EECS 510 Social Media Mining
Lecture 9: Topic Modeling and Generative Models

Spring 2016

Rosanne Liu

rosanne.liu@northwestern.edu

Electrical Engineering and Computer Science
Northwestern University

April 28, 2016

Administrative

Project proposals looking good! (Prelim report due in 2 weeks)

- It better has a linkage to social media mining.
- Make sure you can showcase your contribution on at least one of:
  - Data acquisition
  - Data preprocessing and analysis
  - Coding up an algorithm
  - Result and performance analysis

Paper presentation starting May 17th (Lecture 14)!

- 2 to 3 presentations each time of class
- $\sim$25 minutes each group
- Expect to see quasi-equal contributions

Lecture Outline

1. The Big Picture
2. Latent Dirichlet Allocation
3. Generative Models

Outline

So far in this class we have covered:

- Graph theory basics
- Data mining essentials, tools
- NLP, Text mining

And we are going to learn:

- Topic modeling
- Deep learning
- Web scraping
- Data science from a business angle

Suppose you are a data scientist...

Problem:

I got this new numeric/categorical data set. Now what?

Suppose you are a data scientist...

Problem:

I got this new numeric/categorical data set. Now what?

Idea:

Exploratory data analysis (or, look at the data)
Statistics
Visualize it

Suppose you are a data scientist...

Problem:

I got this new **text corpus**. Now what?

Suppose you are a data scientist...

Problem:

I got this new **text corpus**. Now what?

Idea #1:

Compute gross statistics (#tokens, #word types, etc.)

Suppose you are a data scientist...

Problem:

I got this new **text corpus**. Now what?

Idea #1:

Compute gross statistics (#tokens, #word types, etc.)

Idea #2:

Label words. Visualize distribution of labels.

Suppose you are a data scientist...

Problem:

I got this new **text corpus**. Now what?

Idea #1:

Compute gross statistics (#tokens, #word types, etc.)

Idea #2:

Label words. Visualize distribution of labels.

Problem: How to set labels? POS tagger? NER? Coref?

What's in my text corpus?

Assume that text corpus is generated from some set of topics in the world (sports, business, politics, etc.)

"Kennedy" $\in$ {politics, movies}
"petroleum" $\approx$ gasoline
"browser" $\notin$ woodworking

What's in my text corpus?

Assume that text corpus is generated from some set of topics in the world (sports, business, politics, etc.)

"Kennedy" $\in$ {politics, movies}
"petroleum" $\approx$ gasoline
"browser" $\notin$ woodworking

Infer the topics/labels from data: topic modeling!

Key assumptions behind topic modeling

- A *document* is a collection of words
- A *topic* is a distribution over a fixed vocabulary
- Documents exhibit multiple topics
- A topic model is a model that *generates* documents
- Only the number of topics is specified in advance

A pioneer paper

DOI:10.1145/2133806.2133826

Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

[Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.]

The problem with information

As more information becomes available, it becomes more difficult to access what we are looking for.

We need new tools to help us organize, search, and understand these vast amounts of information.

Topic Modeling

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Uncover the hidden topical patterns that pervade the collection.
2. Annotate the documents according to those topics.
3. Use the annotations to organize, summarize, and search the texts.

Discover topics from a corpus (*Science*)

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Further modeling of topics

With topics discovered from documents, one can ...

- Model the evolution of topics over time.
- Model connections between topics.
- Annotate images.

Further topics of topic modeling

From a machine learning perspective, topic modeling is a case study in applying hierarchical Bayesian models to grouped data, like documents or images.

Topic modeling research touches on:

- Directed graphical models
- Conjugate priors and nonconjugate priors
- Mixed membership models
- Hierarchical Bayesian methods
- Fast approximate posterier inference (MCMC, variational methods)
- Exploratory data analysis
- Model selection and nonparametric Bayesian methods

Outline

Probabilistic modeling

1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using posterior inference
   - What are the topics that describe this collection?
3. Situate new data into the estimated model.
   - How does this query or new document fit into the estimated topic structure?

