DSC250: Advanced Data Mining

Topic Models

Zhiting Hu

Lecture 6, October 17, 2023

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Bonus credits
  - Interacting with LMTutor for 2 bonus credits!
  - http://lmtutor.org

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

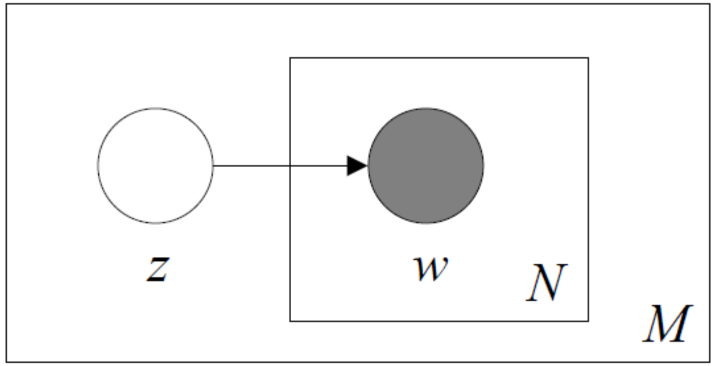
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

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

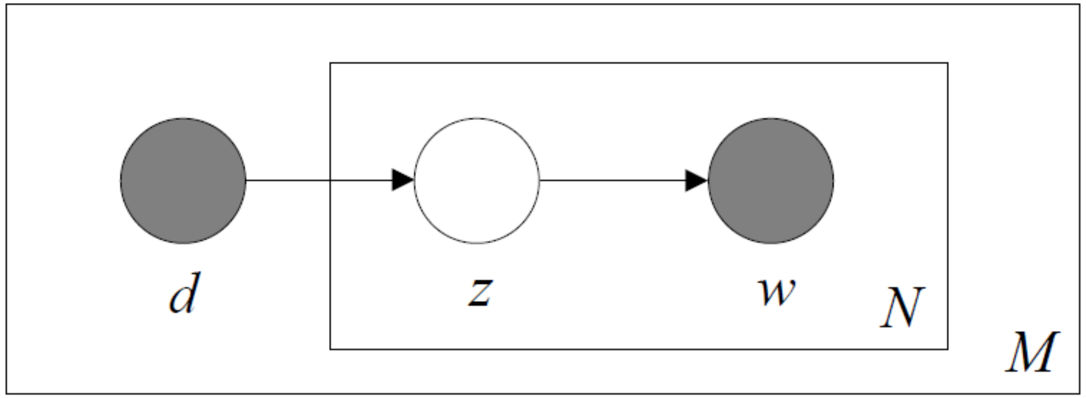
Recap: Likelihood Function

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

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

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

Recap: Generative Model for pLSA



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

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

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

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

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

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

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)

Review: Dirichlet Distribution

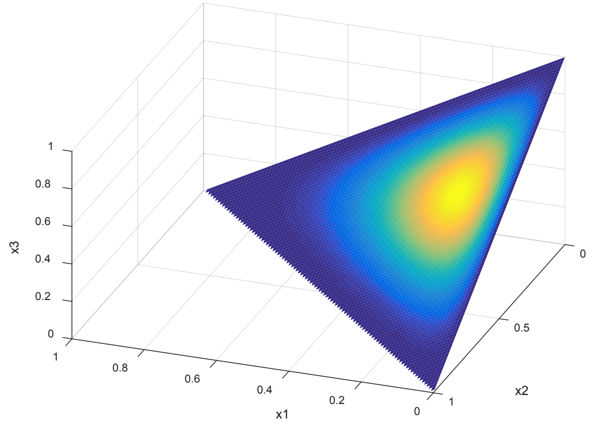
- Dirichlet distribution: $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
  - $i.e., p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$, where $\alpha_k > 0$
    - $\Gamma(\cdot)$ *is gamma function*: $\Gamma(z)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

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:

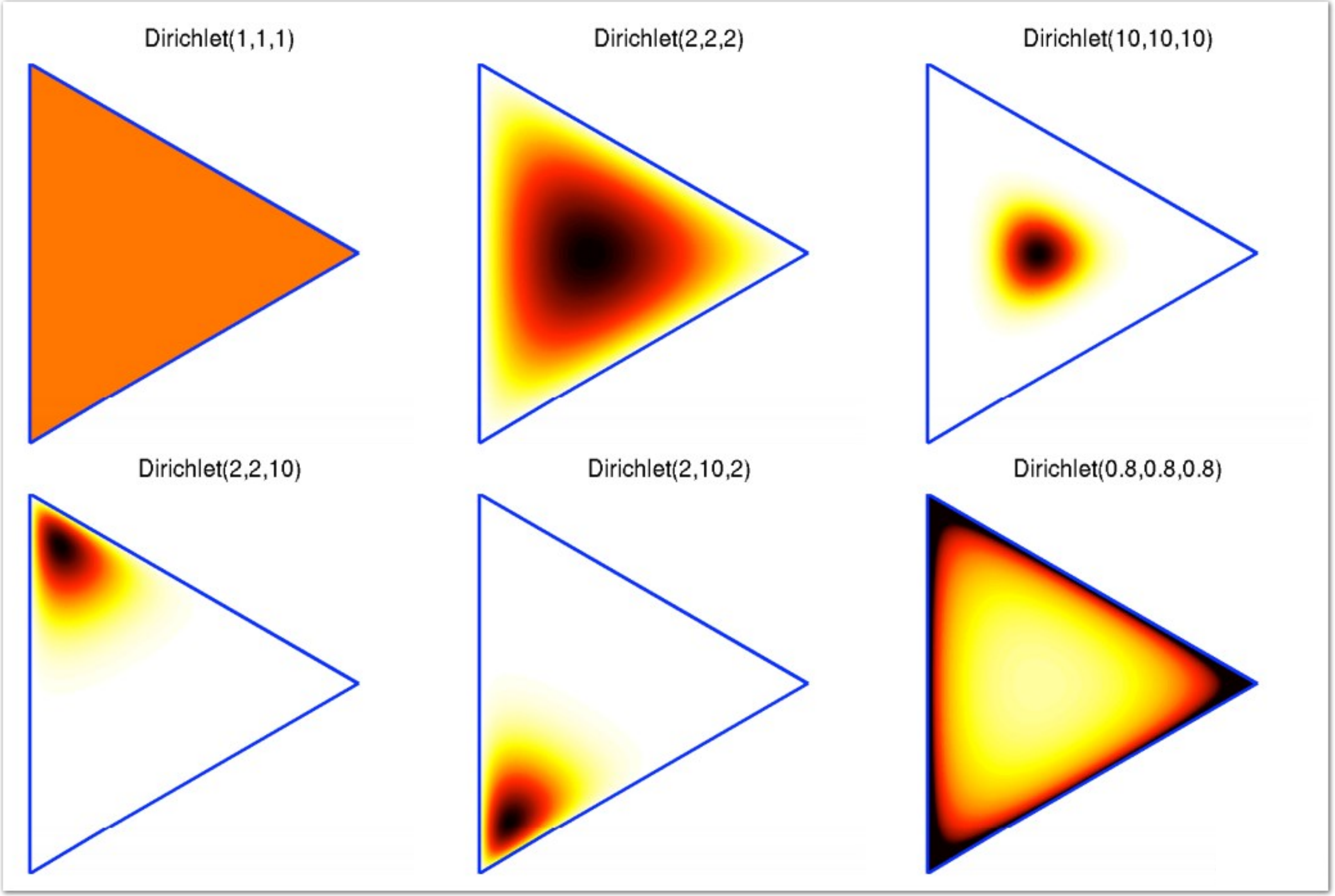
- $x = x_1(1,0,0) + x_2(0,1,0) + x_3(0,0,1)$
  - Where $0 \leq x_1, x_2, x_3 \leq 1$ *and* $x_1 + x_2 + x_3 = 1$



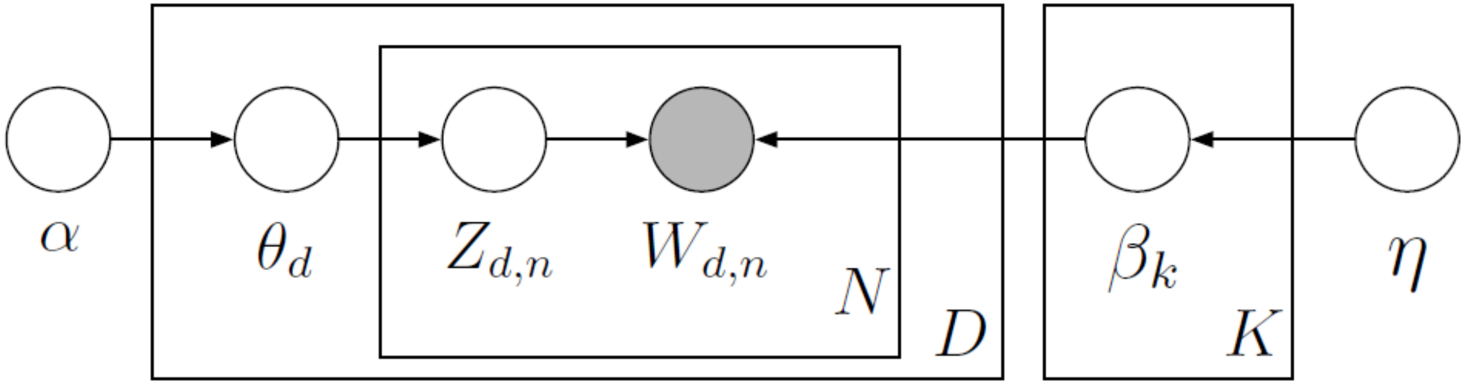
$x|\alpha \sim Dir(\alpha), \alpha = (2,3,4)$

