# Reasoning (III)

Mingsheng Long

Tsinghua University

Spring 2023

# Outline

- **Probabilistic Topic Model**
  - **Dirichlet-Multinomial Model**
  - Latent Dirichlet Allocation
  - Document Generation Process
- Variational Inference
  - Mean Field Variational Inference
  - Variational Inference for LDA
  - Parameter Estimation for LDA

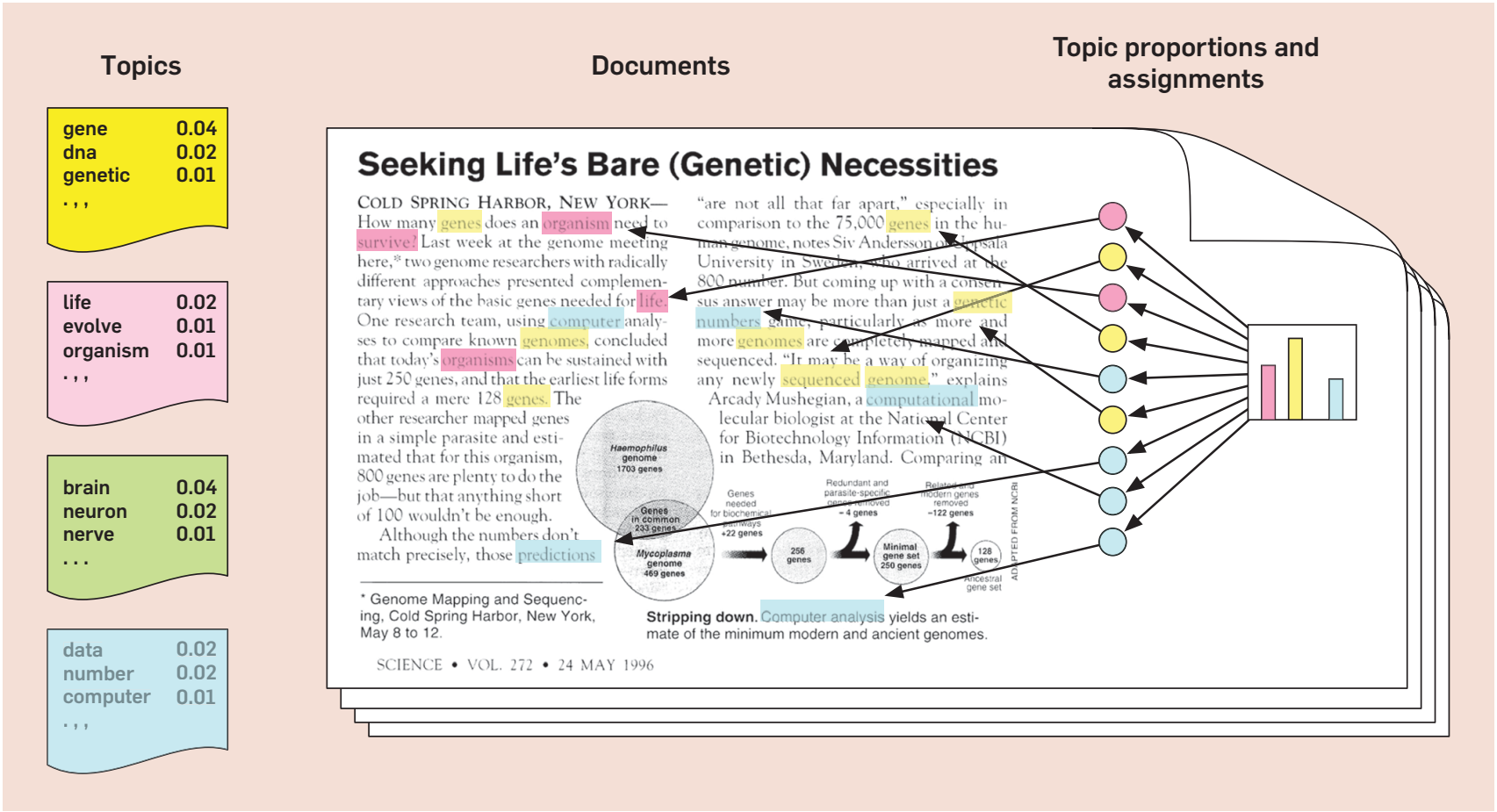
# Content Analysis

- Suppose you're given a massive corpora and asked to carry out the following tasks:
  - Organize the documents into thematic categories
  - Enable a domain expert to analyze and understand the content
  - Find relationships between the categories
  - Understand how authorship influences the content…

- Key problem: How documents are generated? Which distribution?

# Probabilistic Topic Model

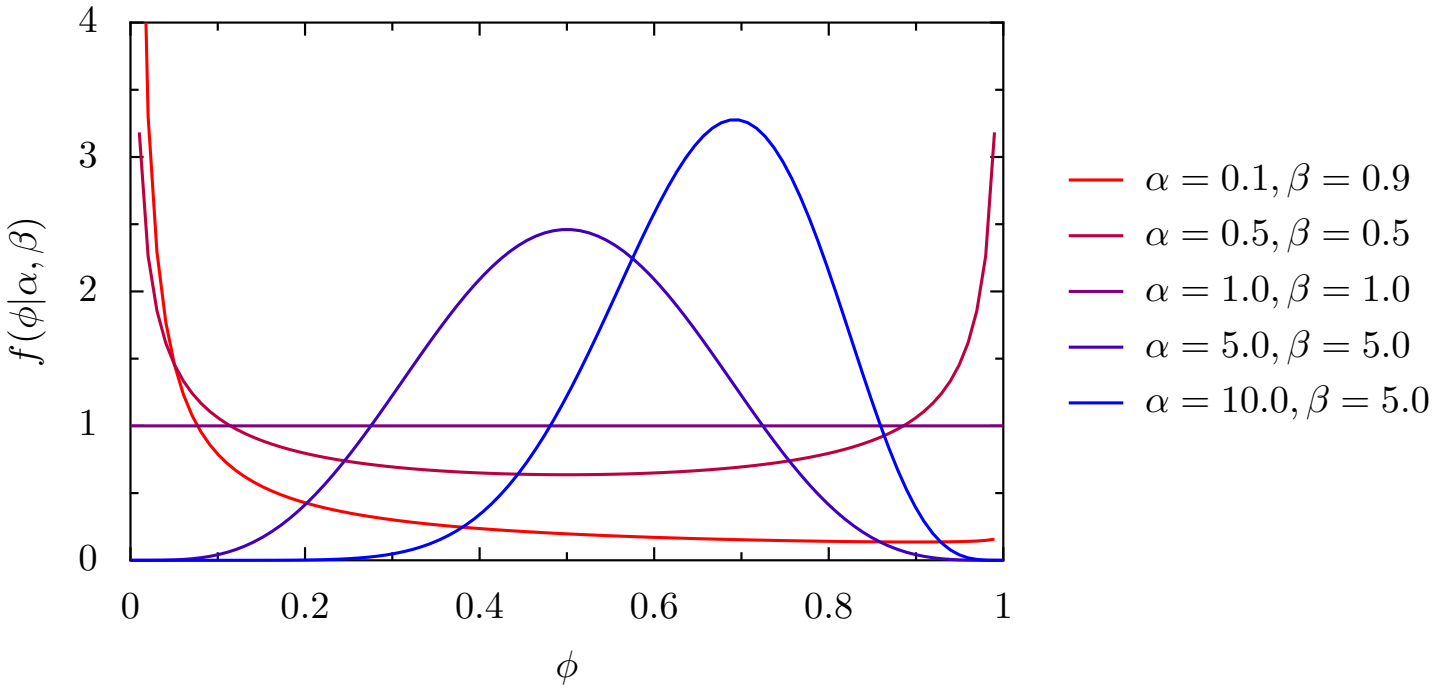
- The answer to content analysis is probabilistic topic model (PTM).



# Beta-Bernoulli Model

- Beta Distribution: conjugate prior for Bernoulli distribution

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$



# Beta-Bernoulli Model

- How are documents generated?
- Generative Process

$\phi \sim \text{Beta}(\alpha, \beta)$ [*draw distribution over words*]
For each word $n \in \{1, \ldots, N\}$
$x_n \sim \text{Bernoulli}(\phi)$ [*draw word*]

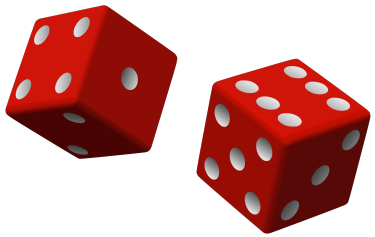
- Example corpus (heads/tails)

| H | T | T | H | H | T | T | H | H | H |
|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ |

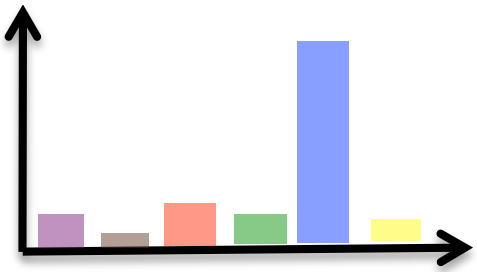
- There are only two words? We need more! Bernoulli is not enough.

# Multinomial Distribution

- Multinomial Distribution models the probability of counts of each side for rolling a $k$-sided dice $n$ times.
  - Each side equips with a probability $\theta_i$.
    - $\sum_{i=1}^{k} \theta_i = 1$.
  - Write by $\mathbf{Mult}(n, \boldsymbol{\theta})$
  - $p(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k}$
    - $\sum_{i=1}^{k} n_i = n$
- When $n = 1$, it is Categorical Distribution, $\mathbf{Cat}(\boldsymbol{\theta}) = \mathbf{Mult}(1, \boldsymbol{\theta})$



