Probabilistic Topic Models
Tutorial: COMAD 2011

Indrajit Bhattacharya
Assistant Professor
Dept of Computer Sc. & Automation
Indian Institute Of Science, Bangalore

My Background

- Interests
  - Topic Models
  - Probabilistic Graphical Models, Non parametric Models
  - Semi-supervised learning; Learning with Side Information
  - Unstructured Data Analysis, Web Data Analysis
- Probabilistic Graphical Models Lab
- Homepage: www.csa.iisc.ernet.in/~indrajit

Center of Machine Learning, CSA Department, IISc

- Members:
  - 6 faculty members,
  - >25 research students
- Interests:
  - Graphical Models, Kernel Learning,
  - Ranking & Information Retrieval, Learning Theory,
  - Reinforcement Learning, Optimization, Pattern recognition
- Application Areas:
  - Biology, Text and Web Data Analysis, Computer Vision, Networks and Systems Analysis

Managing Text Data

- More text data than structured data
- Immense importance in various applications

- Tasks:
  - Browse: Analyze / Understand
  - Similarity search / Semantic Search
  - Information Extraction
  - Predict properties of new documents

Text Data: Challenges

- Large no of dimensions
  - More than 250K words in Oxford English Dictionary
- Sparse nature
  - Very few distinct words in individual documents
- Dependence among dimensions
  - Synonymy, Polysemy

- Linguistic approaches need resources, maintenance
- Language independent statistical models ?

Latent Dirichlet Allocation

- **"Latent Dirichlet Allocation",** Blei, Ng, Jordan, *Journal Machine Learning Research*, 2003
- ~ 4000 citations (Google Scholar)
- Numerous Applications
  - Document Analysis and Natural Language Processing,
  - Image Processing and Computer Vision,
  - Social Network Analysis, Music Analysis, Biology …
- Research Area:
  - Papers in leading conferences every year

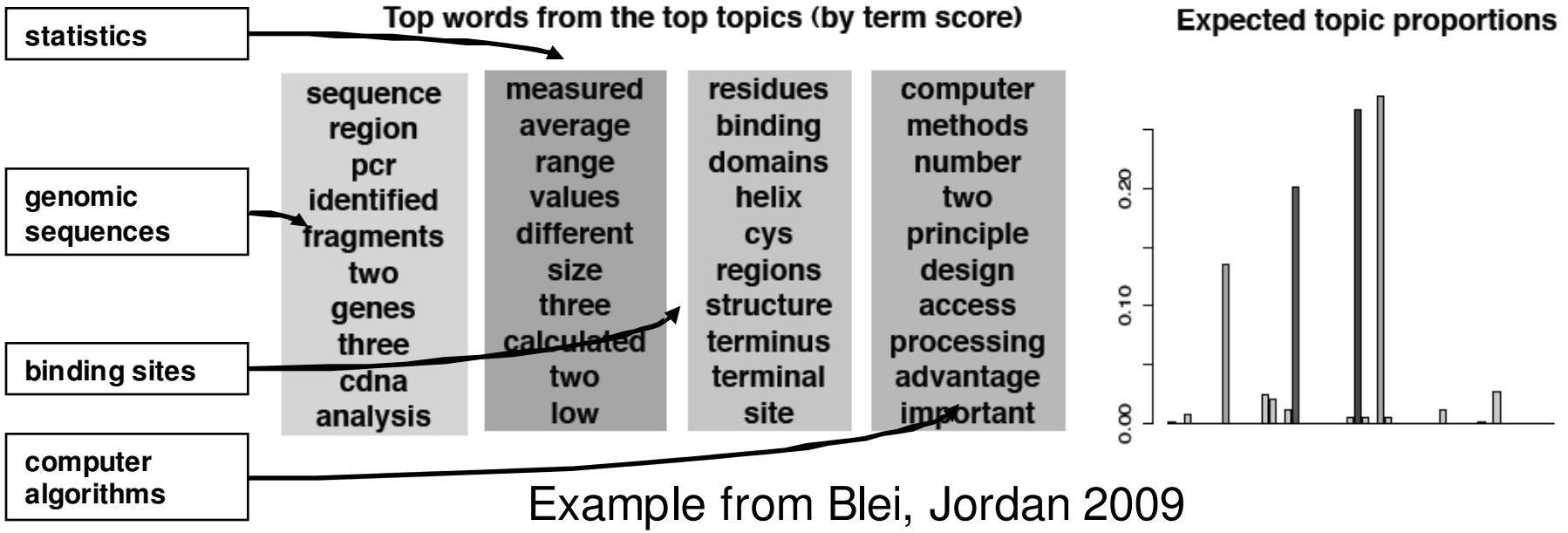
Basic Intuition

- Embed documents in lower dimension space
  - Topic space
- Topic: Combination of word dimensions
  - Brings together “related” words
  - Statistical relations
- Document: Combination of topics
- Perform tasks in the topic space