36

More Examples in the Simplex View

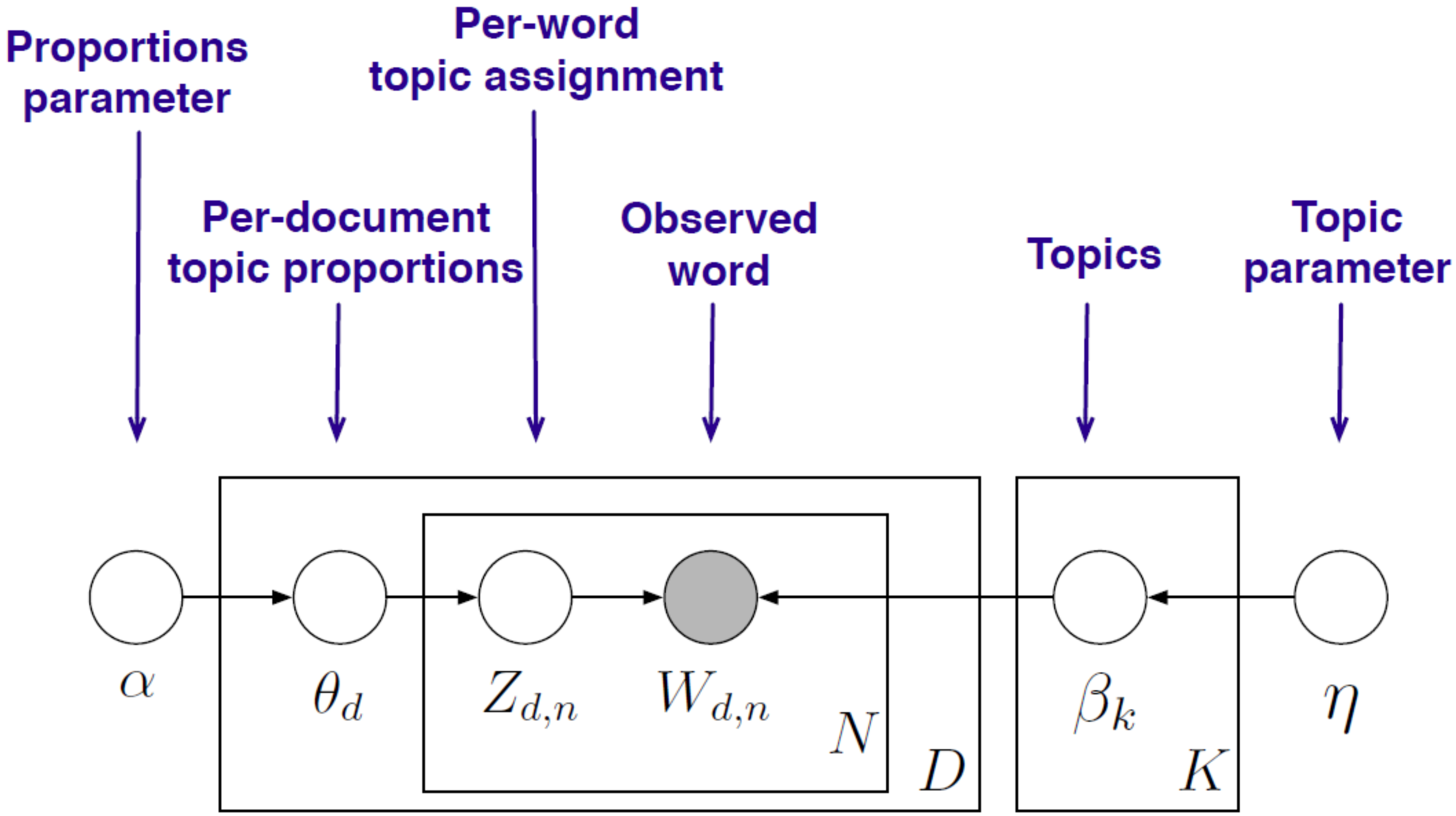


Topic Model v3: Latent Dirichlet Allocation (LDA)



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

Topic Model v3: Latent Dirichlet Allocation (LDA)

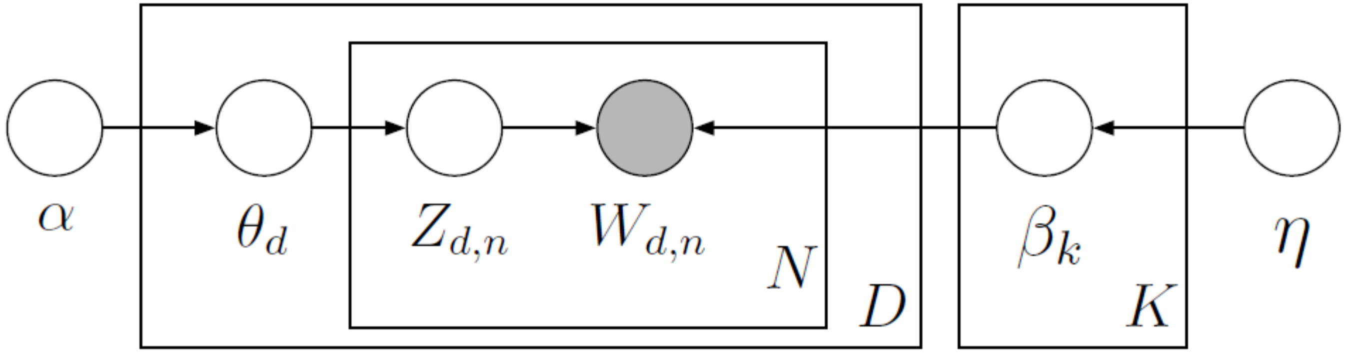


$\boldsymbol{\theta_d \sim Dirichlet(\alpha)}$: **address topic distribution for unseen documents**

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

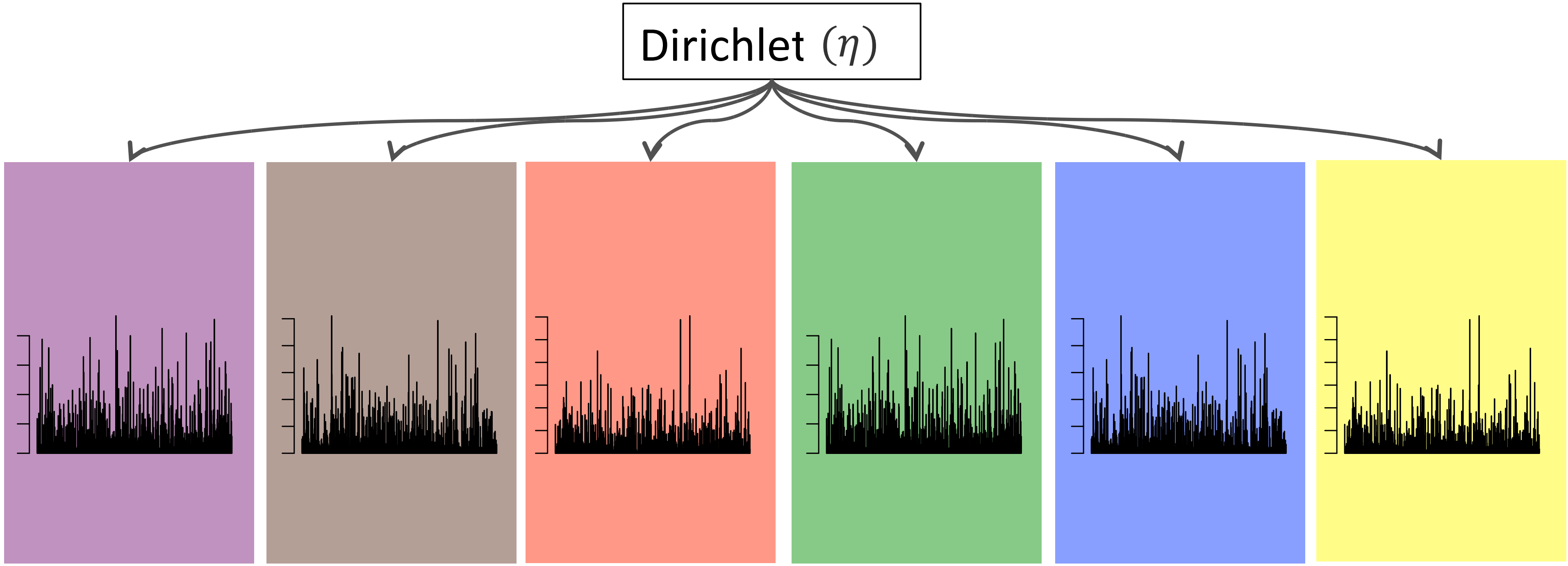
Generative Model for LDA

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*]



(Blei, Ng, & Jordan, 2003)

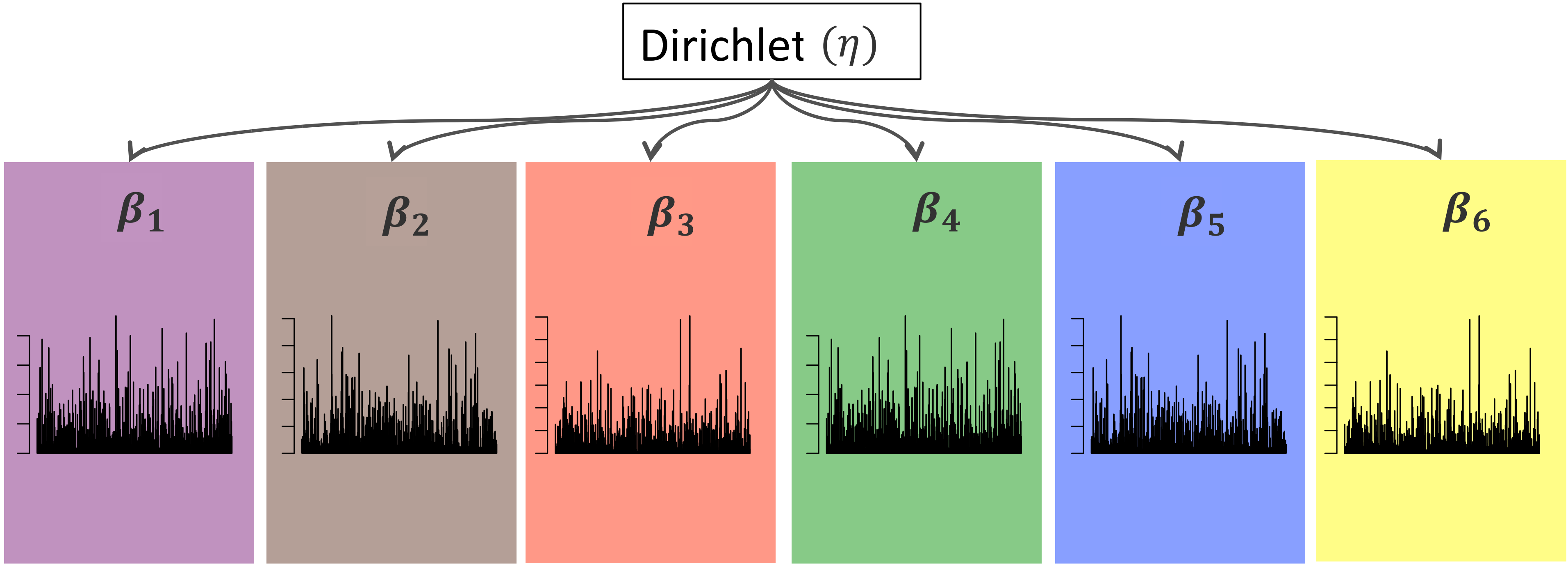
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