Flipping dice:
$\sum_{l=1}^{6} \theta_l = 1$

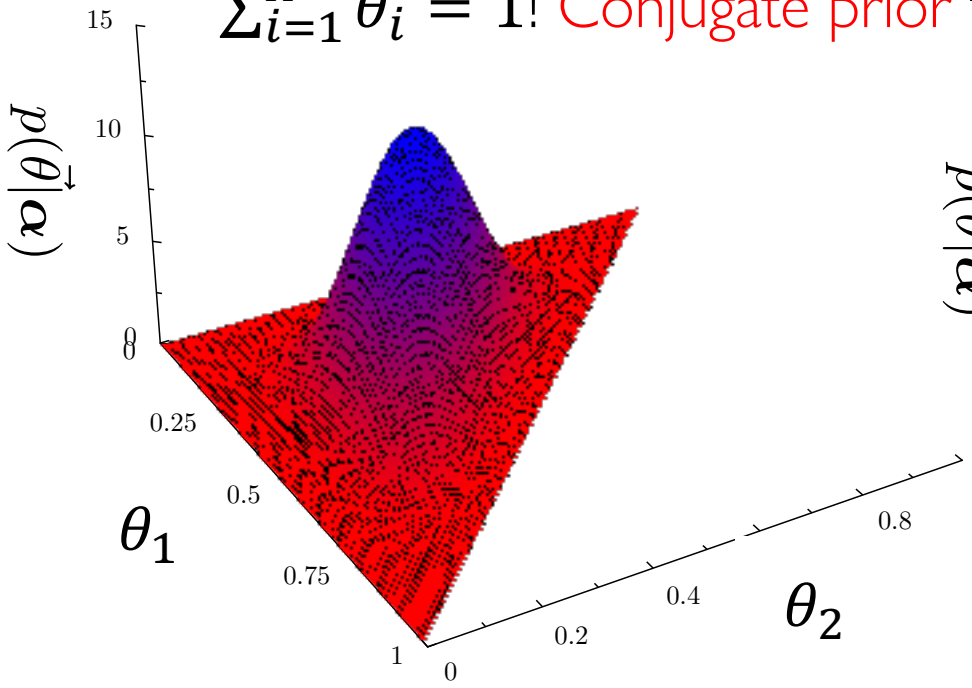


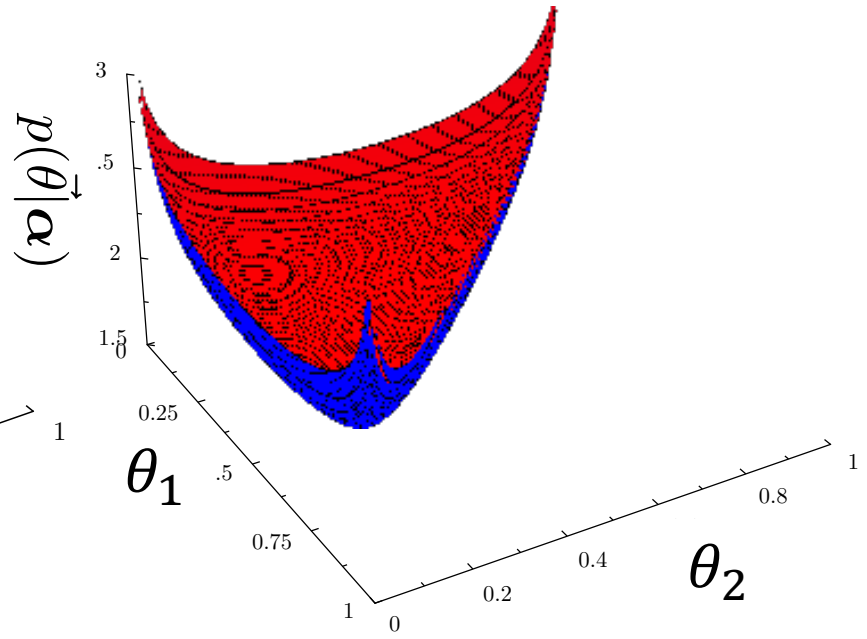
# Dirichlet-Multinomial Model

- Dirichlet Distribution: Multi-dimensional version of Beta Distribution

$$p(\vec{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \quad \text{Where } B(\alpha) = \frac{\Pi_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

$\sum_{i=1}^{K} \theta_i = 1$! Conjugate prior for Multinomial Distribution.



# Dirichlet-Multinomial Model

- Assume that we have a vocabulary of $V$ words.
- Generative Process

What is the dimension of $\boldsymbol{\theta}$ here?

$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ [draw distribution over words]
For each word $n \in \{1, \ldots, N\}$ [draw word]
$w_n \sim \text{Mult}(1, \boldsymbol{\theta})$ [draw word]

Categorical Distribution

- Example corpus

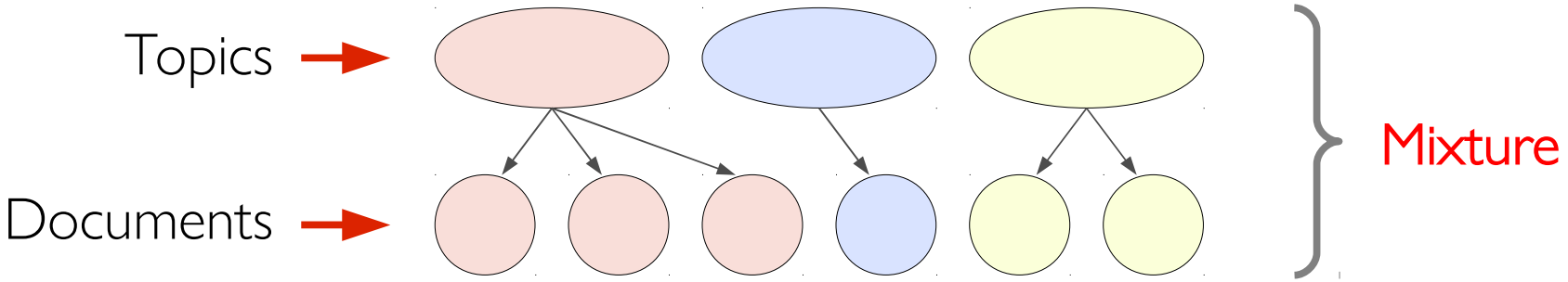
| the | he | is | the | and | the | she | she | is |
|---|---|---|---|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $w_8$ | $w_9$ |

- You will see the benefit of using conjugate prior for a distribution.

# Dirichlet-Multinomial Mixture Model

- Generative Process

We have $K$ topics.




- Example corpus

We have $V$ words.

| the | he | is |
|---|---|---|
| $w_{11}$ | $w_{12}$ | $w_{13}$ |

Document 1

| the | and | the |
|---|---|---|
| $w_{21}$ | $w_{22}$ | $w_{23}$ |

Document 2

| the | she | she | is |
|---|---|---|---|
| $w_{31}$ | $w_{32}$ | $w_{33}$ | $w_{34}$ |

Document 3

# Dirichlet-Multinomial Mixture Model

What is the dimension of $\boldsymbol{\theta}$?

What is the dimension of $\boldsymbol{\beta}_k$?

- Generative Process

For each topic $k \in \{1, \ldots, K\}$ :
  $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta})$ [draw distribution over words]
$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ [draw distribution over topics]
For each document $d \in \{1, \ldots, D\}$
  $z_d \sim \text{Mult}(1, \boldsymbol{\theta})$ [draw topic assignment]
For each word $n \in \{1, \ldots, N_d\}$
  $w_{dn} \sim \text{Mult}(1, \boldsymbol{\beta}_{z_d})$ [draw word]

- Example corpus

| the | he | is |
|---|---|---|
| $w_{11}$ | $w_{12}$ | $w_{13}$ |

Document 1

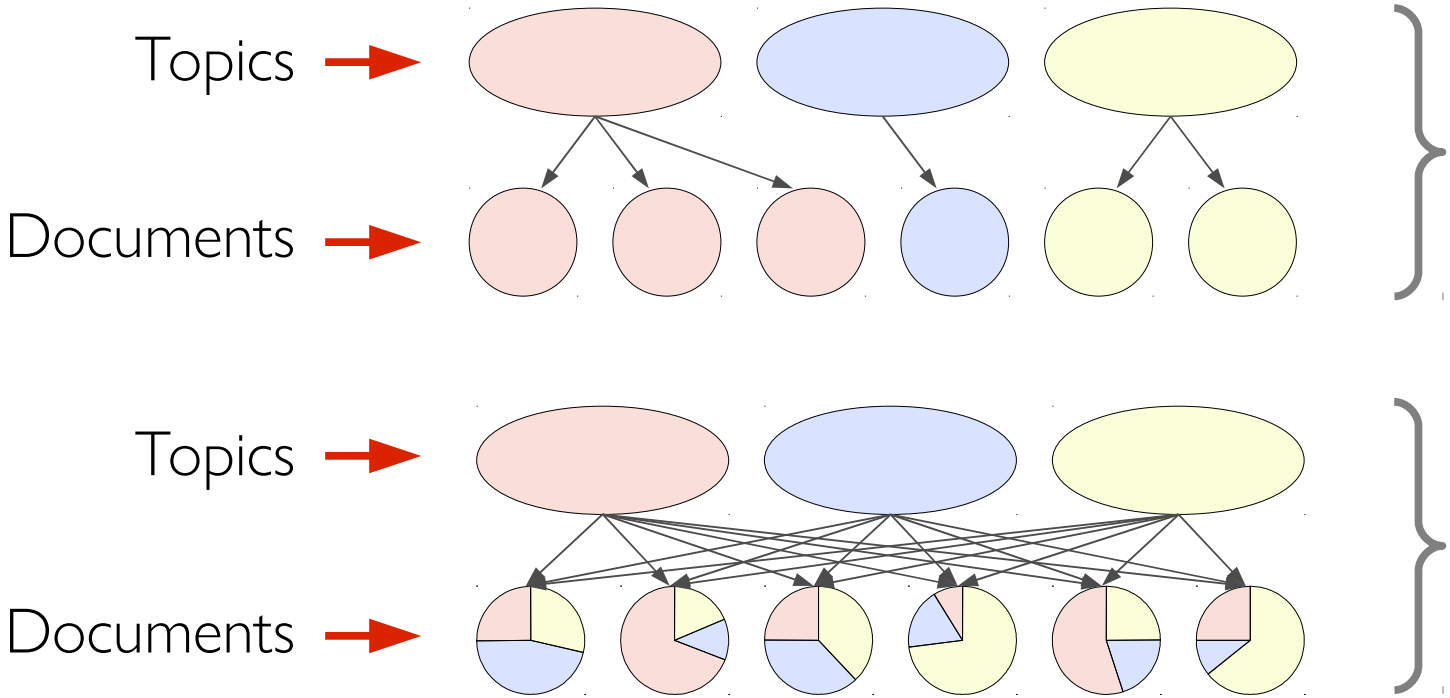
| the | and | the |
|---|---|---|
| $w_{21}$ | $w_{22}$ | $w_{23}$ |

Document 2

| the | she | she | is |
|---|---|---|---|
| $w_{31}$ | $w_{32}$ | $w_{33}$ | $w_{34}$ |

Document 3

# Mixture vs. Admixture




There should be multiple topics in a document:
Each document can be modeled by a new distribution of topics!