Intuition behind LDA

Simple intuition: Documents exhibit multiple topics.

Optimal Brain Damage

Yann Le Cun, John S. Denker and Sara A. Solla
AT&T Bell Laboratories, Holmdel, N. J. 07733

ABSTRACT

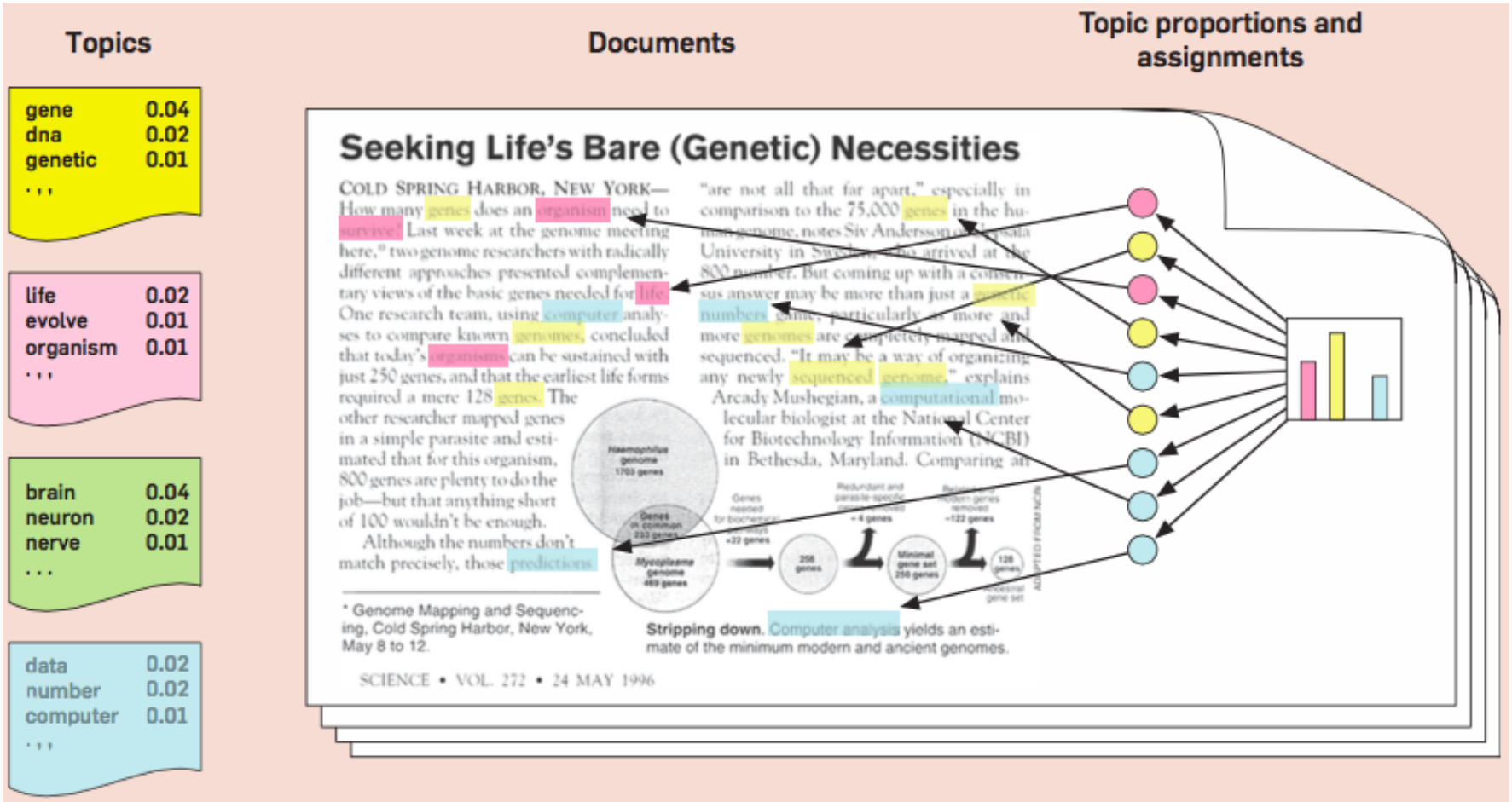
We have used information-theoretic ideas to derive a class of practical and nearly optimal schemes for adapting the size of a neural network. By removing unimportant weights from a network, several improvements can be expected: better generalization, fewer training examples required, and improved speed of learning and/or classification. The basic idea is to use second-derivative information to make a tradeoff between network complexity and training set error. Experiments confirm the usefulness of the methods on a real-world application.