Example: Topic Analysis of 'Science' Paper

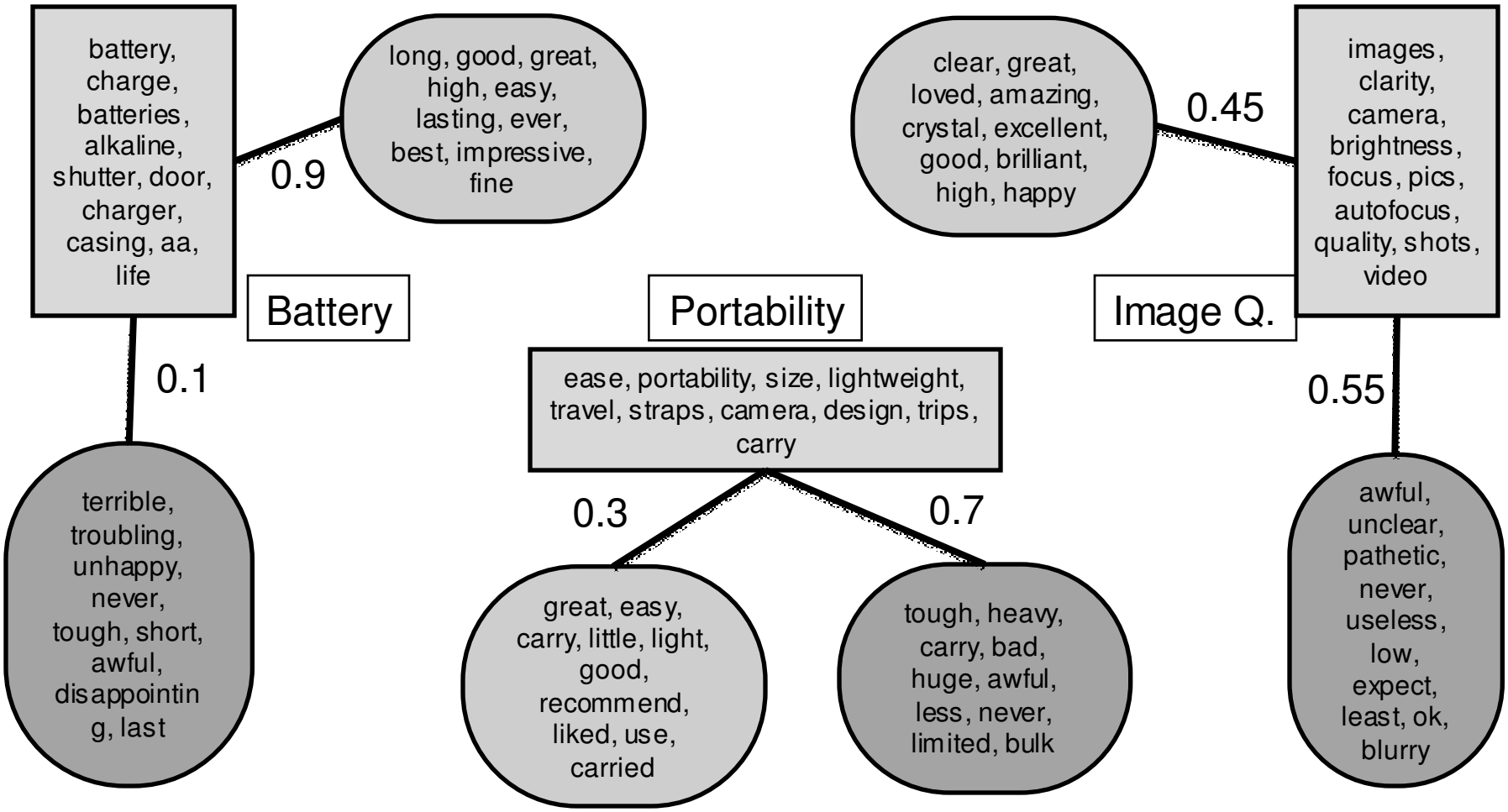
Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.