# Outline

- **Probabilistic Topic Model**
  - Dirichlet-Multinomial Model
  - **Latent Dirichlet Allocation**
  - Document Generation Process
- Variational Inference
  - Mean Field Variational Inference
  - Variational Inference for LDA
  - Parameter Estimation for LDA

# Latent Dirichlet Allocation (LDA)

Each document has a single topic $z$

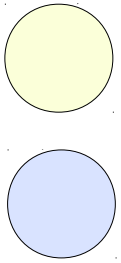
Dirichlet-Multinomial Mixture Model

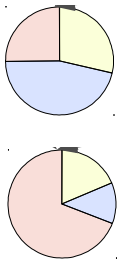
For each topic $k \in \{1, \ldots, K\}$ :
  $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta})$ [draw distribution over words]
$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ [draw distribution over topics]
For each document $d \in \{1, \ldots, D\}$
  $z_d \sim \text{Mult}(1, \boldsymbol{\theta})$ [draw topic assignment]
For each word $n \in \{1, \ldots, N_d\}$
  $w_{dn} \sim \text{Mult}(1, \boldsymbol{\beta}_{z_d})$ [draw word]

Each document has a topic distribution $\boldsymbol{\theta}$
Each word has a topic $z$

Latent Dirichlet Allocation

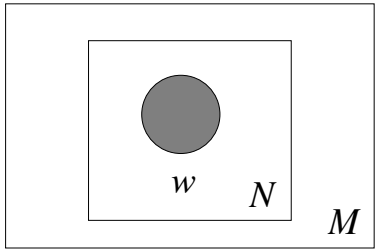
For each topic $k \in \{1, \ldots, K\}$ :
  $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta})$ [draw distribution over words]
For each document $d \in \{1, \ldots, D\}$
  $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$ [draw distribution over topics]
  For each word $n \in \{1, \ldots,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



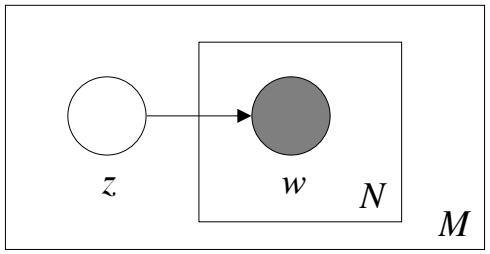


# Probabilistic Graphical Model

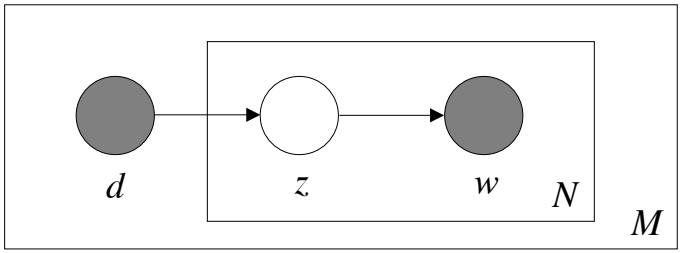
- **Box** with number on the bottom right:
  - Hierarchical samples and Number of samples.
- White/Gray **circle**: Hidden/Observed Variables.
- Arrow: Generating path (conditional prob).
- Read (b):
  - We generate M samples of $z$.
  - Each contains N kid samples $w$.
  - E.g. M clusters, each with N samples in it.
  - $w$ is observed, but $z$ is hidden.




(a) unigram



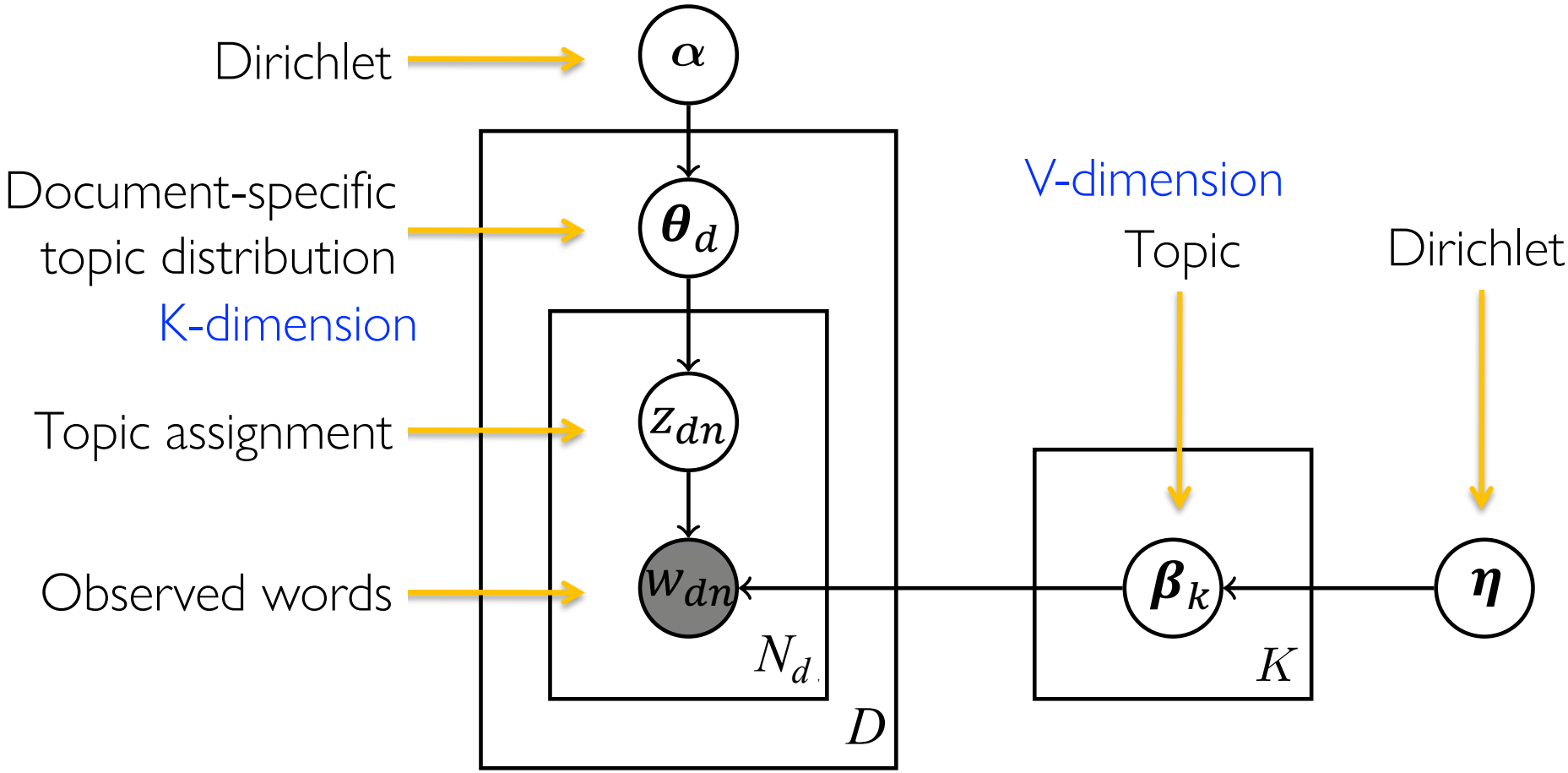

(b) mixture of unigrams




(c) pLSI/aspect model

- Can you read (c) Probabilistic Latent Semantic Indexing (PLSI) now?

# Latent Dirichlet Allocation (LDA)



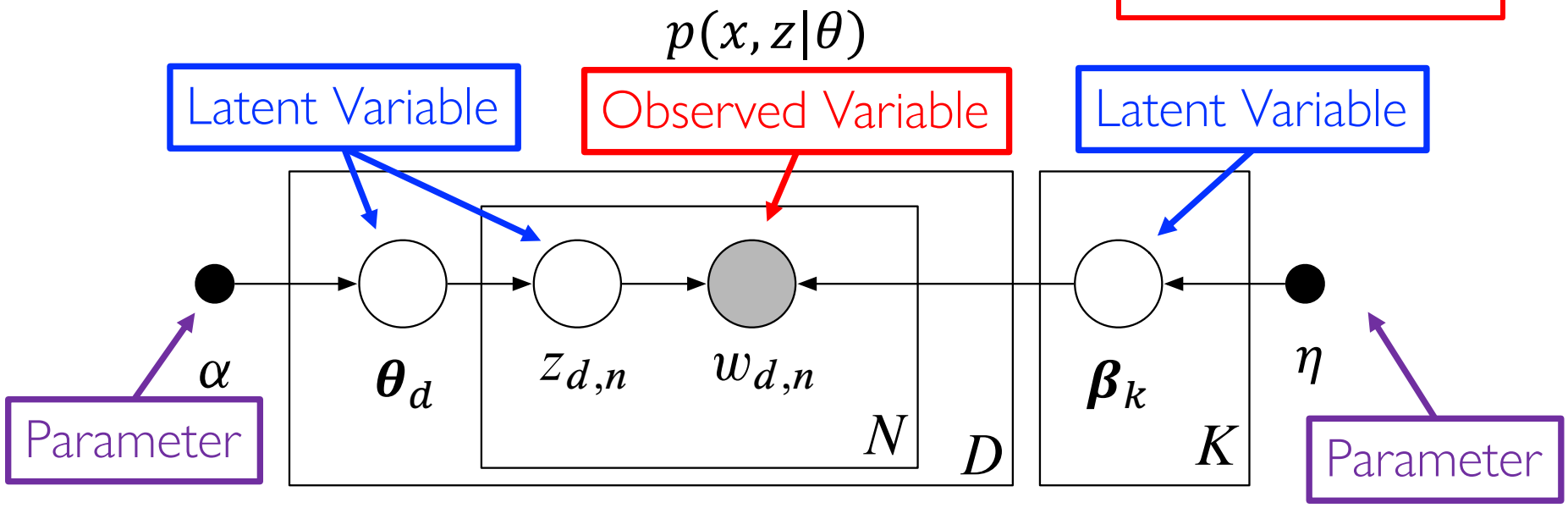

Can you explain graph of LDA now?

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = p(\boldsymbol{\beta}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})$$

# LDA as Latent Variable Model

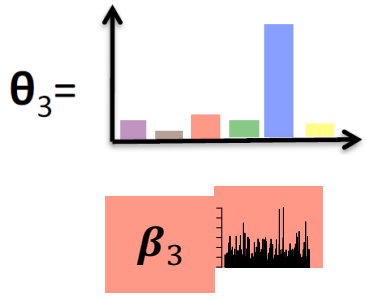
- Two sets of random variables: $z$ and $x$
  - $z$ consists of unobserved hidden variables.
  - $x$ consists of observed variables.
- Joint probability model parameterized by $\theta \in \Theta$:

Where is $x, z, \theta$ in LDA?



# How to Use LDA

- After we learn the model, we can do inference of all latent variable.
  - **Learning:** estimate parameters $\boldsymbol{\alpha}, \boldsymbol{\eta}$
  - **Inference:** for latent variable $\boldsymbol{\theta}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{1:D,1:N_d}$
- From latent variable $\boldsymbol{\theta}_d$:
  - We can know the main topic of this document.
- From latent variable $\boldsymbol{\beta}_k$:
  - We will know the hot words in this topic.
  - We can infer what this topic is about.
- Then we classify the documents, with no supervision.
- An important unsupervised learning approach to content analysis.




