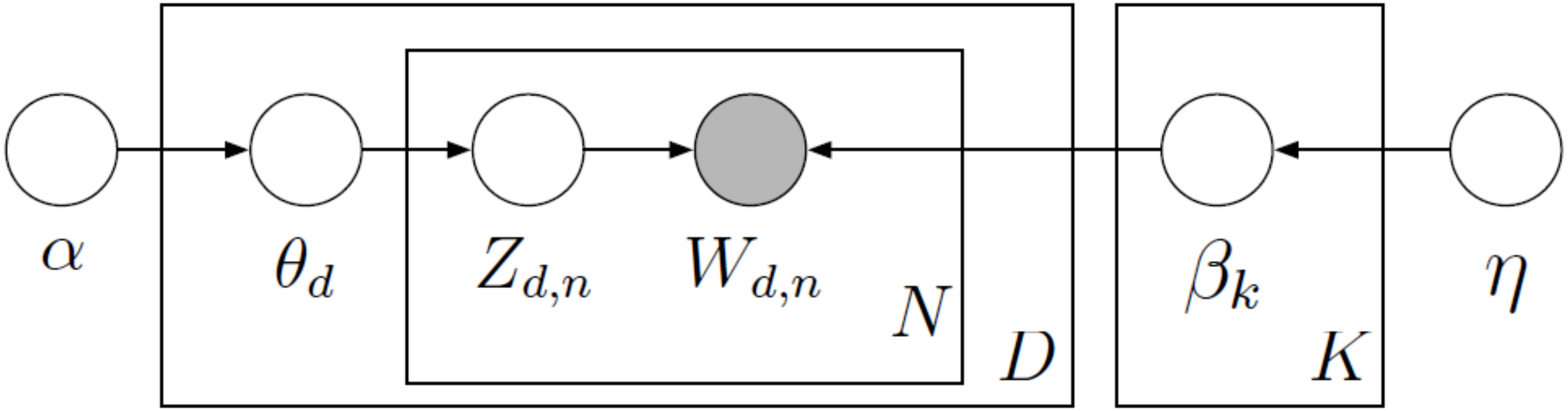


The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon…

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished…

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball…

Joint Distribution for LDA

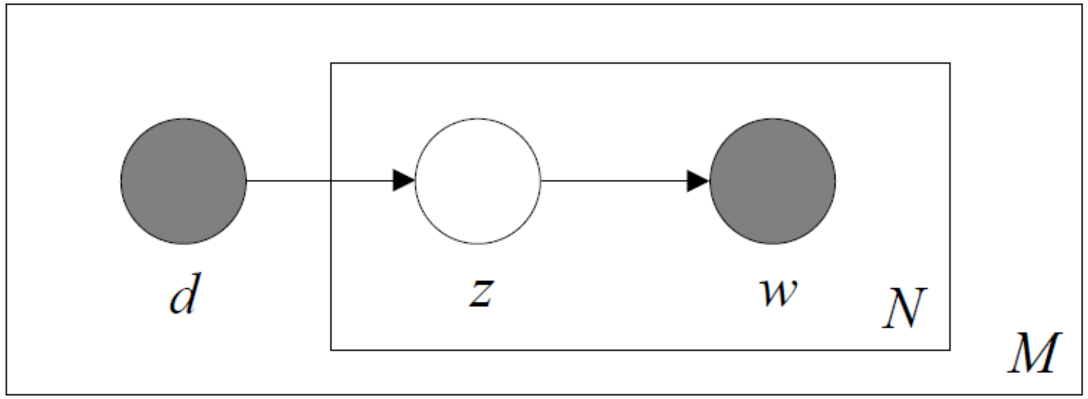


- Joint distribution of latent variables and documents is: $p(\boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{w}_{1:D} | \alpha, \eta) =$

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

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:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

[Content adapted from CMU 10-708]