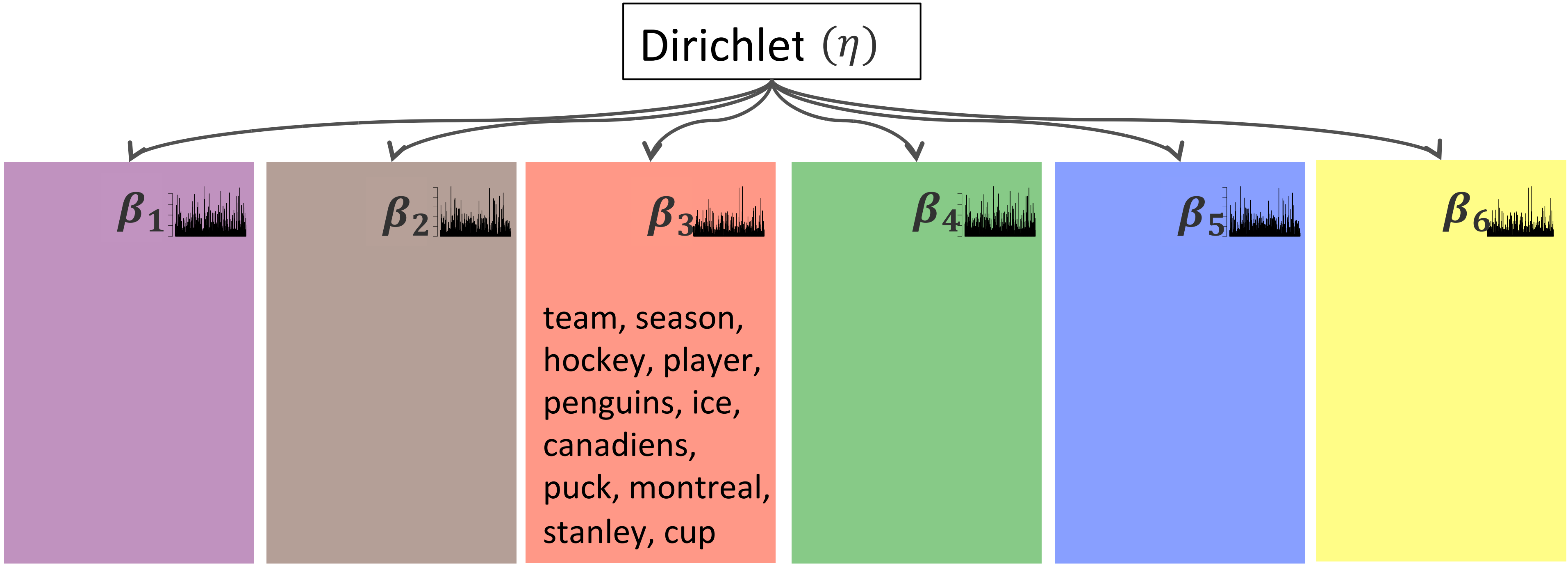
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

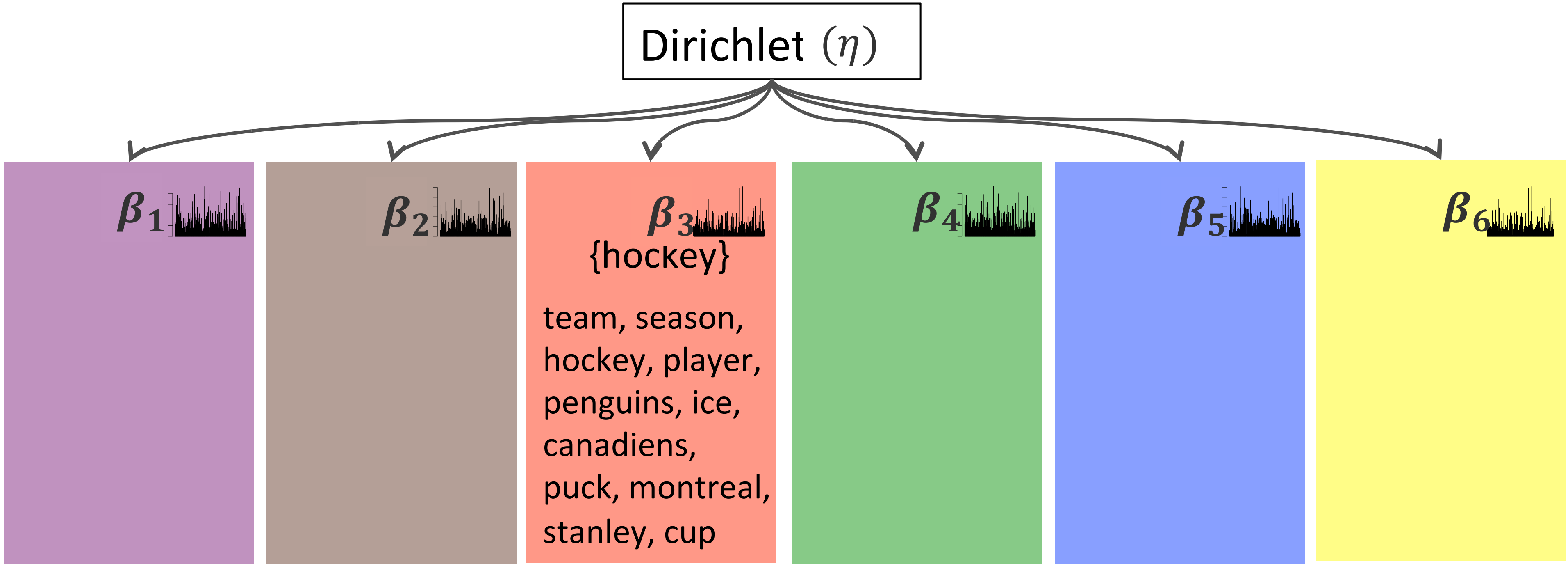
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

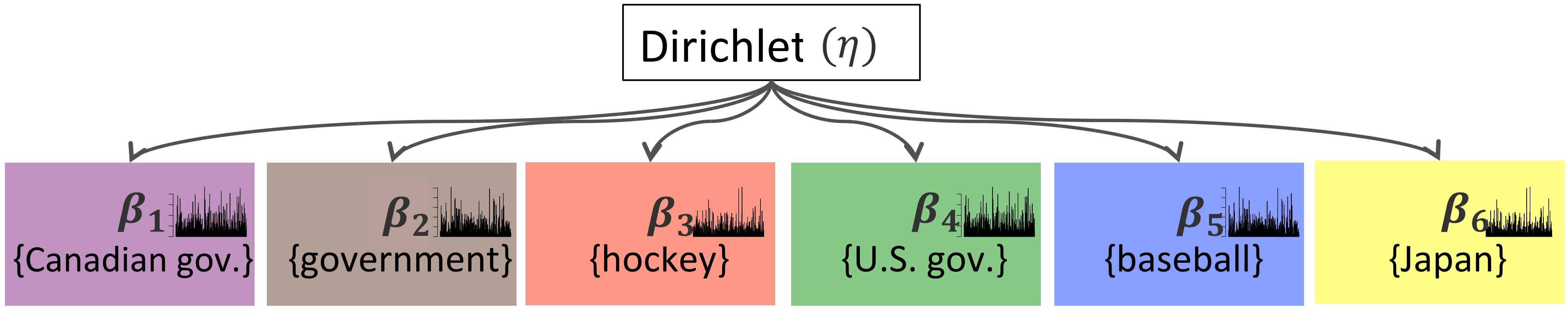
LDA for Topic Modeling



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

(Blei, Ng, & Jordan, 2003)