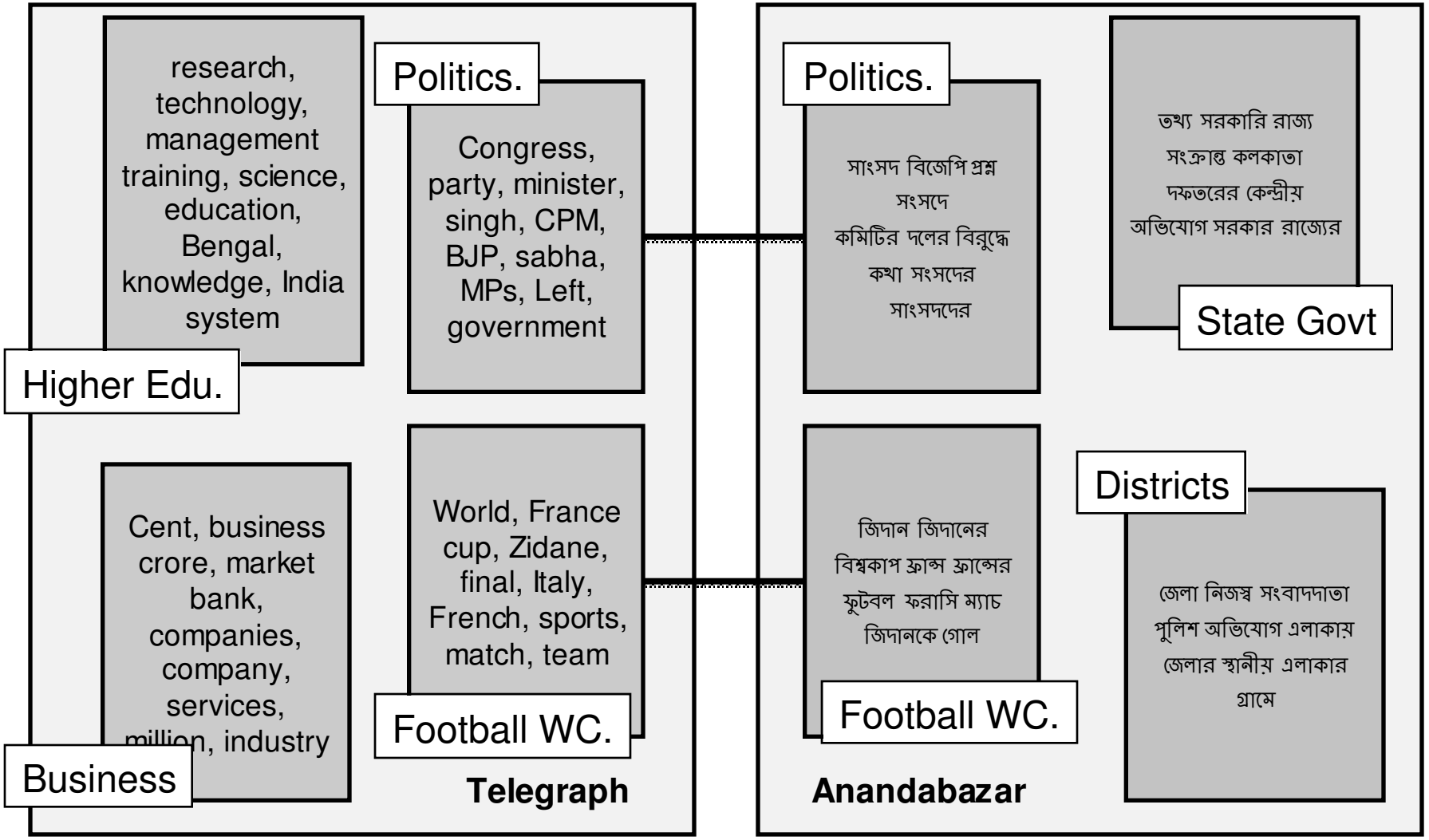
Example from Blei, Jordan 2009

Example: Sentiment Analysis of Product Reviews



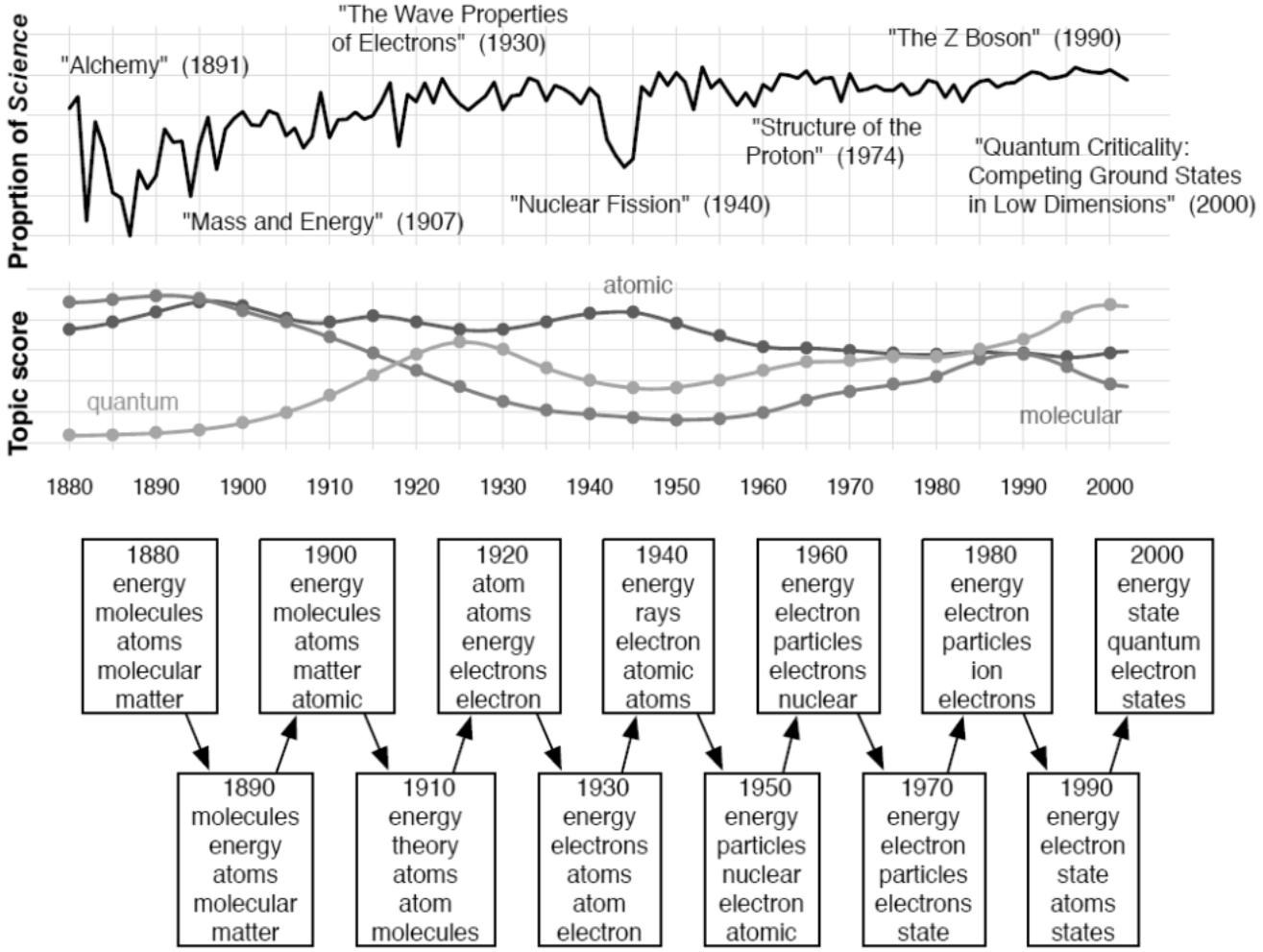
Example from Himabindu et al 2011

Example: Multilingual News Analysis



Example from Dubey et al 2011

Example: Topic Evolution in Science Archive (1880-2002)



Example from Blei, Jordan 2009

Discovering Topics

- Matrix Factorization Approaches
- Latent Semantic Analysis [Deerwester 90]