{hockey}

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

# Latent Dirichlet Allocation (LDA)

**[PDF] Latent dirichlet allocation**

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

We describe **latent Dirichlet allocation** (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in …

☆ Save  Cite  Cited by 46501  Related articles  All 90 versions  »

David Blei

Michael I. Jordan

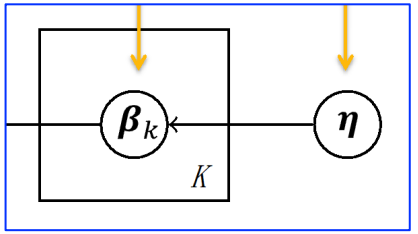
Andrew Ng

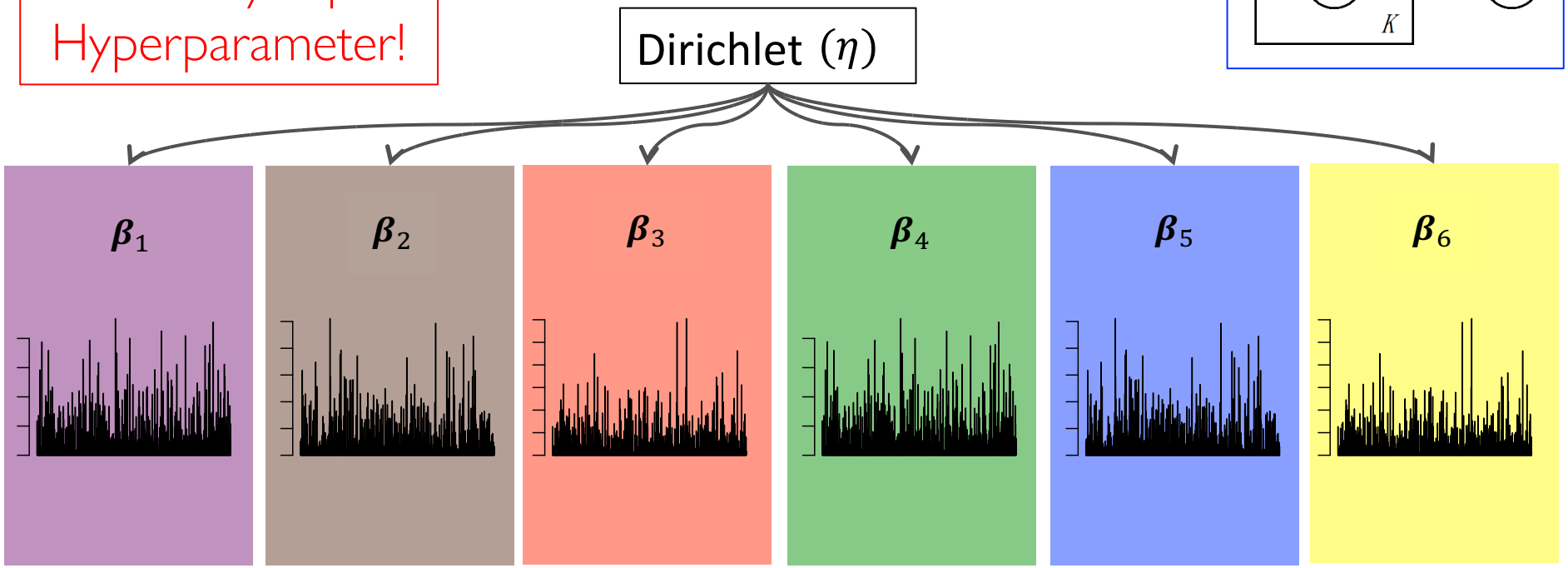
# Outline

- **Probabilistic Topic Model**
  - Dirichlet-Multinomial Model
  - Latent Dirichlet Allocation
  - **Document Generation Process**
- Variational Inference
  - Mean Field Variational Inference
  - Variational Inference for LDA
  - Parameter Estimation for LDA

# LDA for Topic Modeling

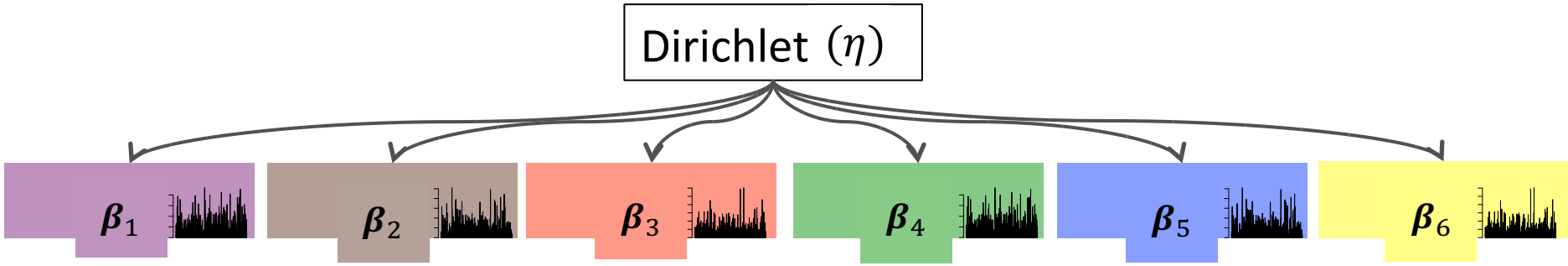
How many topics?
Hyperparameter!



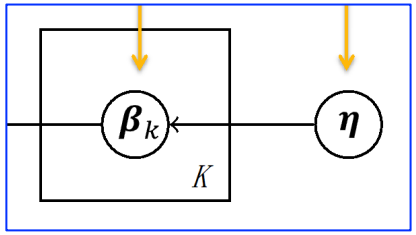

- The generative story begins with a **Dirichlet prior** over the topics.
- Each topic is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\beta}_k$.

# LDA for Topic Modeling



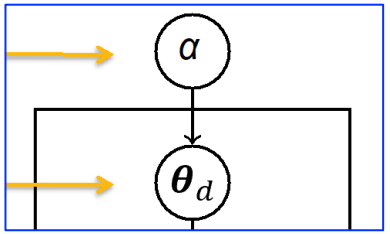

{hockey}

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

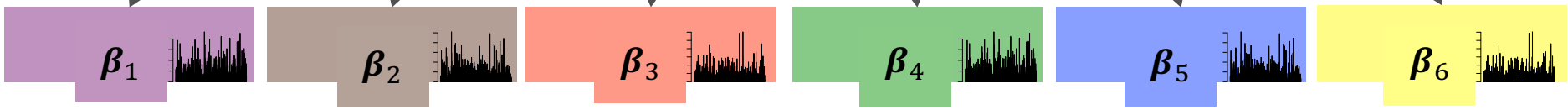



- A topic is visualized as its **high probability words**.
- A pedagogical **label** is used to identify the topic.

# LDA for Topic Modeling

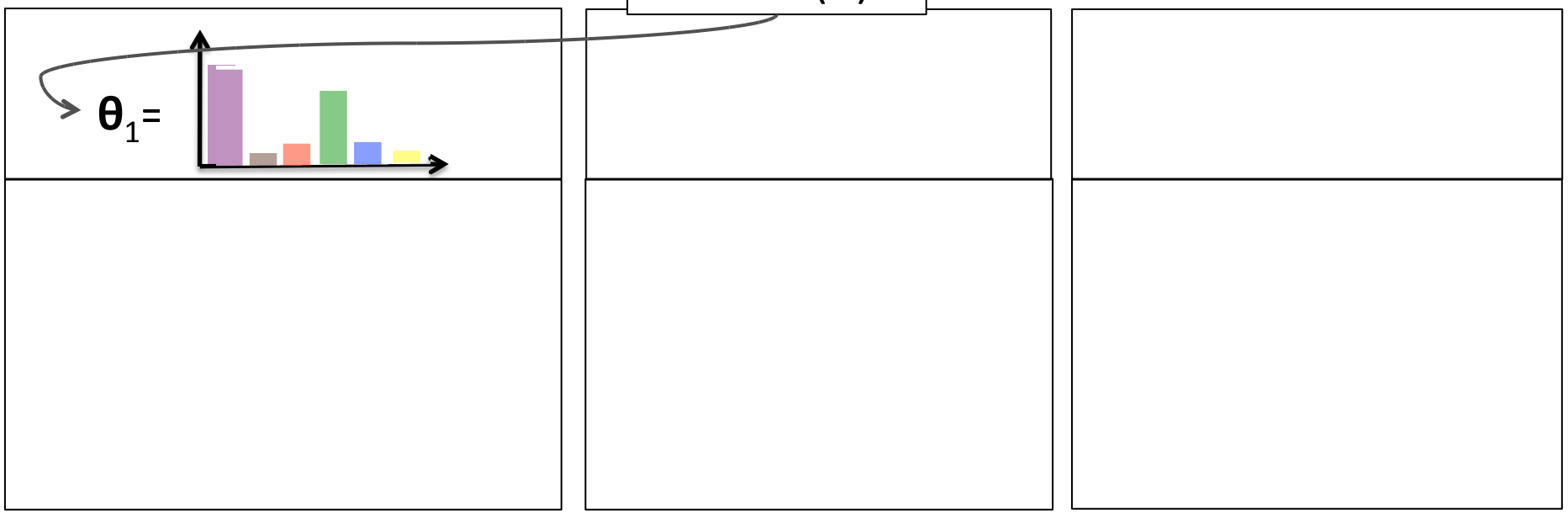



Dirichlet $(\eta)$

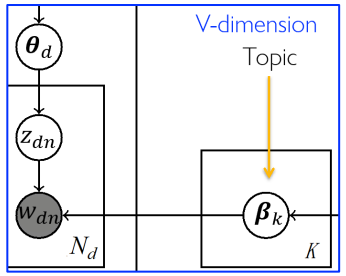



{Canadian gov.} {government} {hockey} {U.S. gov.} {baseball} {Japan}

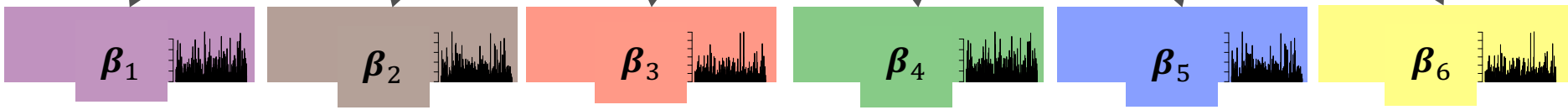
Dirichlet($\alpha$)