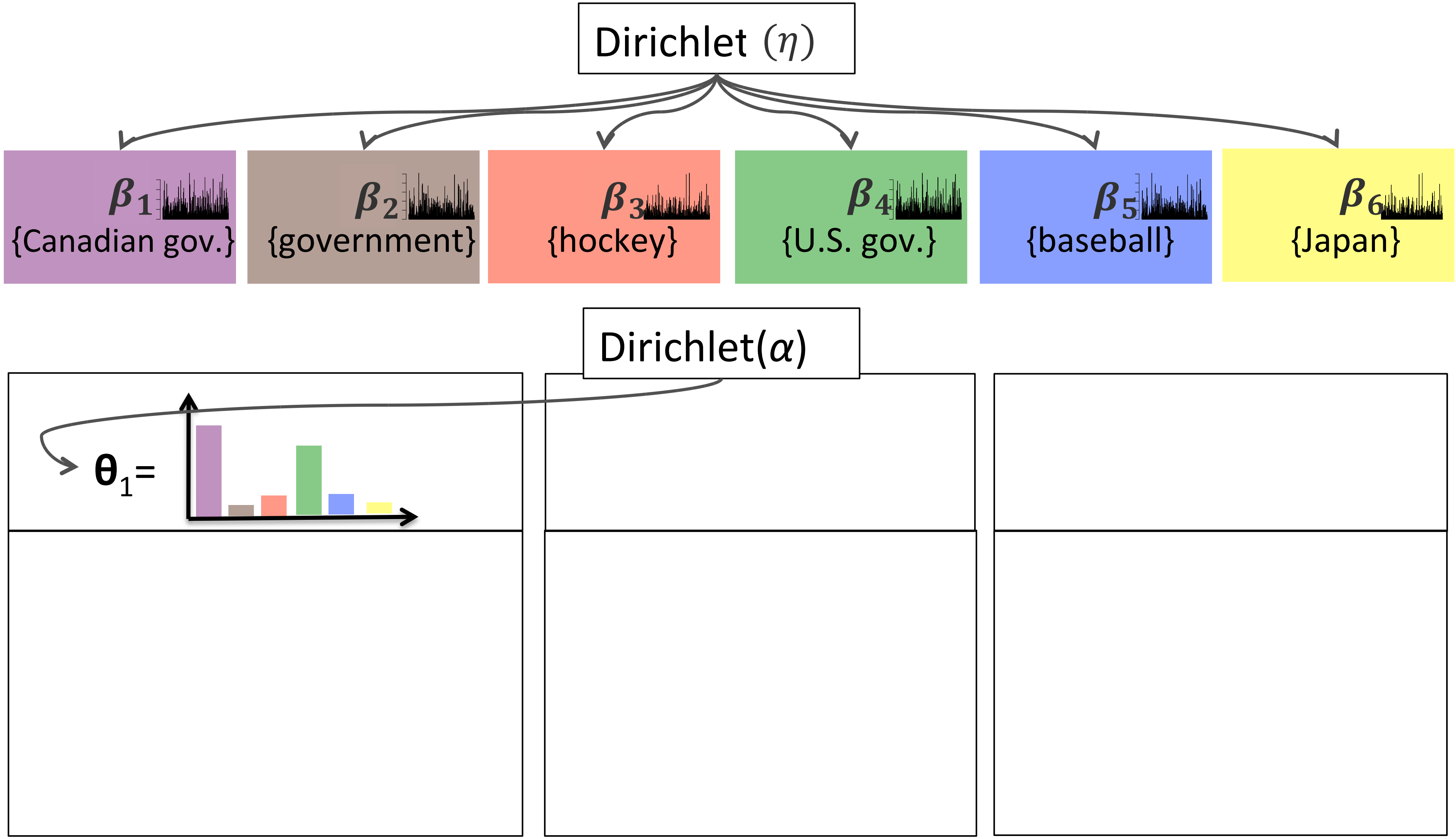
LDA for Topic Modeling



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

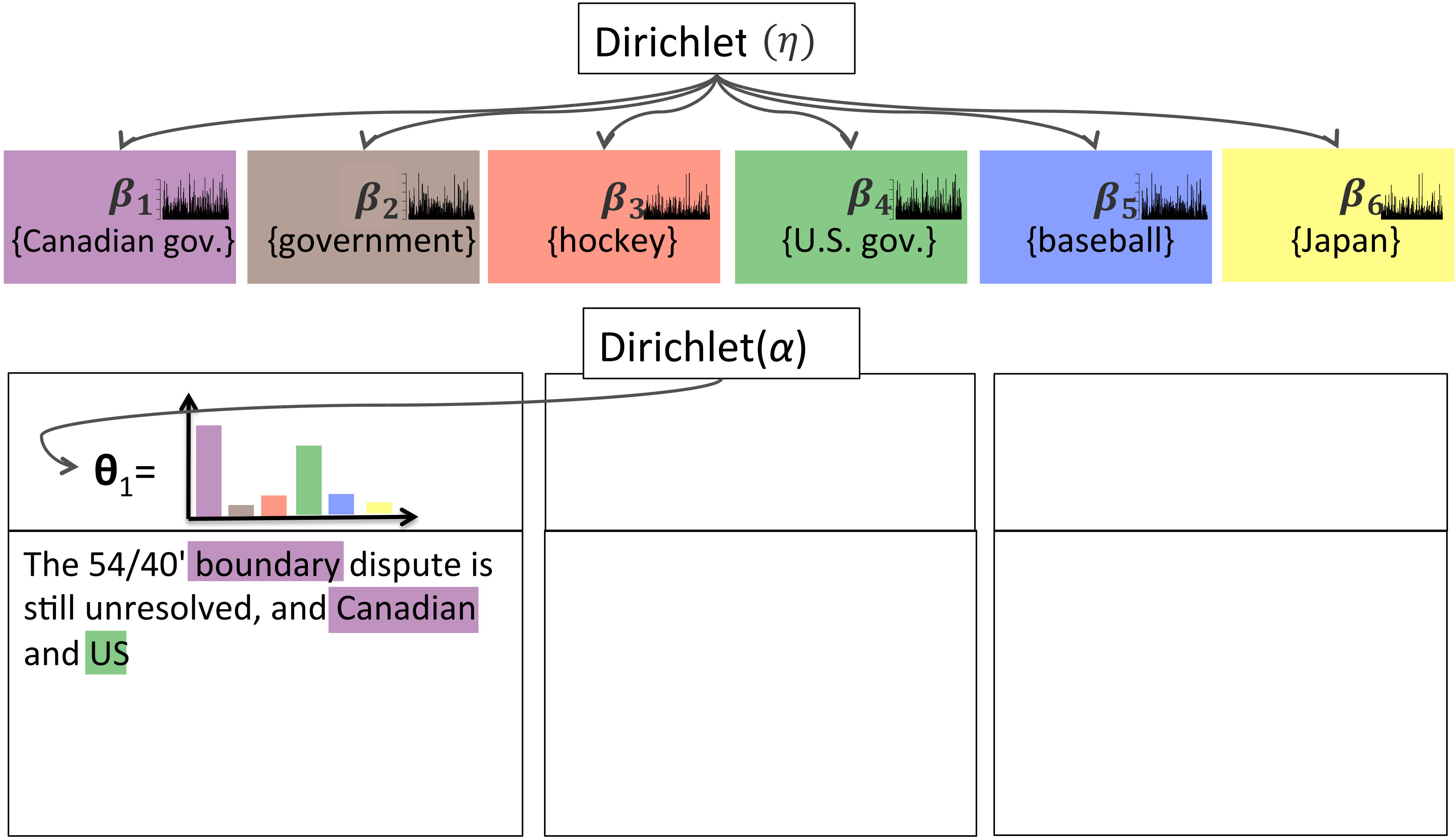
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



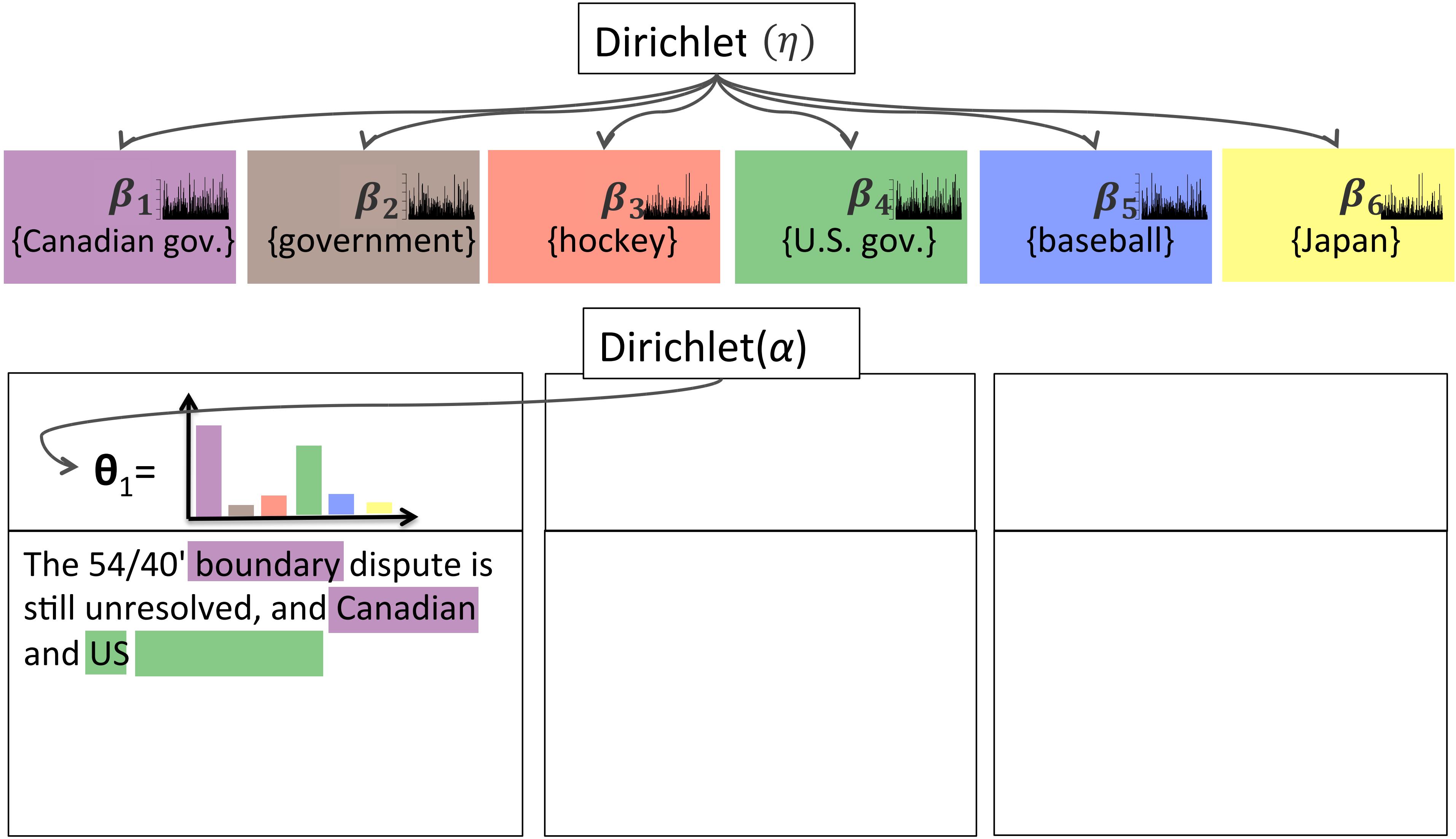
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