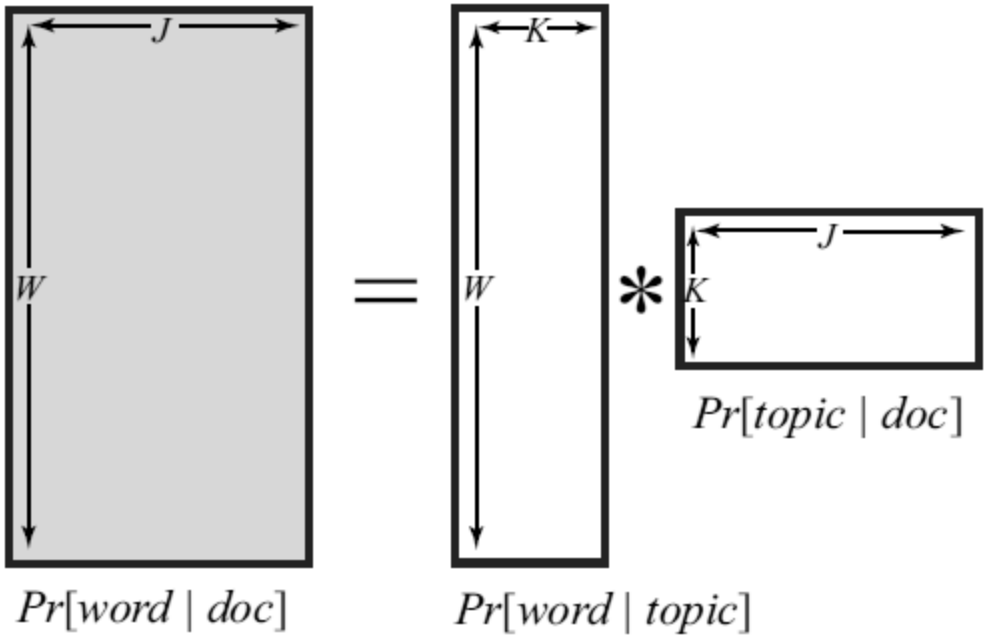


Figure from Sudderth 2006

Why Probabilistic?

- Interpretability
- Host of tools and techniques available
- Priors for limited data settings
- It's fun!

A little bit of background …

- Probabilistic Models, Inference, Learning
- Bayesian, Priors, Conjugates
- Graphical Models, Plate Notation

Basics: Probabilistic Formulation

- Given a sequence of n 0's and 1's
- Predict n+1th value

- Probabilistic Formulation:
- Introduce random variables and distributions
  - Values $x_0 ... x_n$ for binary random variables $X_1 ... X_n$ following unknown joint distribution $P(X_1 ... X_n)$
  - Find $P(X_{n+1} | X_1 ... X_n)$

Basics: Probabilistic Model

- Assume appropriate **independences**

$$X_i \text{ independent of } X_j,\ j \neq i, \Rightarrow P(X_1 \ldots X_n) = P(X_1) \ldots P(X_n)$$