# LDA for Topic Modeling

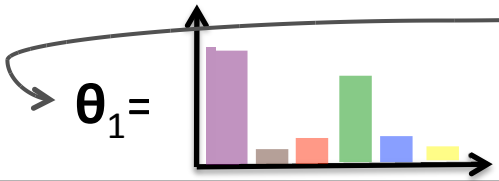



Dirichlet $(\eta)$



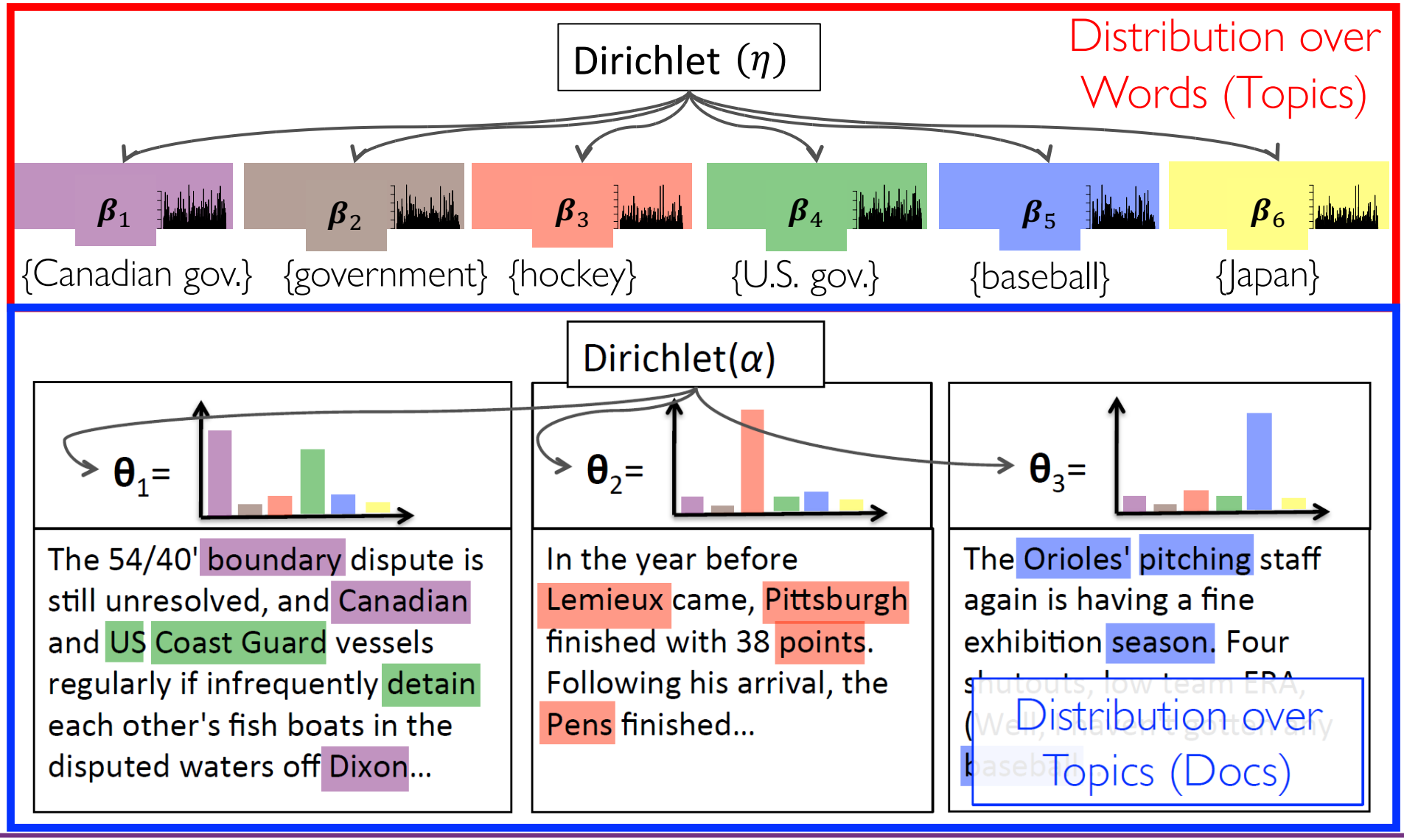

{Canadian gov.} {government} {hockey} {U.S. gov.} {baseball} {Japan}

Dirichlet($\alpha$)

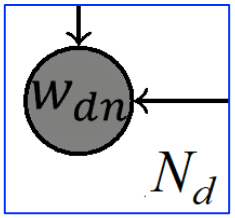



The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard

# LDA for Topic Modeling



# LDA for Topic Modeling




Inference and Learning start with only the data

Dirichlet ( )

$\beta_1 =$ $\beta_2 =$ $\beta_3 =$ $\beta_4 =$ $\beta_5 =$ $\beta_6 =$

Dirichlet ( )

| $\theta_1 =$ | $\theta_2 =$ | $\theta_3 =$ |
|---|---|---|
| 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' itching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball… |

# LDA for NIPS Papers

- https://proceedings.neurips.cc/paper/2015

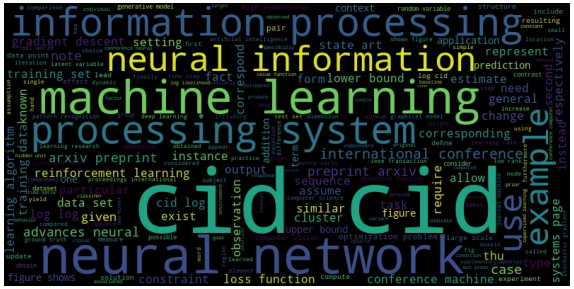
NeurIPS Proceedings

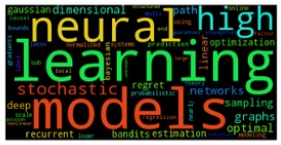
## Advances in Neural Information Processing Systems 28 (NIPS 2015)

Edited by: *C. Cortes and N. Lawrence and D. Lee and M. Sugiyama and R. Garnett*

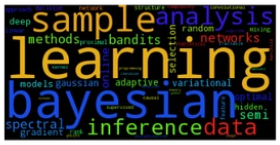
---

- Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling *Zheng Qu, Peter Richtarik, Tong Zhang*
- Associative Memory via a Sparse Recovery Model *Arya Mazumdar, Ankit Singh Rawat*
- Policy Gradient for Coherent Risk Measures *Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, Shie Mannor*
- A fast, universal algorithm to learn parametric nonlinear embeddings *Miguel A. Carreira-Perpinan, Max Vladymyrov*

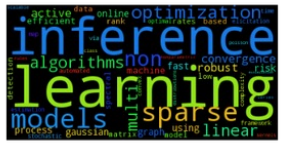


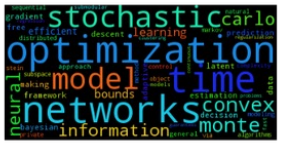

Topic 1: Neural




Topic 2: Bayesian




Topic 3: Inference




Topic 4: Optimization

# LDA for NIPS Papers

- https://cs.stanford.edu/people/karpathy/nips2015/

Below every paper are TOP 100 most-occuring words in that paper and their color is based on LDA topic model with k = 7.
(This is very hard but it looks like 0 = graphical models?, 1 = reinforcement learning?, 2 = deep learning, 3 = kernels?, 4 = theory?, 5 = optimization, 6 = matrix factorization?)

Toggle LDA topics to sort by: TOPIC0 TOPIC1 TOPIC2 TOPIC3 TOPIC4 TOPIC5 TOPIC6

Also note that you can filter by the day each poster appears: click on the day next to a paper to filter by that day.

**A Nonconvex Optimization Framework for Low Rank Matrix Estimation**
Tuo Zhao, Zhaoran Wang, Han Liu

[pdf] [rank by tf-idf similarity to this] [abstract]

Poster: Mon Dec 7th 07:00 - 11:59 PM @ 210 C #101

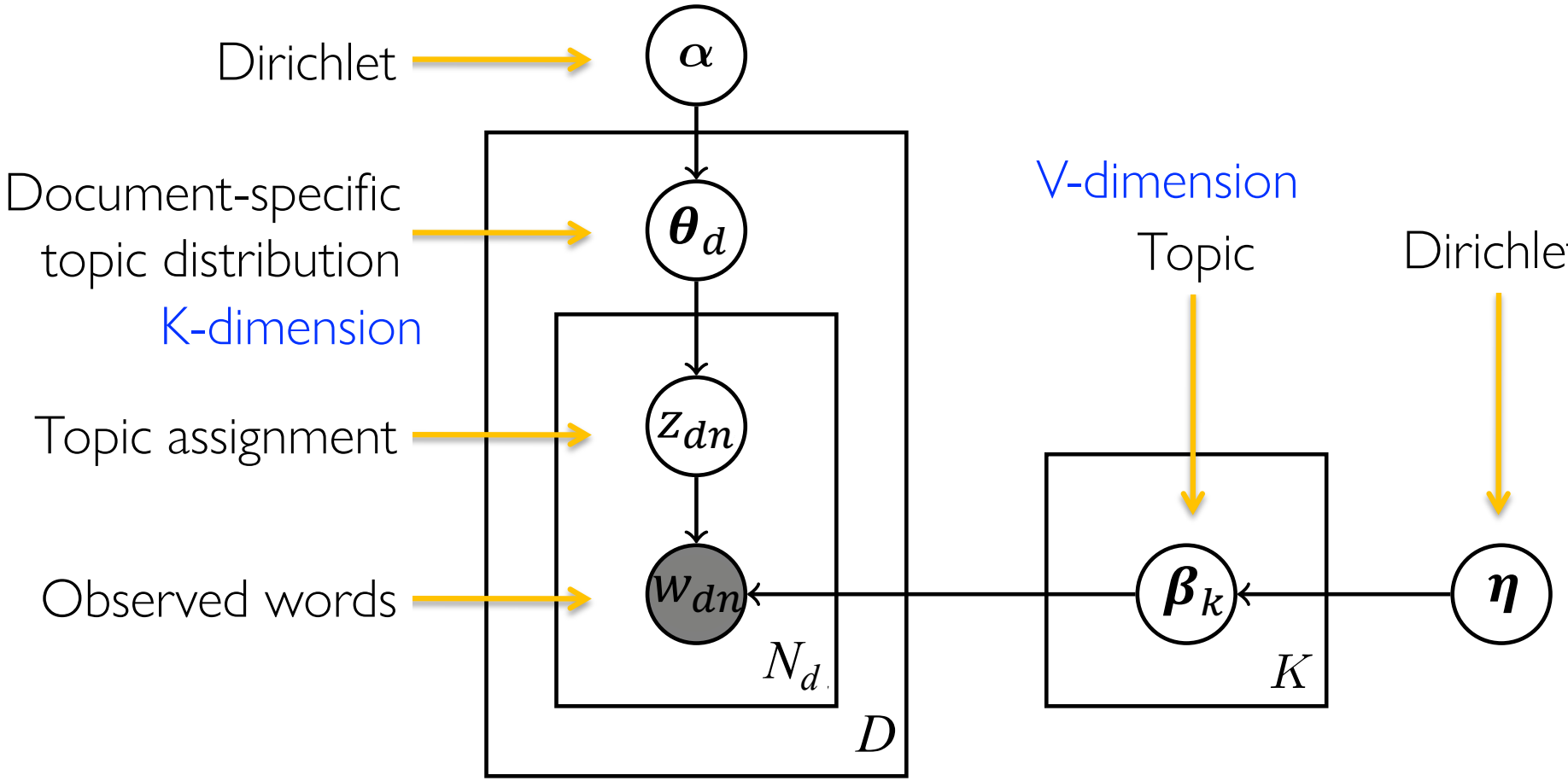
[exact, observation, key] [framework, based, existing, computationally, class, set] [arxiv, preprint] [log, computational, divergence, sample, family, parameter, introduce, analytic, point] [lemma, proof, algorithm, corollary, implies, prove, exists, complexity, number] [optimization, nonconvex, gradient, minimization, oracle, convex, projected, descent, convergence, rate, problem, global, max, satisfy, main, assumption, iteration, coordinate, assume, large, linear, updating, converge, general, step, established, optimum] [matrix, alternating, rank, establish, low, singular, theorem, completion, min, estimation, decomposition, appendix, geometric, provided, suppose, factorization, sensing, condition, error, high, orthonormal, result, broad, ieee, benjamin, entry, nsf, smaller, establishes, theoretical, including, solving]

0 =graphical models?, 1 = reinforcement learning?, 2 = deep learning, 3 = kernels?, 4 = theory?, 5 = optimization, 6 = matrix factorization?