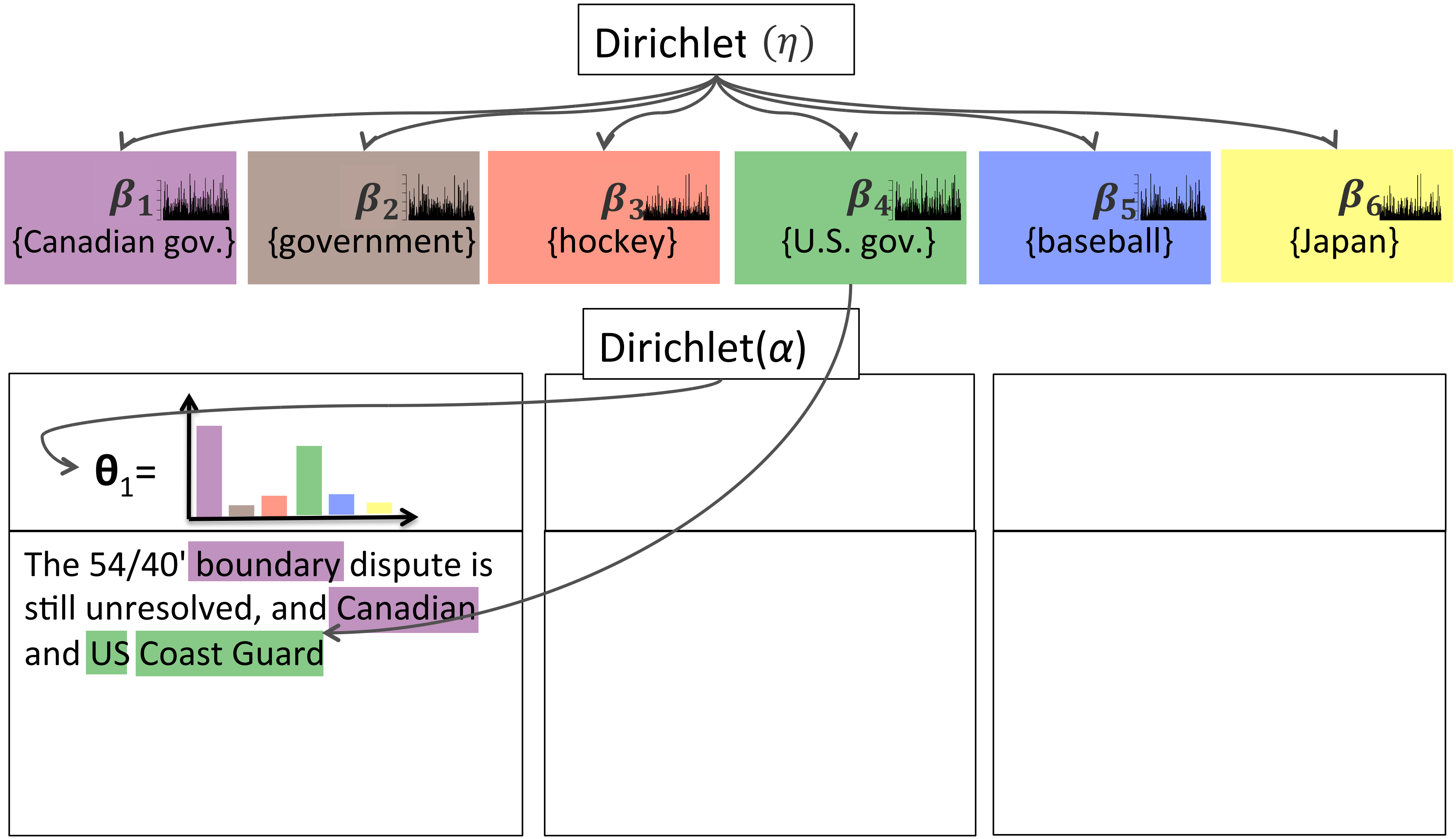
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



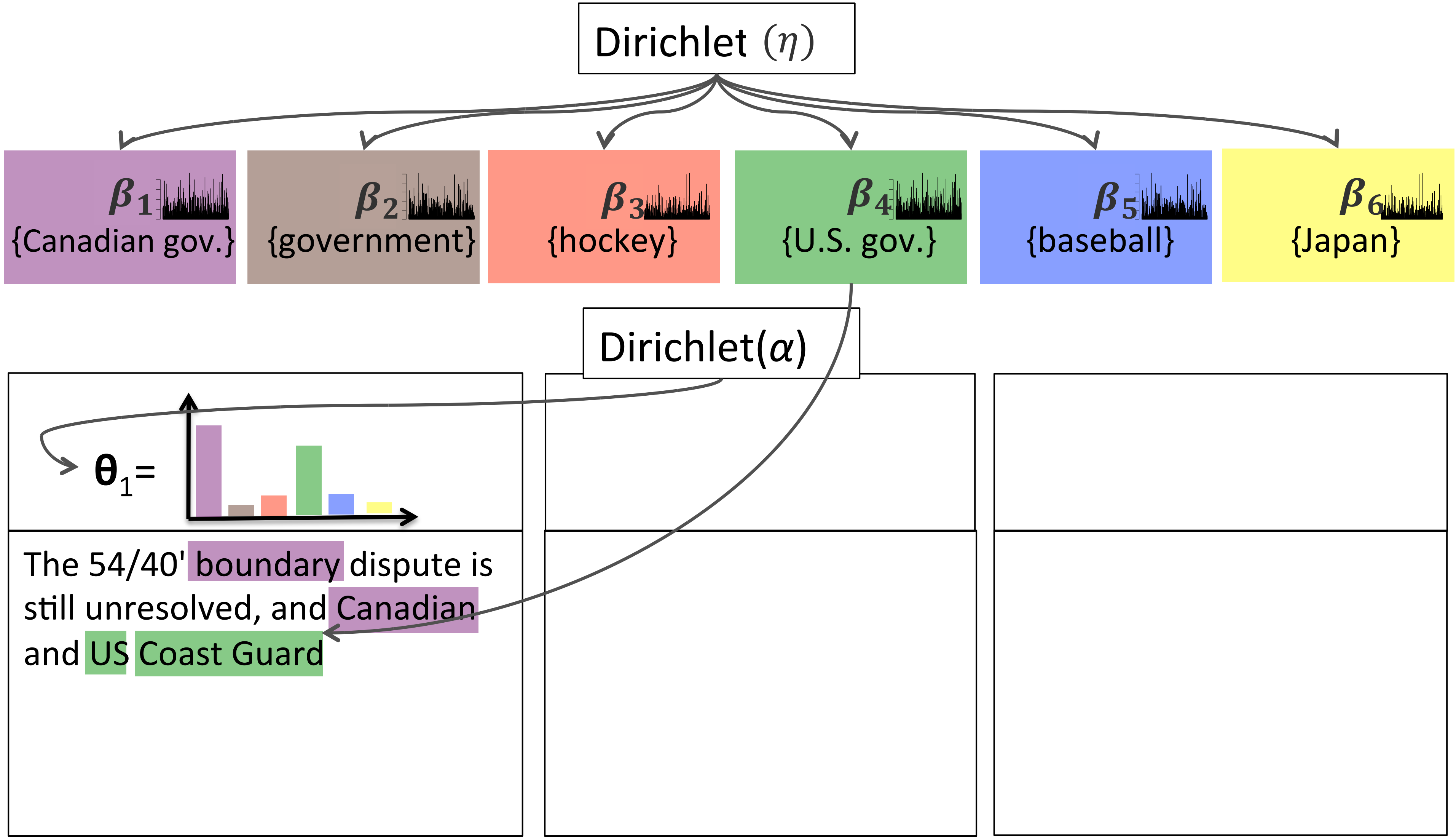
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



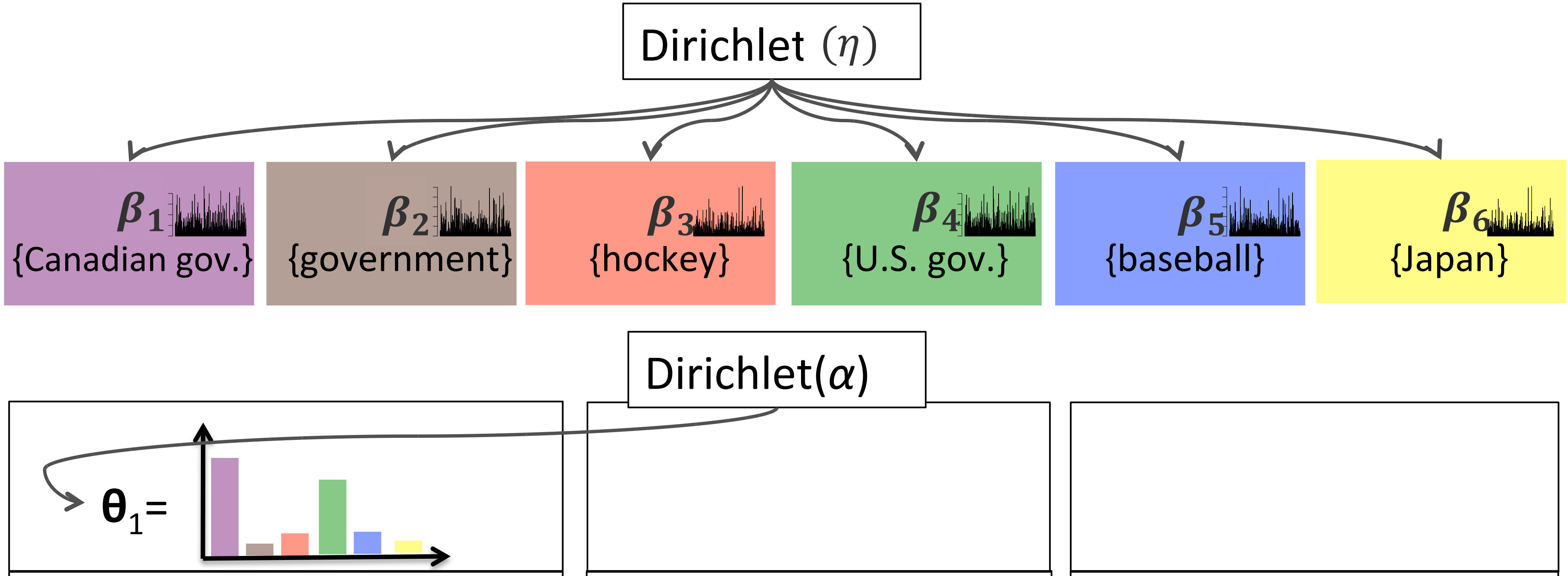
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