1 INTRODUCTION

Most successful applications of neural network learning to real-world problems have been achieved using highly structured networks of rather large size [for example (Waibel, 1989; LeCun et al., 1990)]. As applications become more complex, the networks will presumably become even larger and more structured. Design tools and techniques for comparing different architectures and minimizing the network size will be needed. More importantly, as the number of parameters in the systems increases, overfitting problems may arise, with devastating effects on the generalization performance. We introduce a new technique called Optimal Brain Damage (OBD) for reducing the size of a learning network by selectively deleting weights. We show that OBD can be used both as an automatic network minimization procedure and as an interactive tool to suggest better architectures.

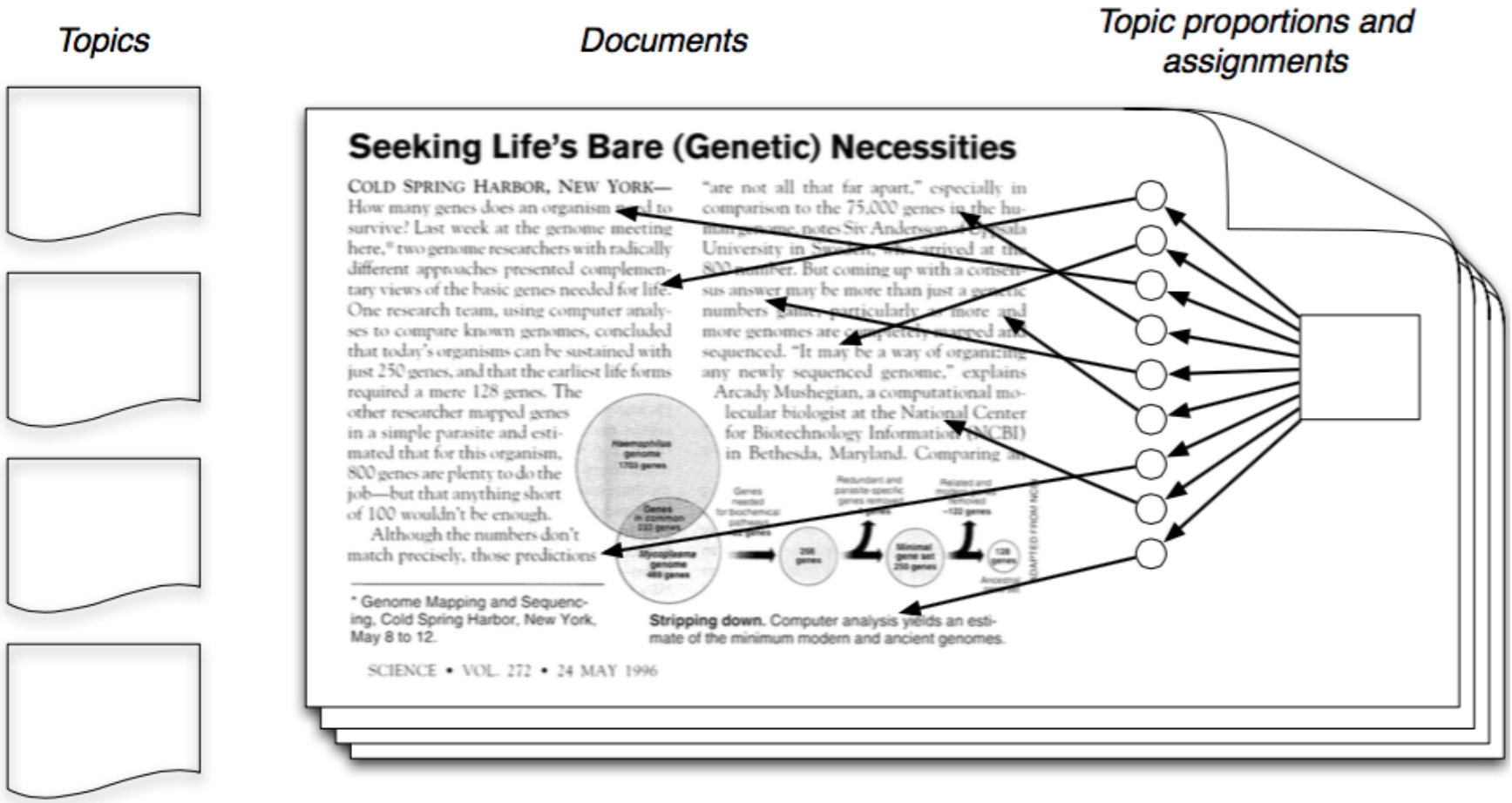
Generative model

- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

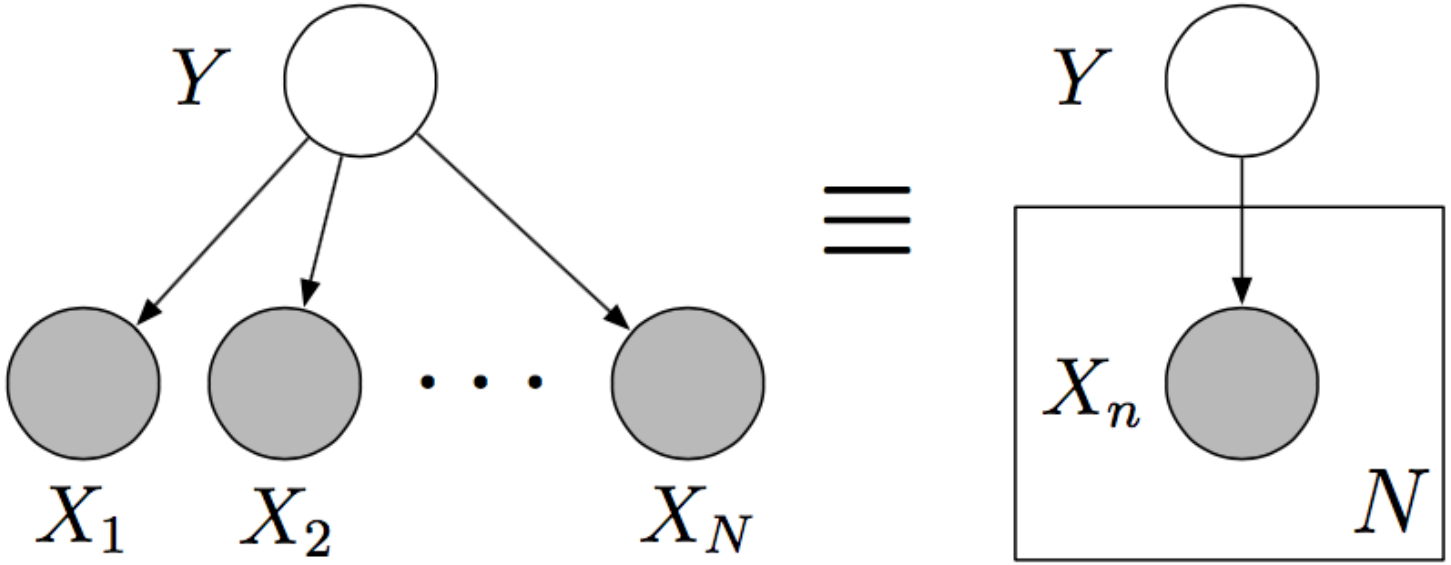


The posterior distribution

- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

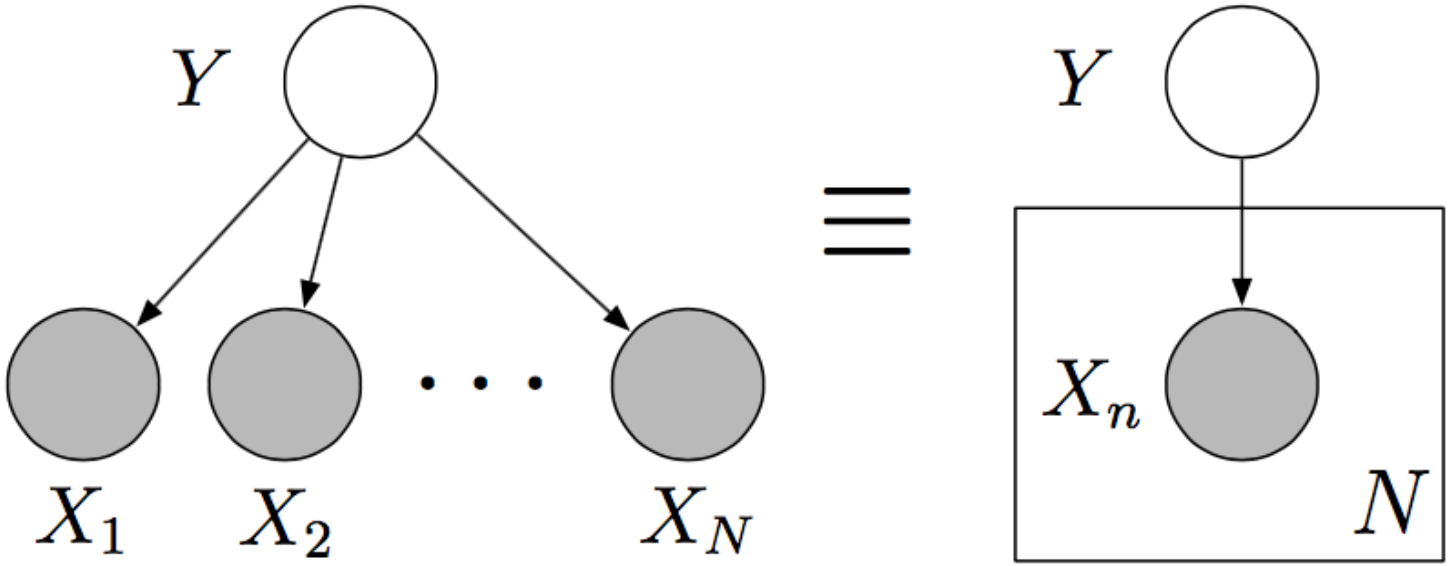


Graphical models (Aside)