- Assume distribution family

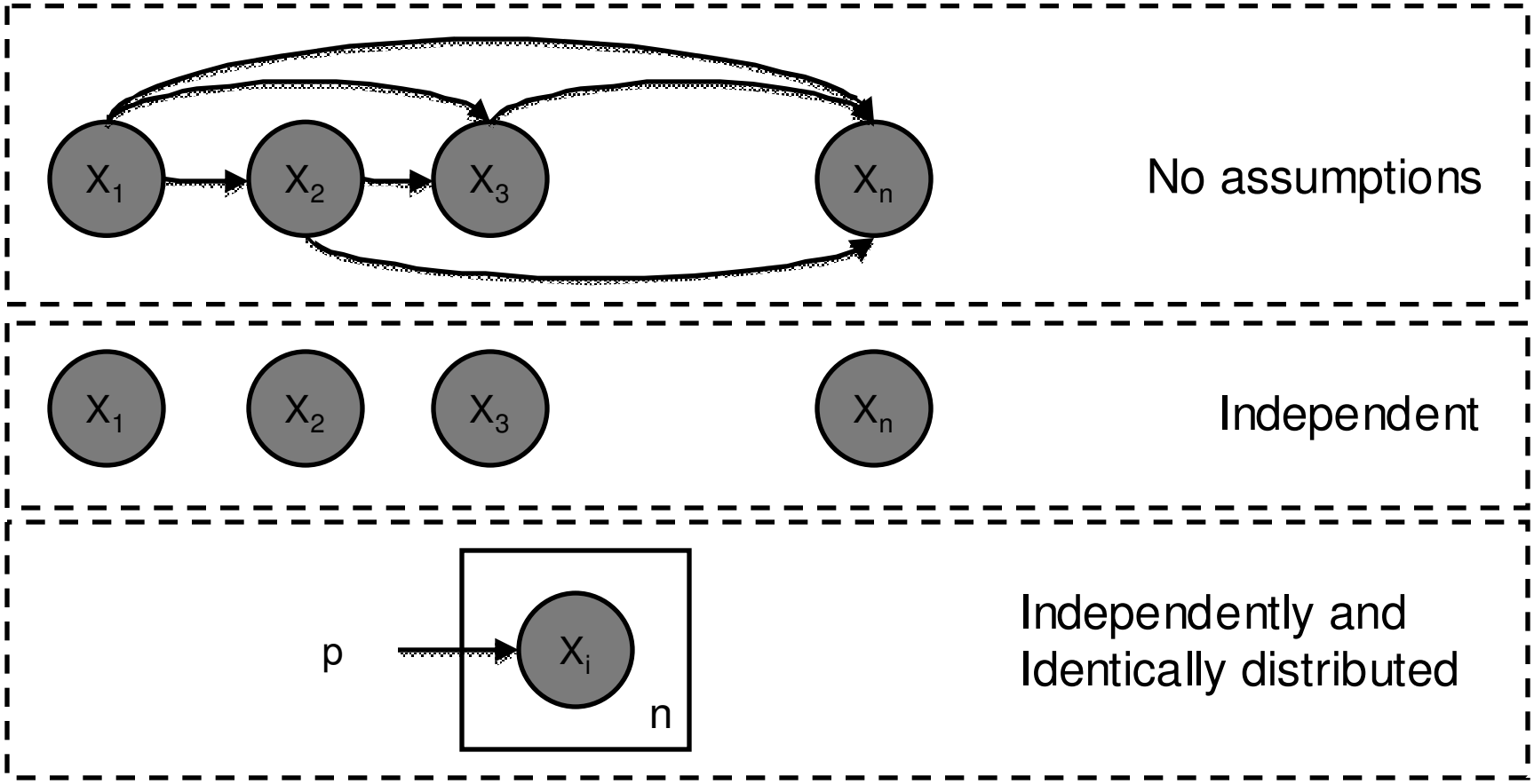
$$(X_n) \sim Ber(p_i)\ \forall i$$

- Assume **identically distributed**

$$(p_i) = p\ \forall i$$

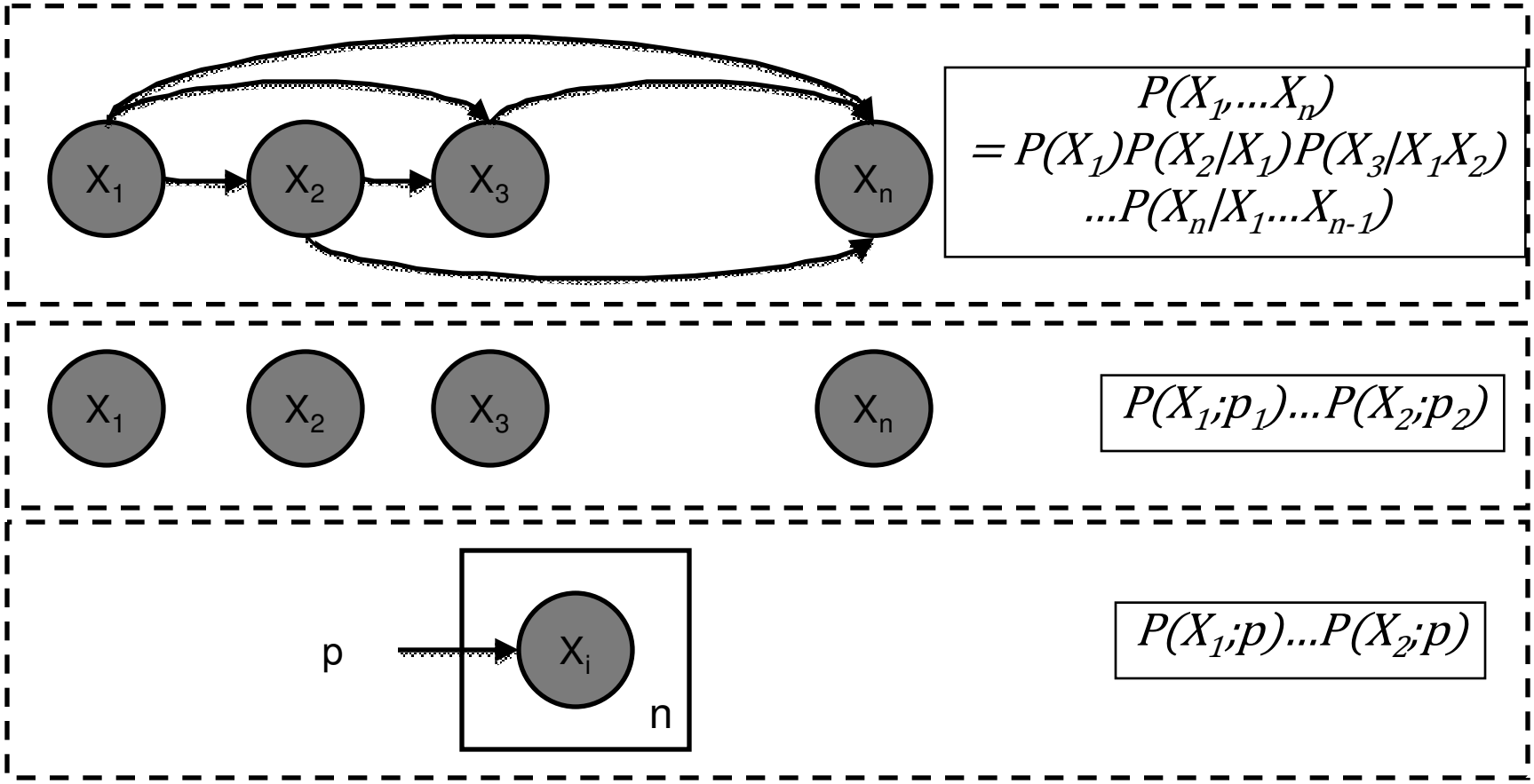
Basics: Graphical Model

- Nodes: Random Variables