# Outline

- Probabilistic Topic Model
  - Dirichlet-Multinomial Model
  - Latent Dirichlet Allocation
  - Document Generation Process
- **Variational Inference**
  - **Mean Field Variational Inference**
  - Variational Inference for LDA
  - Parameter Estimation for LDA

# Latent Dirichlet Allocation (LDA)

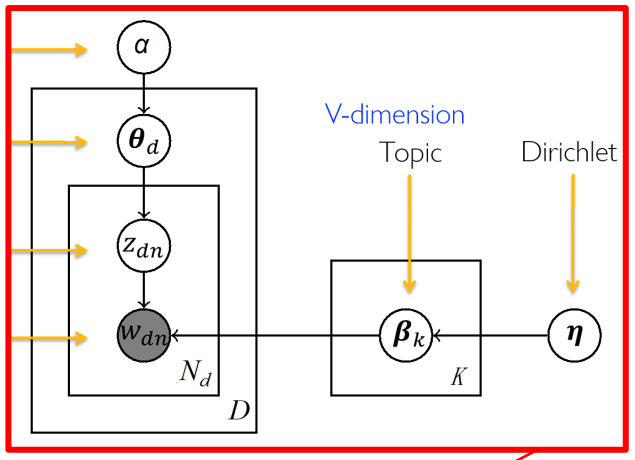



Can you explain graph of LDA now?

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = p(\boldsymbol{\beta}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})$$

# Maximum Likelihood Estimation

- If we know all the **latent variables:**
  - (We do not know them in fact.)
  - Let's try to compute MLE as an exercise!




$$\log p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} | \boldsymbol{\alpha}, \boldsymbol{\eta})$$

$$= \log p(\boldsymbol{\beta} | \boldsymbol{\eta}) + \log p(\boldsymbol{\theta} | \boldsymbol{\alpha}) + \log p(\boldsymbol{z} | \boldsymbol{\theta}) + \log p(\boldsymbol{w} | \boldsymbol{z}, \boldsymbol{\beta})$$

给定yita，K个主题独立

$$= \sum_{k=1}^{K} \log p(\vec{\boldsymbol{\beta}}_k | \boldsymbol{\eta}) + \sum_{d=1}^{D} \log p(\vec{\boldsymbol{\theta}}_d | \boldsymbol{\alpha}) + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log p(z_{d,n} | \vec{\boldsymbol{\theta}}_d)$$

$$+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log p(w_{d,n} | z_{d,n}, \vec{\boldsymbol{\beta}}_{1:K})$$

PGM to simplify learning and inference!

# Complex MLE Objective

- Recall the form of Dirichlet Distribution: $p(\vec{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$

词表上有V个词

$$= \sum_{k=1}^{K}\left(\sum_{v=1}^{V}(\eta_v - 1)\log\beta_{kv} - \log B(\boldsymbol{\eta})\right) + \sum_{d=1}^{D}\sum_{k=1}^{K}(\alpha_k - 1)\log\theta_{dk} - \log B(\boldsymbol{\alpha})$$

beta_z:V维

$$+ \sum_{d=1}^{D}\sum_{n=1}^{N_d}\log\theta_{d,z_{d,n}} + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\log\beta_{z_{d,n}w_{d,n}}$$

- We will meet some more complex terms later 😐
- We need to solve for parameters $\boldsymbol{\alpha}, \boldsymbol{\eta}$.

Newton Method

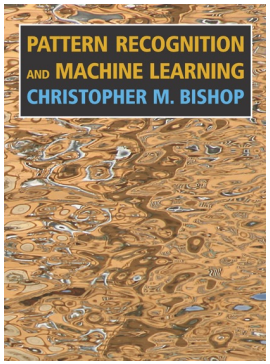
$$\underset{\boldsymbol{\eta}}{\text{argmax}}\sum_{k=1}^{K}\left(\sum_{v=1}^{V}(\eta_v - 1)\log\beta_{kv} - \log B(\boldsymbol{\eta})\right)$$

# ELBO and EM Algorithm

- Evidence Lower Bound (ELBO):

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

$$= -\mathrm{KL}[q(z)||p(z|x,\theta)] + \log p(x|\theta)$$




PRML
Chapter 10

- EM Algorithm: 必考

Minimize

1. Choose initial $\theta^{\mathbf{old}}$
2. E-step: Let $q^* = \underset{q}{\mathrm{argmax}}\,\mathcal{L}(q,\theta^{\mathbf{old}})$
3. M-step: Let $\theta^{\mathbf{new}} = \underset{\theta}{\mathrm{argmax}}\,\mathcal{L}(q^*,\theta)$
4. Go to step 2, until converged.

Optimal choice for E-step
$q^*(z) = p(z|x,\theta^{\mathbf{old}})$

# Intractable E-Step

- We need to solve: Optimal choice for E-step $q^*(z) = p(z|x, \theta^{\text{old}})$

In LDA: $p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \dfrac{p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}$

- But the denominator is intractable!

$$p(\mathbf{w}|\alpha, \eta) = \int \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})\, d\boldsymbol{\theta}\, d\boldsymbol{\beta}$$

Inference is hard for general Bayesian Models!

- We need approximation here!
  - There are two ways to solve E-step approximately:
    - Variational Inference (this lecture)
    - Sampling – Markov Chain Monte Carlo (next lecture) 会考

# Variational Methods

- Suppose E-step is difficult:
  - Hard to take expectation w.r.t. $q^*(z) = p(z|x, \theta^{\text{old}})$
  - For example, topic models.
- **Solution:** Restrict to distributions $\mathcal{Q}$ that are easy to work with.
- The ELBO now looser:

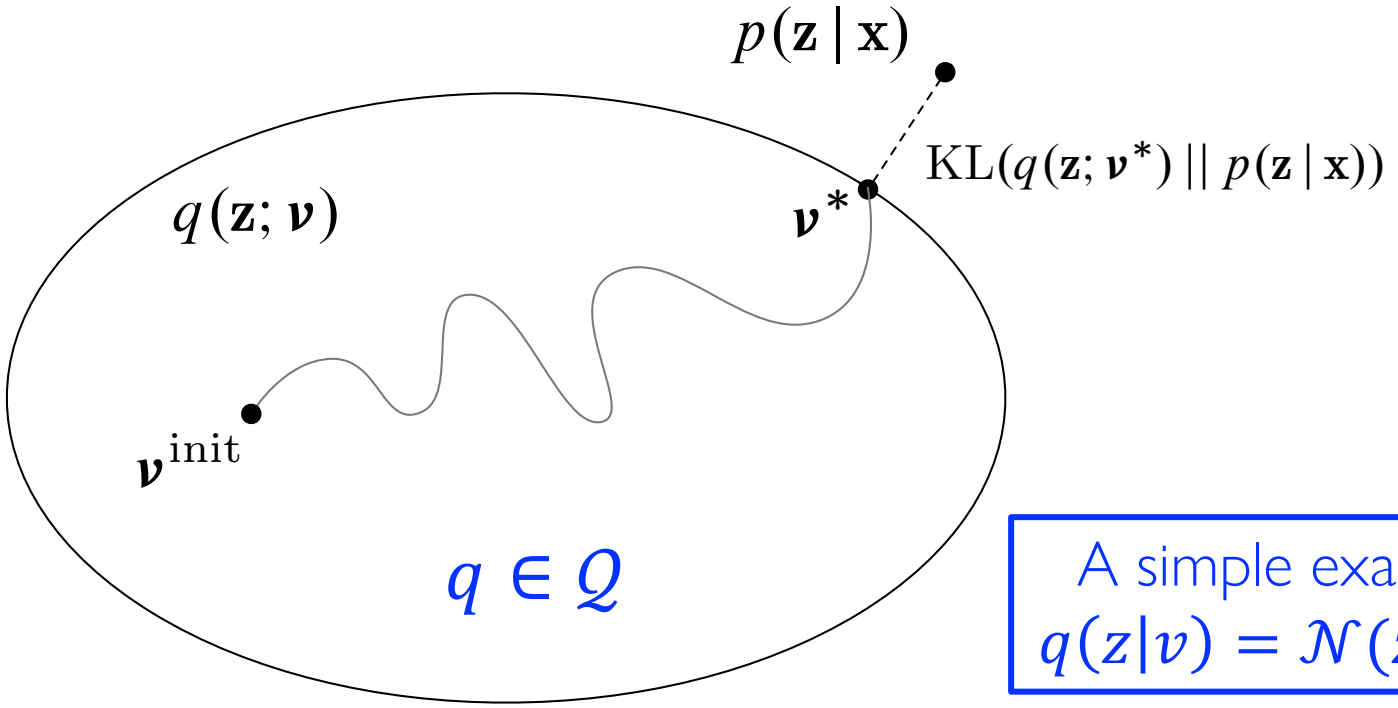
Intractable!

We compute this

$$q^* = \underset{q \in \mathcal{Q}}{\text{argmin}}\, \text{KL}\left[q(z), p\left(z|x, \theta^{\text{old}}\right)\right] = \underset{q \in \mathcal{Q}}{\text{argmax}}\, \mathcal{L}\left(q, \theta^{\text{old}}\right)$$

- Find an easy-to-work variational distribution $q^*$ to approximate the inference distribution $p(z|x, \theta^{\text{old}})$.
  - This group of methods are called Variational Methods.

# Variational Inference (VI)




A simple example:
$q(z|v) = \mathcal{N}(z|\mu, \sigma)$

- A variational distribution $q(z; v)$ along with its parameter $v$ is used.
  - Objective in VI: Find $v$ to maximize the ELBO.
- $v$ may have no relation with $\theta^{\text{old}}$, which makes VI easier to solve.

# Mean Field Theory

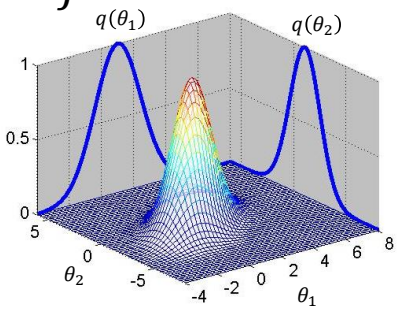
不要求

- Assume $\mathcal{Q}$ a distribution $q(z)$ that factorizes $q(z) = \prod_{i=1}^{m} q_i(z_i)$.
- Variational inference of $q(z)$ corresponds to the mean field theory.
- Seek all factors $\{q_i = q_i(z_i)\}$ to maximize ELBO: $q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmax}}\, \mathcal{L}(q)$

$$\mathcal{L}(q) = \int \prod_i q_i \log \frac{p(x,z)}{\prod_i q_i} \mathrm{d}z = \int \prod_i q_i \left\{ \log p(x,z) - \sum_i \log q_i \right\} \mathrm{d}z$$

$$= \int q_j \left\{ \int \log p(x,z) \prod_{i \neq j} q_i \mathrm{d}z_i \right\} \mathrm{d}z_j - \int q_j \log q_j \, \mathrm{d}z_j + C$$

$$= \int q_j \{ \log \tilde{p}(x, z_j) \} \mathrm{d}z_j - \int q_j \log q_j \, \mathrm{d}z_j + C$$