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

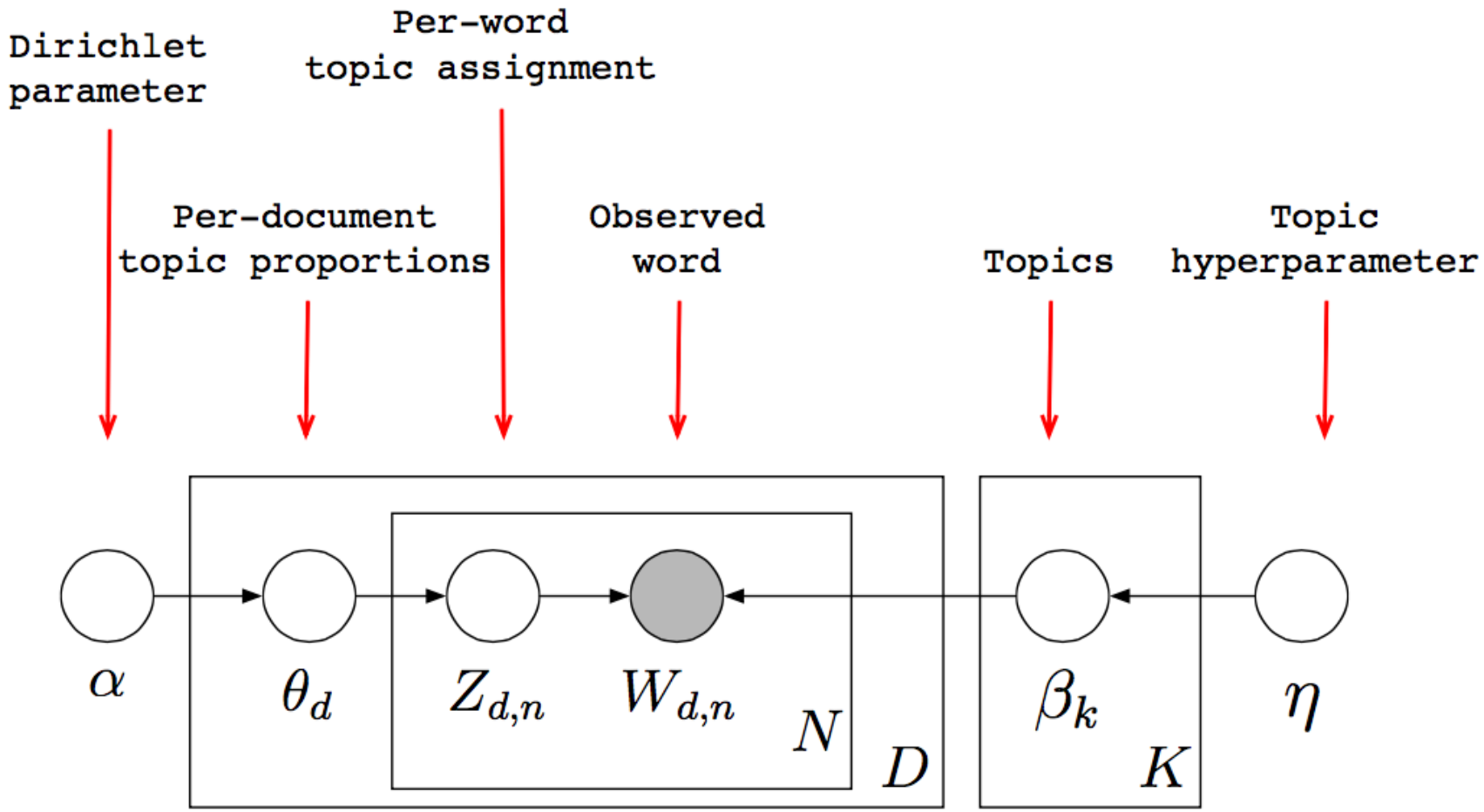
Graphical models (Aside)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n|y)$$

Latent Dirichlet allocation



Each piece of the structure is a random variable.

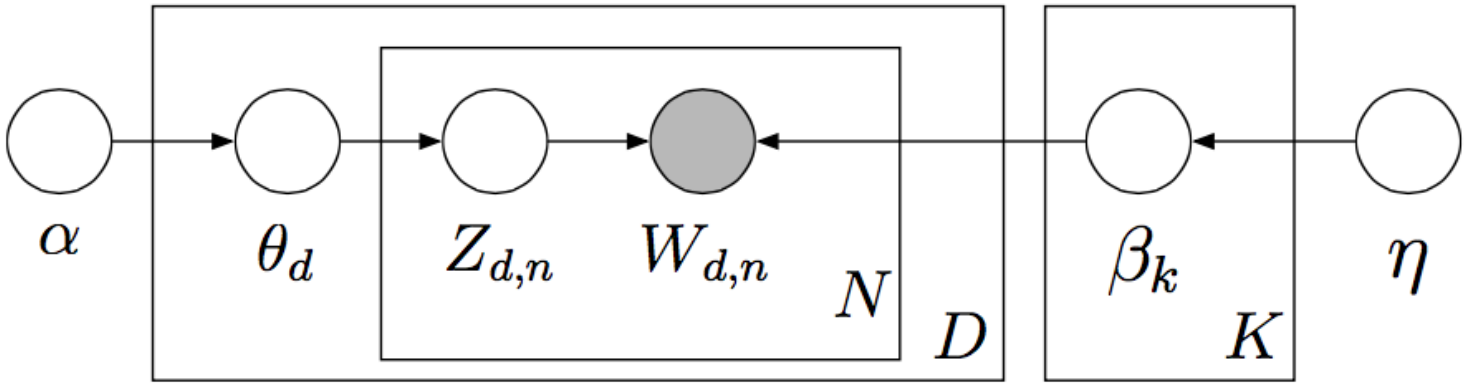
The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