- (Missing) Edges: Conditional Independences



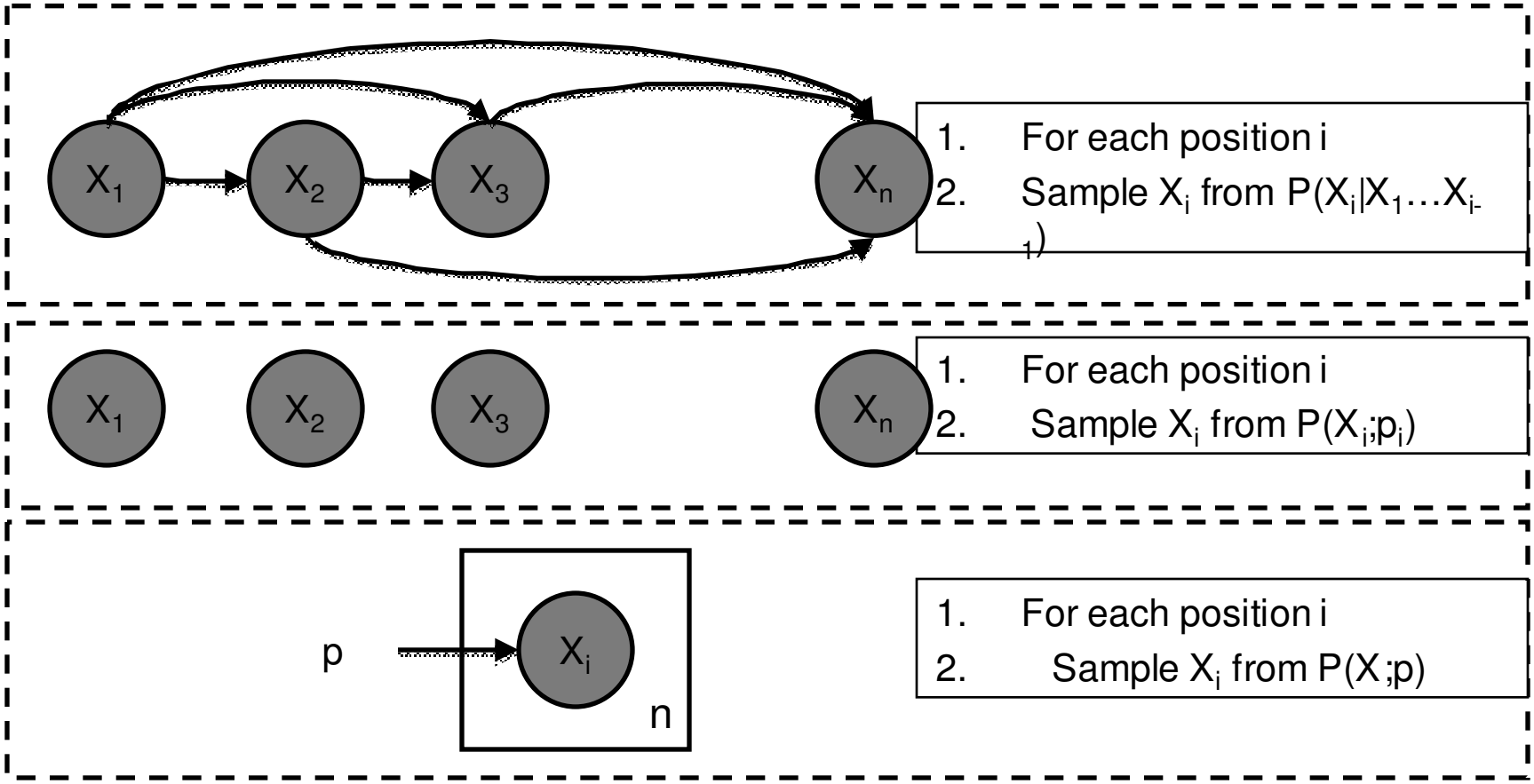
Basics: Factorization

- Graphical model ≡ Factorization of joint distribution



Basics: Generative Process

- Generative Process: Sample from joint distribution
- Directed graphical models ≡ Generative process