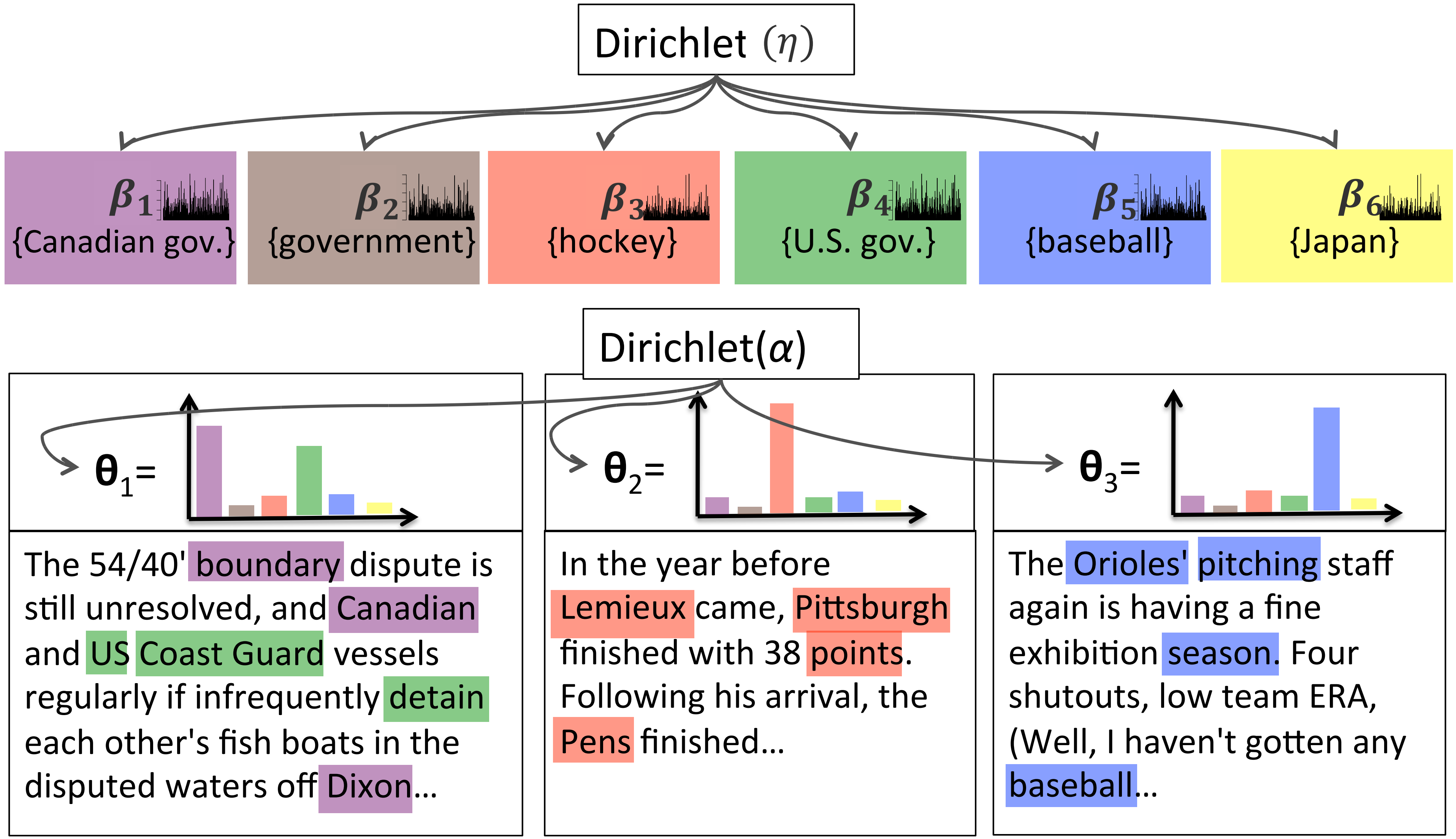
LDA for Topic Modeling



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…

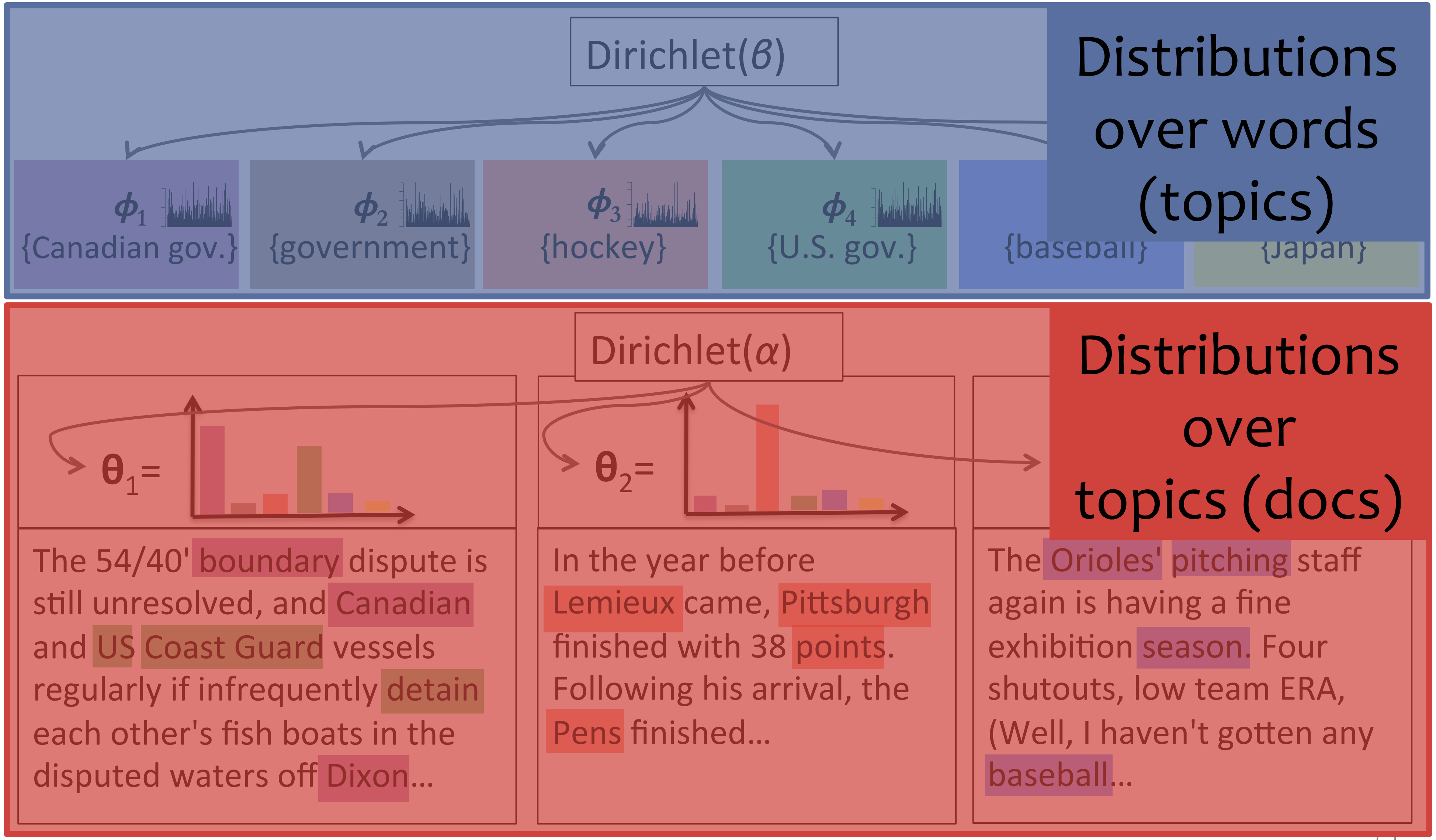
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling

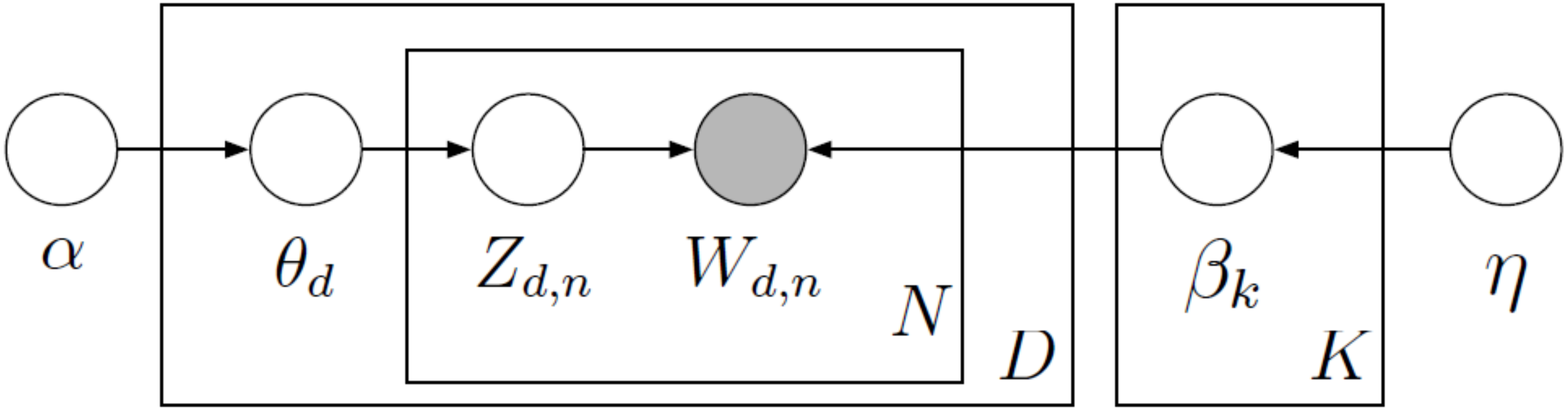


(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



Joint Distribution for LDA



- Joint distribution of latent variables and documents is:

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

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

Learning of Topic Models

Unsupervised Learning

- 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., speech recognition models, mixture models, …