$$p(\theta|\vec{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

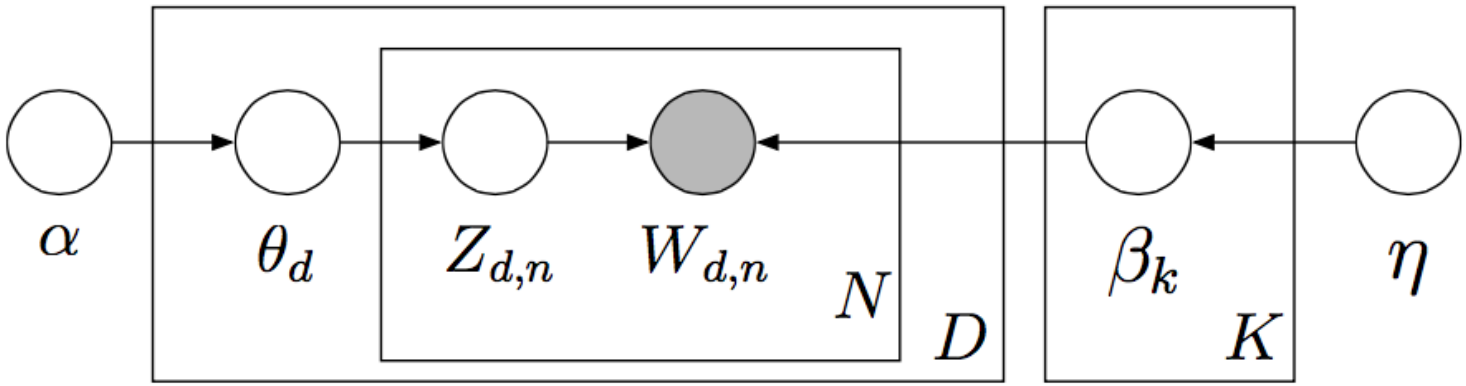
- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet. The topics are a V dimensional Dirichlet.

Latent Dirichlet allocation



- LDA is a mixed membership model.
- For document collections and other grouped data, this might be more appropriate than a simple finite mixture.
- The same model was independently invented for population genetics analysis .

Latent Dirichlet allocation



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distribution β_k
- Use posterier expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

Why does LDA work?

Why does the LDA posterior put "typical" words together?

- Word probabilities are maximized by dividing the words among the topics.
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet on the topic proportions can encourage sparsity, i.e., a document is penalized for using many topics.
- Loosely, this can be thought of as softening the strict definition of "co-occurrence" in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

LDA is modular, general, useful

- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- The data generation distribution can be changed.
- The posterior can be used in creative ways
- Loosely, this can be thought of as softening the strict definition of "co-occurrence" in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

Supervised topic models

- But LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of "diggs"
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised topic models** are topic models of documents and responses, fit to find topics predictive of the response.
  - sLDA

Outline

Generative vs Discriminative

Two approaches to classification:

- Discriminative classifiers estimate parameters of decision boundary/class separator directly from labeled sample
  - learn boundary parameters directly (logistic regression models, $p(y|x)$)
  - learn mappings from inputs to classes (least-squares, neural nets)
- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier)
  - build a model of $p(x|y)$
  - apply Bayes rule

The generative process

To generate a document:

1. Randomly choose a distribution over topics
2. For each word in the document
   a. randomly choose a topic from the distribution over topics
   b. randomly choose a word from the corresponding topic (distribution over vocabulary)

- Note that we need a distribution over a distribution
- Note that words are generated independently of other words (unigram bag-of-words model)

Thank You!

Next lecture: Deep Learning