不要求

# Mean Field Theory

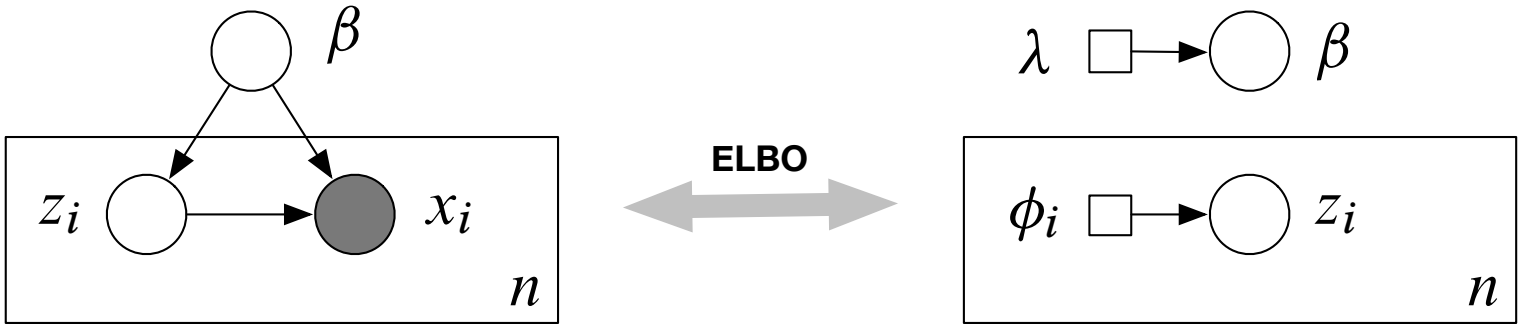
- ELBO: $\mathcal{L}(q) = \int q_j \{\log \tilde{p}(x, z_j)\} \mathrm{d}z_j - \int q_j \log q_j \, \mathrm{d}z_j + C$
  - Where $\log \tilde{p}(x, z_j) = \int \log p(x, z) \prod_{i \neq j} q_i \mathrm{d}z_i \triangleq \mathbb{E}_{i \neq j}[\log p(x, z)]$
- We keep $\{q_{j \neq i}\}$ fixed and maximize $\mathcal{L}(q)$ for all $q_j(z_j)$:
  - Note that $\mathcal{L}(q) = -\mathrm{KL}(q_j(z_j) \| \tilde{p}(x, z_j)) + C$
    - The maximizer of $\mathcal{Q}$ to ELBO is $q_j(z_j) = \tilde{p}(x, z_j)$
  - A general expression for the optimal solution $q_j^*(z_j)$:

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j}[\log p(x, z)] + C$$

- In physics, the mean field theory studies the behavior of high-dimensional stochastic models by studying a simpler model that approximates the original one by averaging over degrees of freedom.

隐变量独立

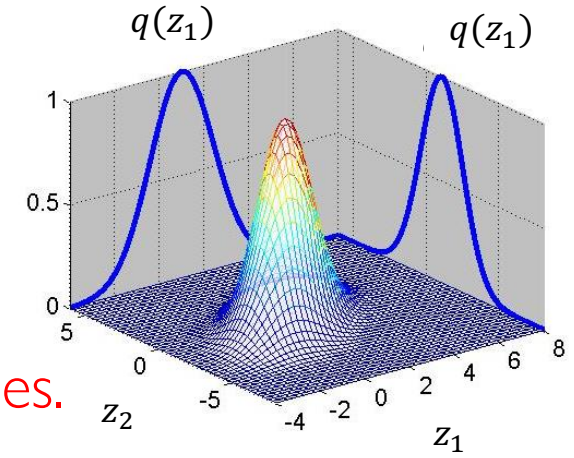
# Mean Field Variational Inference




- Consider a simpler graphical model.

$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- The difficulty to tackle this distribution:
  - Latent variable depends on other latent variables.




- **Mean Field Assumption** tries to disentangle latent variables to simplify:

$$q(\beta, z; \lambda, \phi) = q(\beta; \lambda) \prod_{i=1}^{n} q(z_i; \phi_i)$$

No dependency!

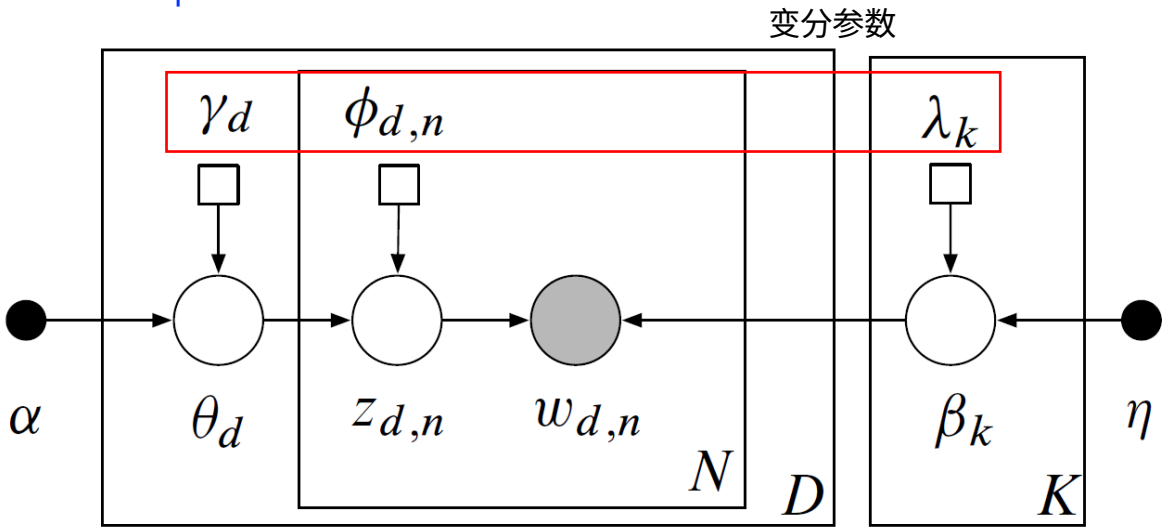
可以假设beta是高斯分布

# Outline

- Probabilistic Topic Model
  - Dirichlet-Multinomial Model
  - Latent Dirichlet Allocation
  - Document Generation Process
- **Variational Inference**
  - Mean Field Variational Inference
  - **Variational Inference for LDA**
  - Parameter Estimation for LDA

# Variational Inference for LDA

- Mean field assumption in LDA:




$$q(\vec{\boldsymbol{\theta}}_{1:D}, \boldsymbol{z}_{1:D,1:N_d}, \vec{\boldsymbol{\beta}}_{1:K}) = \prod_{k=1}^{K} q(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{\lambda}}_k) \prod_{d=1}^{D} q(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{\gamma}}_d) \prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\vec{\boldsymbol{\phi}}_{d,n})$$

Dirichlet Distribution

Dirichlet Distribution

Multinomial Distribution

# Variational Inference for LDA

- Recall the form of ELBO:

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right) = \mathbb{E}_{z\sim q}\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

- In LDA, we have known the form of $\boldsymbol{q}$, and what left is:

$$p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta}) = p(\boldsymbol{\beta}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})$$

- The objective function $\mathcal{L}$ in VI of LDA (Solvable!) turns out to be:

$$\mathcal{L} = \sum_{k=1}^{K} E_q\left[\log p(\vec{\boldsymbol{\beta}}_k|\boldsymbol{\eta})\right] + \sum_{d=1}^{D} E_q\left[\log p(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{\alpha}})\right] +$$
$$\sum_{d=1}^{D}\sum_{n=1}^{N} E_q\left[\log p(Z_{d,n}|\vec{\boldsymbol{\theta}}_d)\right] +$$
$$\sum_{d=1}^{D}\sum_{n=1}^{N} E_q\left[\log p(w_{d,n}|Z_{d,n},\vec{\boldsymbol{\beta}}_{1:K})\right] - E_q \log q$$

# Variational Inference for LDA

- We solve one of the terms as an example:

All latent variables are gone!

$$
\begin{aligned}
& E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] \\
= \; & E_q[\log\left(\exp\left\{\left(\sum_{i=1}^{K}(\alpha_i-1)\log\theta_i\right)+\log\Gamma(\sum_{i=1}^{K}\alpha_i)-\sum_{i=1}^{K}\log\Gamma(\alpha_i)\right\}\right)], \\
= \; & \left(\sum_{i=1}^{K}(\alpha_i-1)E_q[\log\theta_i]\right)+\log\Gamma(\sum_{i=1}^{K}\alpha_i)-\sum_{i=1}^{K}\log\Gamma(\alpha_i), \\
= \; & \left(\sum_{i=1}^{K}(\alpha_i-1)(\Psi(\gamma_i)-\Psi(\sum_{j}\gamma_j))\right)+\log\Gamma(\sum_{i=1}^{K}\alpha_i)-\sum_{i=1}^{K}\log\Gamma(\alpha_i).
\end{aligned}
$$

- $\Psi$ is the first derivative of the log Gamma function $\Gamma$.
- About the last step please refer to Section A.1 of the original paper 😐

# Variational Inference for LDA

- The objective function $\mathcal{L}$ in VI of LDA turns out to be something like:

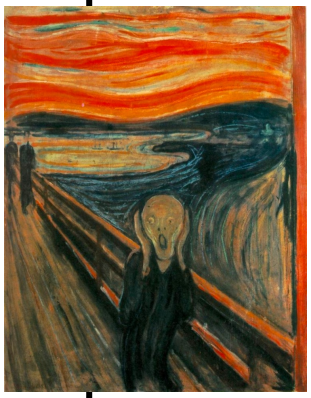
$$L(\gamma,\phi;\alpha,\beta) = \log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

$$+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

$$+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij}$$

$$- \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

$$- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni},$$



You may read the original paper to find the details

- Bayesians face something like this everyday 😐
- Although this is scaring, Just use Calculus or SGD! $\nabla L = 0$

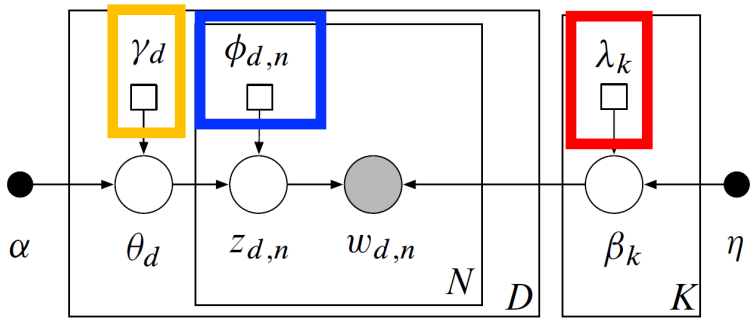
# Variational Inference for LDA

- Iterate the following steps (Descend each variable iteratively):

(1) For each topic $k$ and term $v$:

$$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

隐变量对应的变分分布




(2) For each document $d$:

(a) Update $\gamma_d$:

$$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word $n$, update $\vec{\phi}_{d,n}$:

$$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{\Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\sum_{v=1}^{V}\lambda_{k,v}^{(t+1)})\right\},$$

where $\Psi$ is the digamma function, the first derivative of the log $\Gamma$ function.

$$\sum_{k=1}^{V} \phi_{dnk} = 1$$

Digamma function $\Psi$ can only be computed by approximation, e.g. Taylor

# Outline

- Probabilistic Topic Model
  - Dirichlet-Multinomial Model
  - Latent Dirichlet Allocation
  - Document Generation Process
- **Variational Inference**
  - Mean Field Variational Inference
  - Variational Inference for LDA
  - **Parameter Estimation for LDA**

# EM Algorithm

- ELBO:

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

$$= -\mathrm{KL}[q(z)||p(z|x,\theta)] + \log p(x|\theta)$$

- EM Algorithm with Variational Inference:
  - Choose initial $\theta^{\mathbf{old}}$
  - Let $q^* = \underset{q}{\mathbf{argmax}}\,\mathcal{L}(q,\theta^{\mathbf{old}})$ We are here
  - Let $\theta^{\mathbf{new}} = \underset{\theta}{\mathbf{argmax}}\,\mathcal{L}(q^*,\theta)$ What we left
  - Go to step 2, until converged.