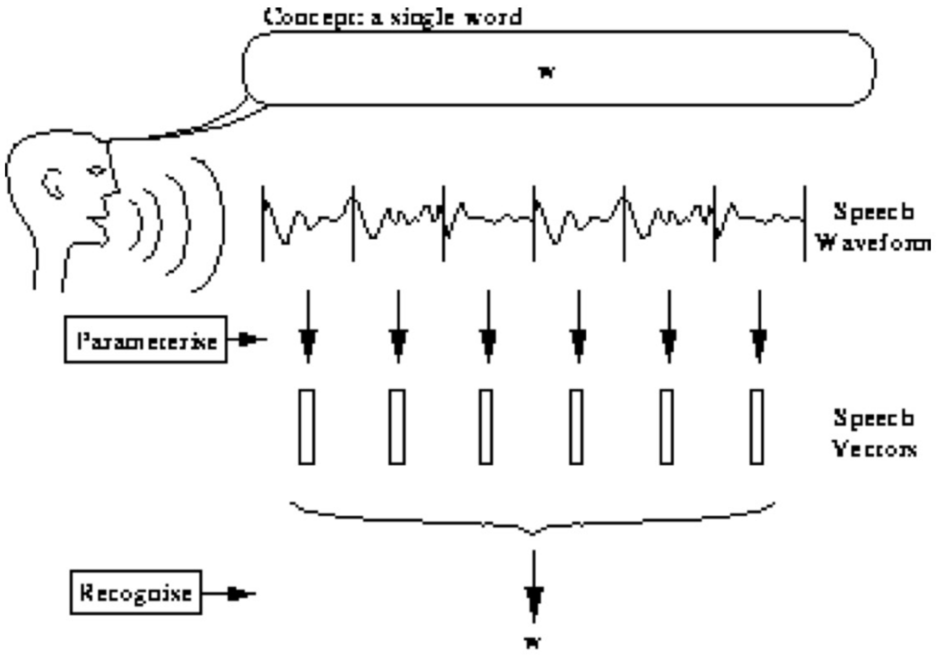
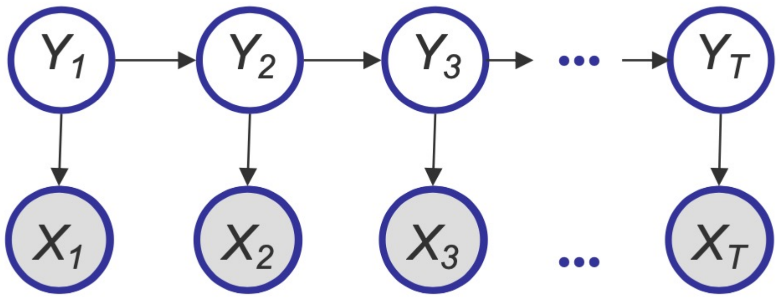


Fig. 1.2 Isolated Word Problem



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., speech recognition models, mixture models, ...

Latent (unobserved) variables

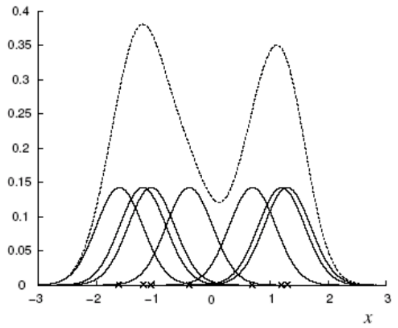
- A variable can be unobserved (latent) because:
  - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
    - e.g., speech recognition models, mixture models, …
  - a real-world object (and/or phenomena), but difficult or impossible to measure
    - 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.)

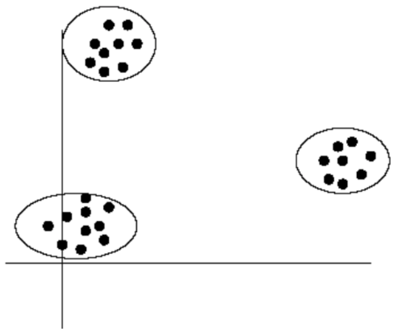
Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:

$$p(x_n|\mu,\Sigma) = \sum_k \pi_k N(x,|\,\mu_k,\Sigma_k)$$

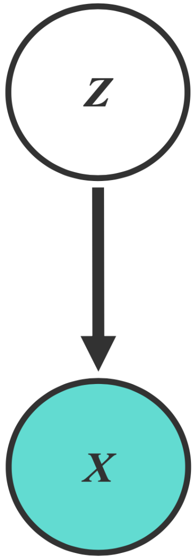
mixture proportion (π_k)   mixture component ($N(x,|\,\mu_k,\Sigma_k)$)





- This model can be used for unsupervised clustering.
  - This model (fit by AutoClass) has been used to discover new kinds of stars in astronomical data, etc.

Example: Gaussian Mixture Models (GMMs)



- Consider a mixture of K Gaussian components:
  - Z is a latent class indicator vector:

$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - X is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n \mid z_n^k = 1, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2}|\Sigma_k|^{1/2}} \exp\left\{-\tfrac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k)\right\}$$

  - The likelihood of a sample:

Parameters to be learned:

mixture component

mixture proportion

$$\begin{aligned} p(x_n|\mu,\Sigma) &= \sum_k p(z^k = 1 \mid \pi)\, p(x, \mid z^k = 1, \mu, \Sigma) \\ &= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k}\right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components: $p(x_n|\mu,\Sigma)=\sum_k \pi_k N(x,|\mu_k,\Sigma_k)$
- Recall MLE for completely observed data
  - Data log-likelihood: $\ell(\boldsymbol{\theta};D)=\log\prod_n p(z_n,x_n)=\log\prod_n p(z_n\mid\pi)p(x_n\mid z_n,\mu,\sigma)$

$$=\sum_n\log\prod_k \pi_k^{z_n^k}+\sum_n\log\prod_k N(x_n;\mu_k,\sigma)^{z_n^k}$$

$$=\sum_n\sum_k z_n^k\log\pi_k-\sum_n\sum_k z_n^k\frac{1}{2\sigma^2}(x_n-\mu_k)^2+C$$

  - MLE:

$$\hat{\pi}_{k,MLE}=\arg\max_\pi \ell(\boldsymbol{\theta};D),$$

$$\hat{\mu}_{k,MLE}=\arg\max_\mu \ell(\boldsymbol{\theta};D)\qquad\Rightarrow\quad\hat{\mu}_{k,MLE}=\frac{\sum_n z_n^k x_n}{\sum_n z_n^k}$$

$$\hat{\sigma}_{k,MLE}=\arg\max_\sigma \ell(\boldsymbol{\theta};D)$$

- What if we do not know z_n?

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 (as we'll see later), marginalization over $\boldsymbol{z}$ is intractable.

Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{z}|\theta)$$

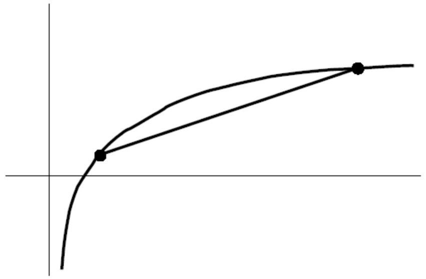
  - A deterministic function of θ
  - Inherit the factorizability of $\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})$
- Use this as the surrogate objective
- Does maximizing this surrogate yield a maximizer of the likelihood?

Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{z}|\theta)$$

- Jensen's inequality



$$\begin{aligned}
\ell(\theta;x) &= \log p(x\mid\theta) \\
&= \log \sum_z p(x,z\mid\theta) \\
&= \log \sum_z q(z\mid x)\frac{p(x,z\mid\theta)}{q(z\mid x)} \\
&\geq \sum_z q(z\mid x)\log\frac{p(x,z\mid\theta)}{q(z\mid x)} \qquad \text{Evidence Lower Bound (ELBO)} \\
&= \sum_z q(z\mid x)\log p(x,z\mid\theta) - \sum_z q(z\mid x)\log q(z\mid x) \\
&= \mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] + H(q)
\end{aligned}$$

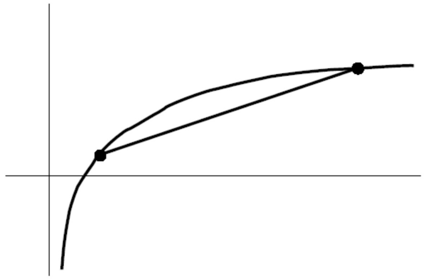
Expectation Maximization (EM)

- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})] = \sum_z q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

- Jensen's inequality

$$\begin{aligned}\ell(\theta; x) &= \log p(x \mid \theta) \\ &= \log \sum_z p(x, z \mid \theta) \\ &= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\ &\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)}\end{aligned}$$



- Indeed we have

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

Lower Bound and Free Energy

- For fixed data $\boldsymbol{x}$, define a functional called the (variational) free energy:

$$F(q,\theta) = -\mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] - H(q) \geq \ell(\theta;\boldsymbol{x})$$

- The EM algorithm is coordinate-decent on F
  - At each step t:
    - E-step: $$q^{t+1} = \arg\min_q F\left(q, \theta^t\right)$$
    - M-step: $$\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right)$$

E-step: minimization of $F(q, \theta)$ w.r.t q

- Claim:

$$q^{t+1} = \text{argmin}_q \, F(q, \theta^t) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$$

  - This is the posterior distribution over the latent variables given the data and the current parameters.

- Proof (easy): recall

$$\ell(\theta^t; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta^t)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)\big)$$

Independent of q | $-F(q, \theta^t)$ | ≥ 0

  - $F(q, \theta^t)$ is minimized when $\text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)\big) = 0$, which is achieved only when $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$

M-step: minimization of $F(q, \theta)$ w.r.t $\boldsymbol{\theta}$

- Note that the free energy breaks into two terms:

$$F(q,\theta) = -\mathbb{E}_q[\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})] - H(q) \geq \ell(\theta; \boldsymbol{x})$$

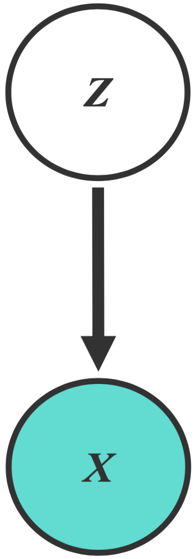
  - The first term is the expected complete log likelihood and the second term, which does not depend on q, is the entropy.