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, …

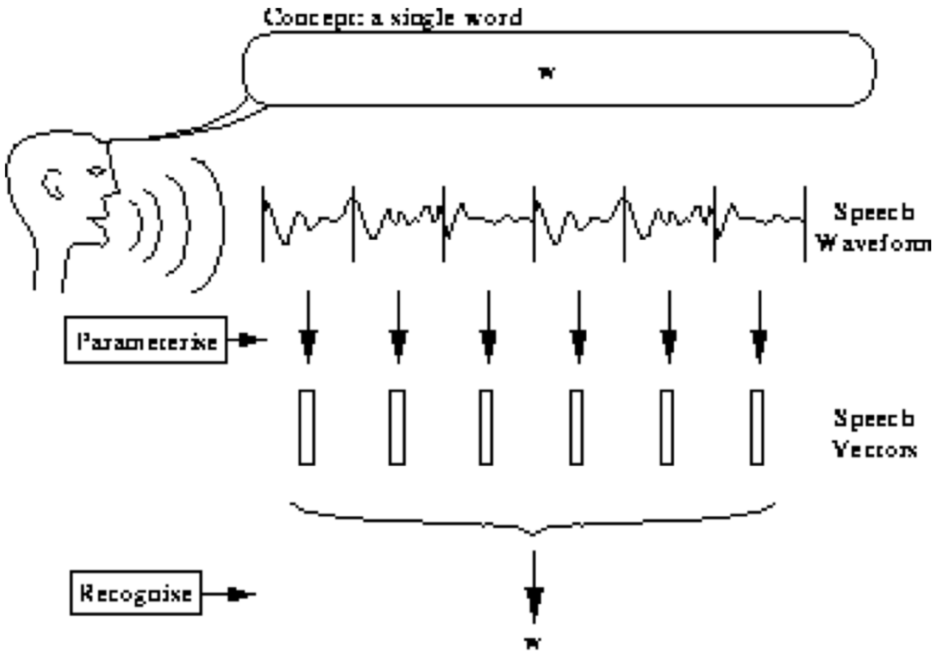
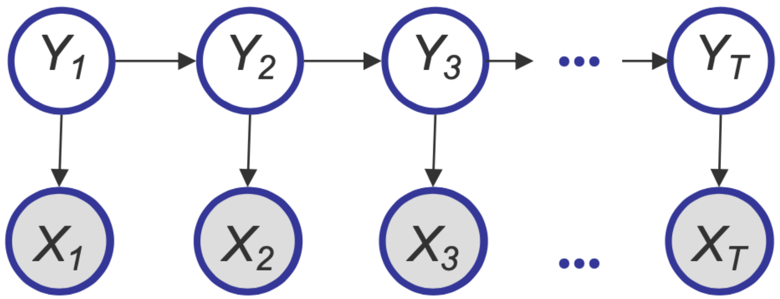


Fig. 1.2 Isolated Word Problem



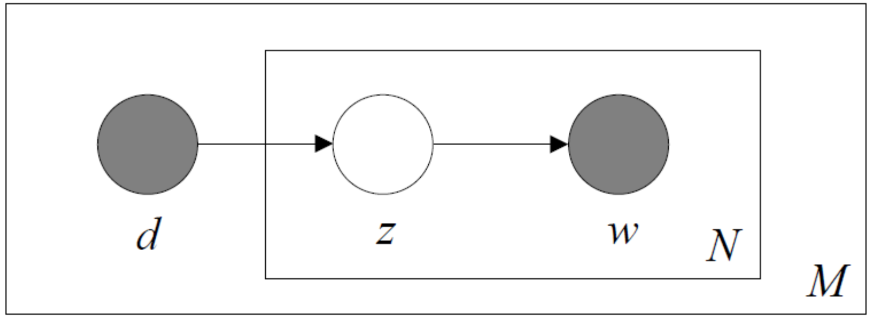
Latent (unobserved) variables

- 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

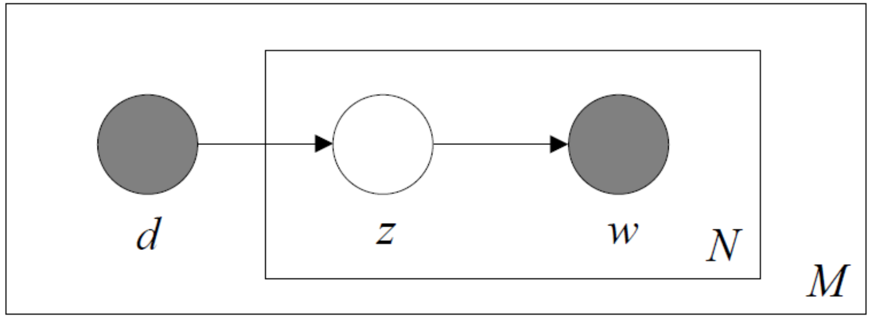


- **Likelihood function of a word w:**

$$p(w|d,\theta,\beta) = \sum_k p(w, z = k|d,\theta,\beta)$$

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

Recap: pLSA Topic Model



- **Likelihood function of a word w:**

$$p(w|d,\theta,\beta) = \sum_k p(w, z = k|d,\theta,\beta)$$

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

- **Learning by maximizing the log likelihood:**

Why is Learning Harder?

Why is Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

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

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

- But given that $\boldsymbol{z}$ is not observed, $\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})$ is a random quantity, cannot be maximized directly

Why is Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

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

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

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

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

Questions?