In LDA:
Solve $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$

# Parameter Estimation for LDA

- **Minimize** the part of ELBO that relates to $\boldsymbol{\alpha}$:

$$L_{[\alpha]} = \sum_{d=1}^{D}\left(\log\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \sum_{i=1}^{K}\log\Gamma\left(\alpha_i\right) + \sum_{i=1}^{K}\left(\left(\alpha_i - 1\right)\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right)\right)\right)\right)$$

- We can't find an analytical solution.
  - Use Newton method.
  - Compute Derivatives and Hessian:

Newton Iteration Function

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

$$\frac{\partial L}{\partial \alpha_i} = D\left(\Psi\left(\sum_{j=1}^{K}\alpha_j\right) - \Psi\left(\alpha_i\right)\right) + \sum_{d=1}^{D}\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right)\right)$$

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i,j)D\Psi'\left(\alpha_i\right) - \Psi'\left(\sum_{j=1}^{K}\alpha_j\right)$$

- The problem of $\boldsymbol{\eta}$ is as the same form of $\boldsymbol{\alpha}$'s problem.

# LDA: Learning

- Iterate between:
  - E-step: Iterative optimization

(1) For each topic $k$ and term $v$:

$$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

(2) For each document $d$:

(a) Update $\gamma_d$:

$$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word $n$, update $\vec{\phi}_{d,n}$:

$$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{\Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)})\right\},$$

where $\Psi$ is the digamma function, the first derivative of the log $\Gamma$ function.

  - M-step:
    - Solve $\boldsymbol{\alpha}^{\text{new}}, \boldsymbol{\eta}^{\text{new}} := \underset{\boldsymbol{\alpha},\boldsymbol{\eta}}{\text{argmax}}\, \mathcal{L}(q^*, \boldsymbol{\alpha}, \boldsymbol{\eta})$ (Newton or SGD)

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i,j) D \Psi'(\alpha_i) - \Psi'\left(\sum_{j=1}^{K} \alpha_j\right)$$

# LDA: Inference

- After computing all parameters in the **mean field approximation**:

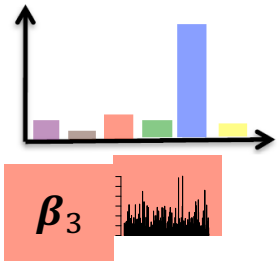
$$q(\vec{\boldsymbol{\theta}}_{1:D}, \boldsymbol{z}_{1:D,1:N_d}, \vec{\boldsymbol{\beta}}_{1:K}) = \prod_{k=1}^{K} q(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{\lambda}}_k) \prod_{d=1}^{D} q(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{\gamma}}_d) \prod_{d=1}^{D}\prod_{n=1}^{N} q(z_{d,n}|\vec{\boldsymbol{\phi}}_{d,n})$$

- How do we compute $\vec{\boldsymbol{\theta}}_{1:D}, \boldsymbol{z}_{1:D,1:N_d}, \vec{\boldsymbol{\beta}}_{1:K}$? Argmax!

  - $\vec{\boldsymbol{\theta}}_d = \underset{\vec{\boldsymbol{\theta}}}{\operatorname{argmax}}\, q(\vec{\boldsymbol{\theta}}|\vec{\boldsymbol{\gamma}}_d), \vec{\boldsymbol{\beta}}_k = \underset{\vec{\boldsymbol{\beta}}}{\operatorname{argmax}}\, q(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{\lambda}}_k)$

  - $z_{d,n} = \underset{z}{\operatorname{argmax}}\, q(z|\vec{\boldsymbol{\phi}}_{d,n})$

- Variational Inference does not only enable EM algorithm.
  - The essential goal of VI is to do inference.

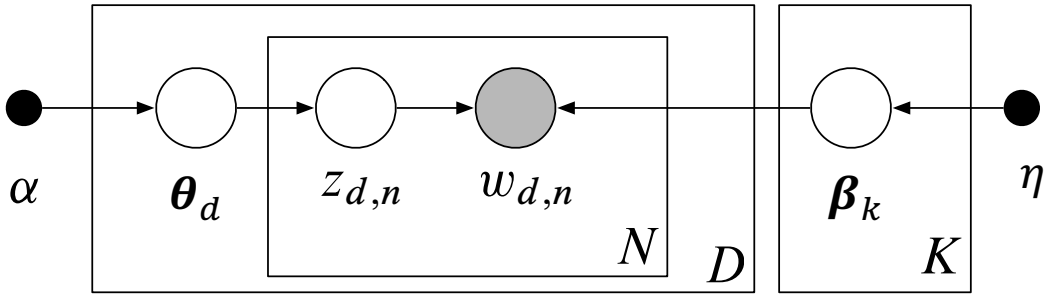



{hockey}

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

# LDA: Evaluation 不要求

- How to evaluation LDA?
  - First, we need a held out corpora for testing.
  - Which component of the model can generalize to unseen data?



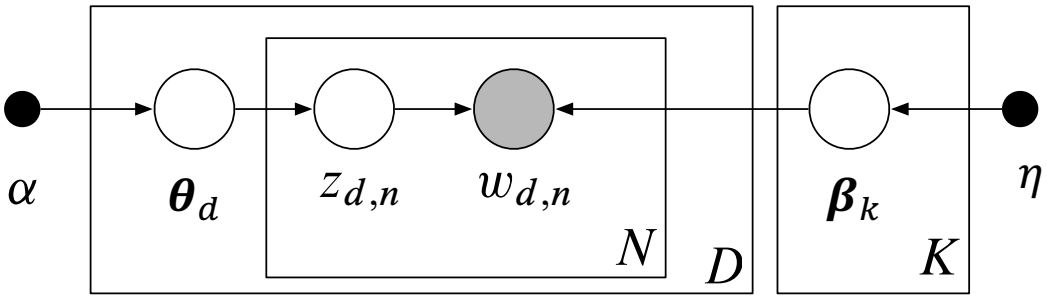

  - Of course, parameter $\boldsymbol{\alpha}, \boldsymbol{\eta}$ can generalize.
    - $\boldsymbol{\beta}_{1:K}$ can also generalize to related corpus!
- Try to compute the marginal likelihood of words on a new dataset:

$$\text{perplexity } (D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{D_{te}} \log p\left(\mathbf{w}_d|\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta}\right)}{\sum_{d=1}^{D_{te}} N_d}\right\}$$

Lower is better

# Perplexity Metric

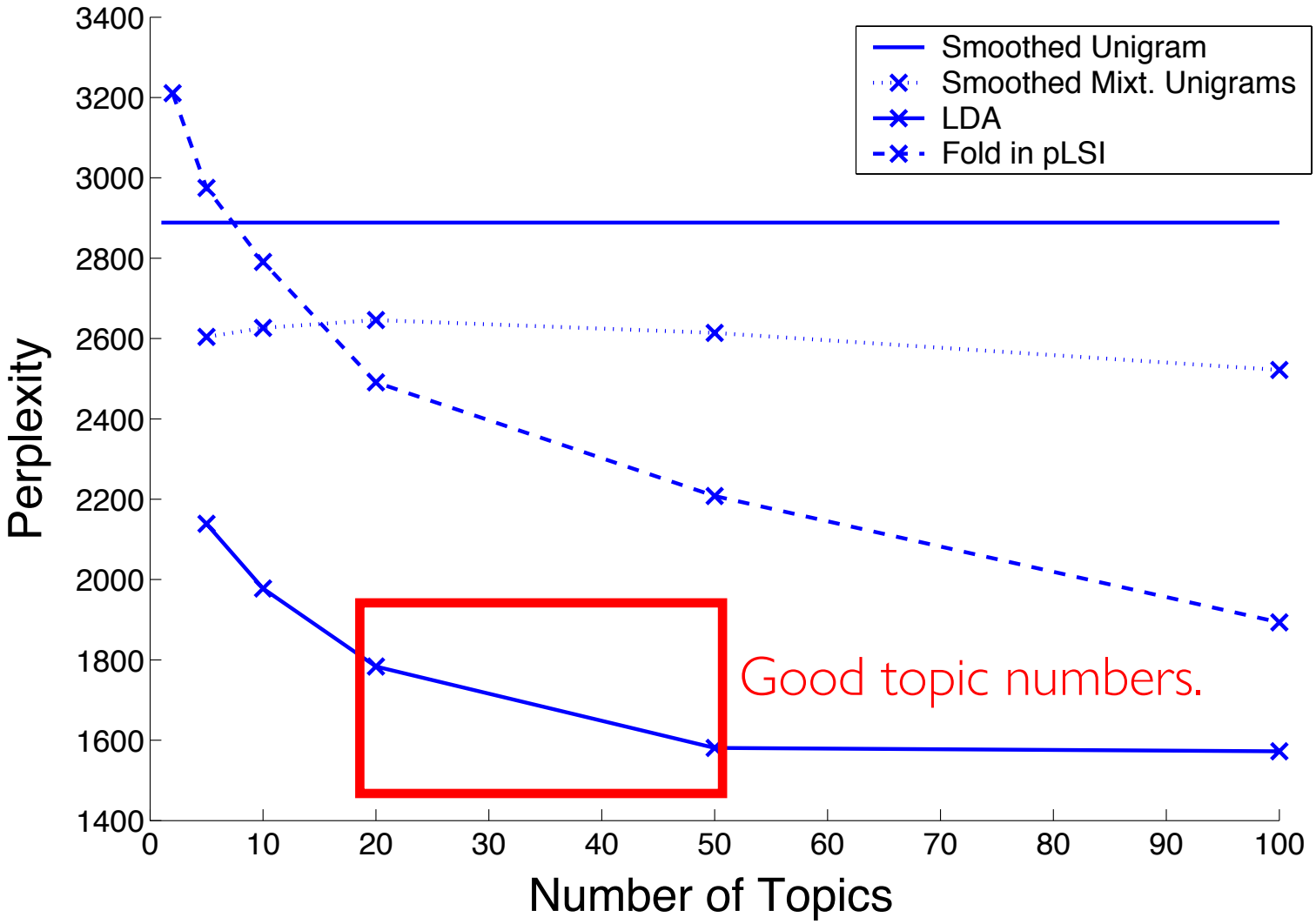
- With $\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta}_{1:K}$, we can use similar method to infer $\boldsymbol{\theta}, \boldsymbol{z}$ on test data.




- After knowing all the latent variables, the marginal likelihood of words can be computed as:

$$\text{perplexity } (D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{D_{te}} \log p\left(\mathbf{w}_d|\boldsymbol{z}_d, \boldsymbol{\beta}\right) p\left(\mathbf{z}_d|\boldsymbol{\theta}_d\right) p\left(\boldsymbol{\theta}_d|\boldsymbol{\alpha}\right)}{\sum_{d=1}^{D_{te}} N_d}\right\}$$

# Performance Evaluation



Thank You

# Questions?

Mingsheng Long
mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong
答疑：东主楼11区413室