- Thus, in the M-step, maximizing with respect to θ for fixed q we only need to consider the first term:

$$\theta^{t+1} = \mathrm{argmax}_\theta \, \mathbb{E}_q[\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z})] = \mathrm{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

  - Under optimal q^{t+1}, this is equivalent to solving a standard MLE of fully observed model $p(\boldsymbol{x}, \boldsymbol{z}|\theta)$, with z replaced by its expectation w.r.t $p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$

Example: Gaussian Mixture Models (GMMs)



- Consider a mixture of K Gaussian components:
  - Z is a latent class indicator vector:

$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

  - X is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n \mid z_n^k = 1, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{ -\tfrac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right\}$$

  - The likelihood of a sample:

mixture component

mixture proportion

$$\begin{aligned} p(x_n | \mu, \Sigma) &= \sum_k p(z^k = 1 \mid \pi) p(x, \mid z^k = 1, \mu, \Sigma) \\ &= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k} \right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components
- The expected complete log likelihood

$$\mathbb{E}_q\left[\ell_c(\boldsymbol{\theta}; x, z)\right] = \sum_n \mathbb{E}_q\left[\log p\left(z_n \mid \pi\right)\right] + \sum_n \mathbb{E}_q\left[\log p\left(x_n \mid z_n, \mu, \Sigma\right)\right]$$

$$= \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \log \pi_k - \frac{1}{2} \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log|\Sigma_k| + C \right)$$

- E-step: computing the posterior of z_n given the current estimate of the parameters (i.e., π , μ, Σ)

$$p(z_n^k = 1 \mid x, \mu^{(t)}, \Sigma^{(t)}) = \frac{\pi_k^{(t)} N(x_n, \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_i \pi_i^{(t)} N(x_n, \mid \mu_i^{(t)}, \Sigma_i^{(t)})}$$

Numerator → $p(z_n^k = 1, x, \mu^{(t)}, \Sigma^{(t)})$

Denominator → $p(x, \mu^{(t)}, \Sigma^{(t)})$

Example: Gaussian Mixture Models (GMMs)

- M-step: computing the parameters given the current estimate of z_n

$$\pi_k^* = \arg\max \langle l_c(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \frac{\partial}{\partial \pi_k} \langle l_c(\boldsymbol{\theta}) \rangle = 0, \forall k, \quad \text{s.t.} \sum_{\text{k}} \pi_k = 1$$

$$\Rightarrow \quad \pi_k^* = \sum_n \langle z_n^k \rangle_{q^{(t)}} \Big/ N = \sum_n \tau_n^{k(t)} \Big/ N = \langle n_k \rangle \Big/ N$$

$$\mu_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

$$\Sigma_k^* = \arg\max \langle l(\boldsymbol{\theta}) \rangle, \qquad \Rightarrow \quad \Sigma_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \tau_n^{k(t)}}$$

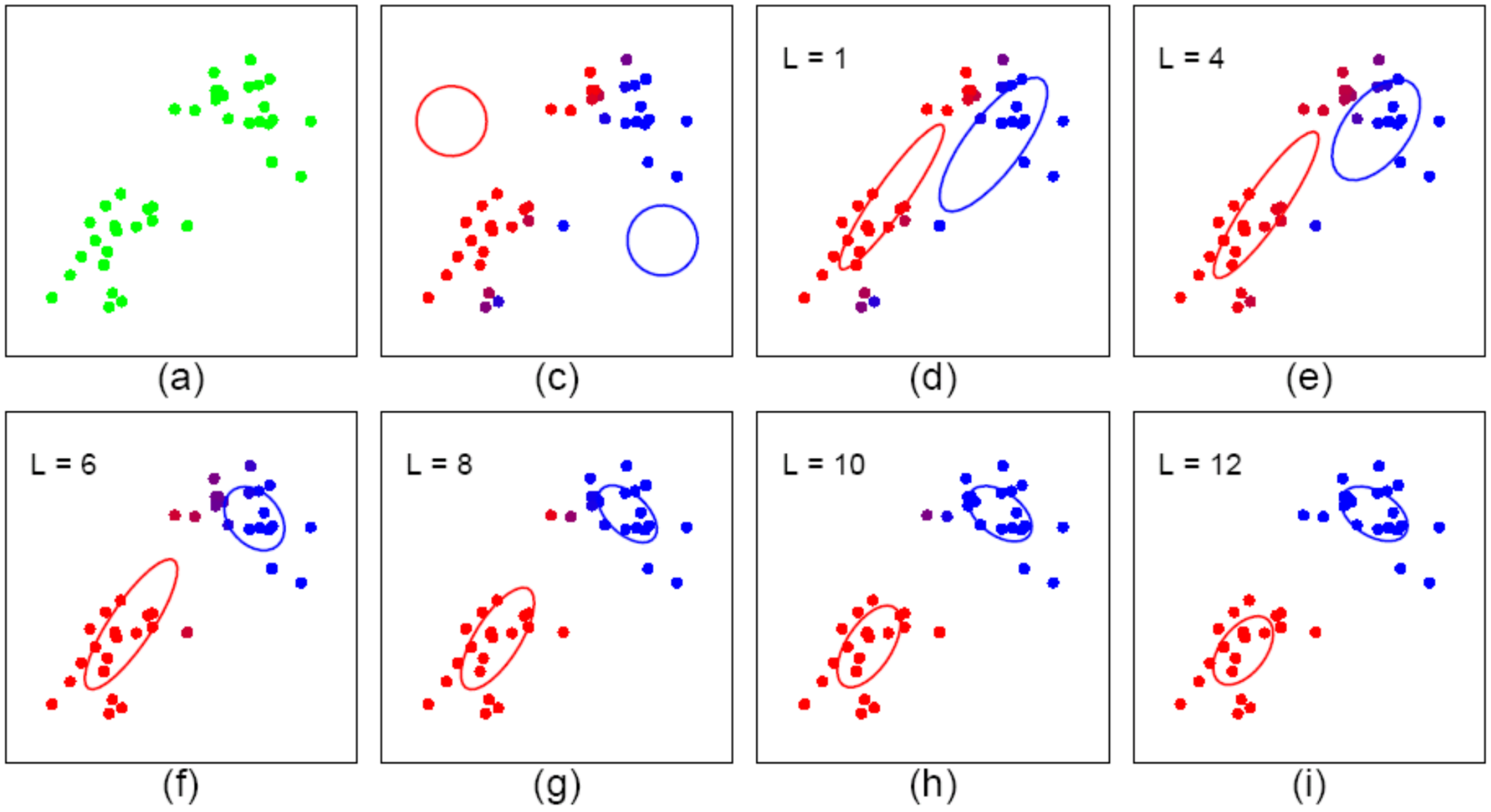
Fact:

$$\frac{\partial \log |\mathrm{A}^{-1}|}{\partial \mathrm{A}^{-1}} = \mathrm{A}^T$$

$$\frac{\partial \mathbf{x}^T \mathrm{A} \mathbf{x}}{\partial \mathrm{A}} = \mathbf{x}\mathbf{x}^T$$

Example: Gaussian Mixture Models (GMMs)

- Start: "guess" the centroid μ_k and covariance Σ_k of each of the K clusters
- Loop:



Summary: EM Algorithm

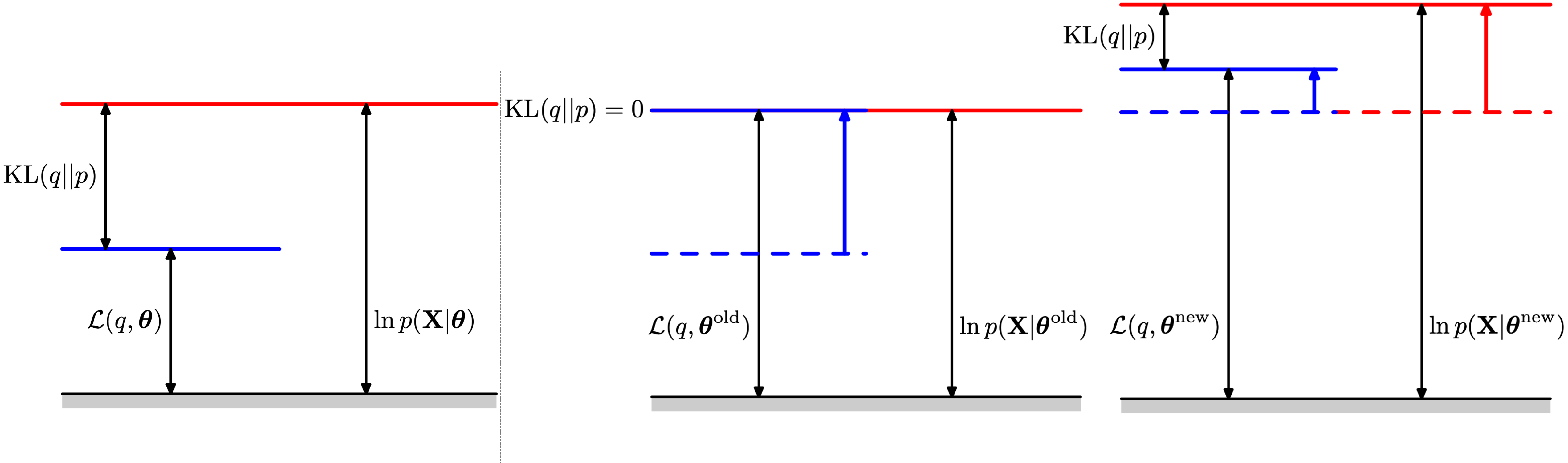
- A way of maximizing likelihood function for latent variable models. Finds MLE of parameters when the original (hard) problem can be broken up into two (easy) pieces
  - Estimate some "missing" or "unobserved" data from observed data and current parameters.
  - Using this "complete" data, find the maximum likelihood parameter estimates.

- Alternate between filling in the latent variables using the best guess (posterior) and updating the parameters based on this guess:

  - E-step: $q^{t+1} = \arg\min_{q} F\left(q, \theta^{t}\right)$
  - M-step: $\theta^{t+1} = \arg\min_{\theta} F\left(q^{t+1}, \theta^{t}\right)$

Each EM iteration guarantees to improve the likelihood

$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$



E-step

M-step

[PRML, Chap 9.4]

EM Variants

- Sparse EM
  - Do not re-compute exactly the posterior probability on each data point under all models, because it is almost zero.
  - Instead keep an “active list” which you update every once in a while.

- Generalized (Incomplete) EM:
  - It might be hard to find the ML parameters in the M-step, even given the completed data. We can still make progress by doing an M-step that improves the likelihood a bit (e.g. gradient step).

Questions?