Basics: Inference and Estimation

- Given (part of) the outcome of a process

1. Guess process parameters
2. Guess remaining part of the outcome

- “Reversal” of the Generative Process

Basics: Inference

- Using known parameters, find conditional distribution for unobserved variables

$$P(X_{n+1} \mid X_1 \dots X_n; p) = P(X_{n+1}, X_n, \dots, X_1; p) \Big/ P(X_n, \dots, X_1; p)$$
$$= P(X_{n+1}, X_n, \dots, X_1; p) \Big/ \sum_x P(X_{n+1} = x, X_n, \dots, X_1; p)$$

  - Ratio of marginal probabilities

- Computationally hard w/o independences
  - Exponential in tree-width of graph

Basics: Learning / Parameter Estimation

- Find most likely value of the parameters from observed data
  - Guess value of p from $X_1 ... X_n$

- Classical Estimation: Pars unknown constants
  - Maximum Likelihood Estimation

$$\hat{p} = \sum_{i=0}^{n} X_i / n$$

  - Counts are sufficient statistics
  - 'Correctness' guarantees given infinite data

Basics:
Bayesian Parameter Estimation

- Model parameter p also as a random variable
- Prior distribution $P(p)$ captures prior belief

$$p \sim Beta(a,b) \equiv P(p) \propto p^{a-1}(1-p)^{b-1}$$

- Find posterior distribution given observed data

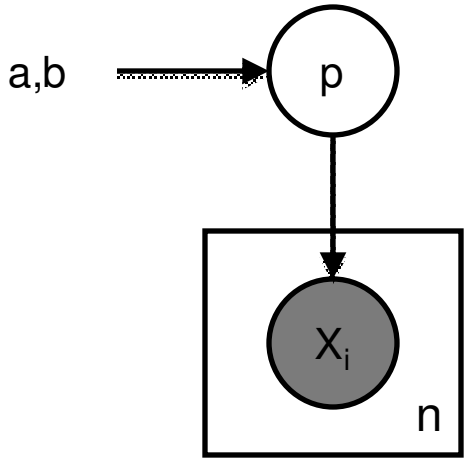
$$P(p|X_1 \dots X_n) = P(p)P(X_1 \dots X_n|p) \Big/ \int P(p)P(X_1 \dots X_n|p)dp$$

- Bayesian Estimator

$$E[p|X_1 \dots X_n] = \sum_{i=0}^{n} X_i + a \Big/ n + a + b$$

Basics:
Bayesian Parameter Estimation

- Graphical model and Generative Process



1. Sample p from Beta(a,b)
2. For each position i
3. Sample X_i from Ber(p)

Basics: Exchangeability

- Can we justify the model assumptions?
- Exchangeability
  - Set of random variables is *exchangeable* if any *permutation* has the *same probability*

- DeFinetti's Theorem
  - The joint distribution of (infinitely) exchangeable random variables can always be expressed as

$$P(X_1 \ldots X_n) = \int_\theta \; p(\theta) \prod_i f(X_i|\theta)\, d\theta$$

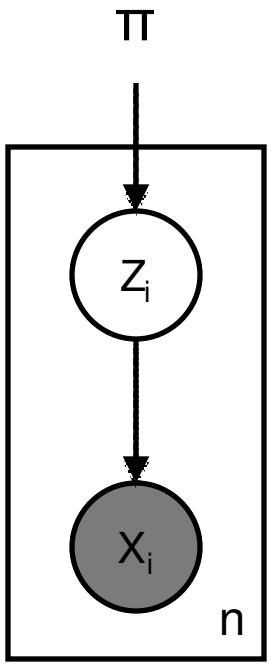
  - Conditional independence given parameter

Basics: Learning from Partial Observations

- Sequence of n 0's and 1's, generated using k different coins
- Mixture of Bernoulli distributions

- Random variables
  - Z_i : Index of coin used for i^{th} toss
  - X_i : Outcome of i^{th} toss

Basics:
Learning from Partial Observations



$$P(X_1, Z_1 \ldots X_n, Z_n, \pi) = P(\pi; a_1 \ldots a_k) \prod_{k=1}^{n} P(Z_k|\pi)\, P(X_k|Z_k)$$

1. Sample π from Dir(a_1,…,a_k)
2. For each position i
3. Sample Z_i *from Mult(π)*
4. Sample X_i from Ber(p_{zi})

Basics:
Learning from Partial Observations

- Mixture labels are unobserved in data!

$$P(X_1 \dots X_n; \pi) = \sum_{Z_1 \dots Z_n} P(Z_1, X_1 \dots Z_n, X_n, \pi)$$

- Parameter Estimation: No closed form solutions

Basics: Expectation Maximization (EM)

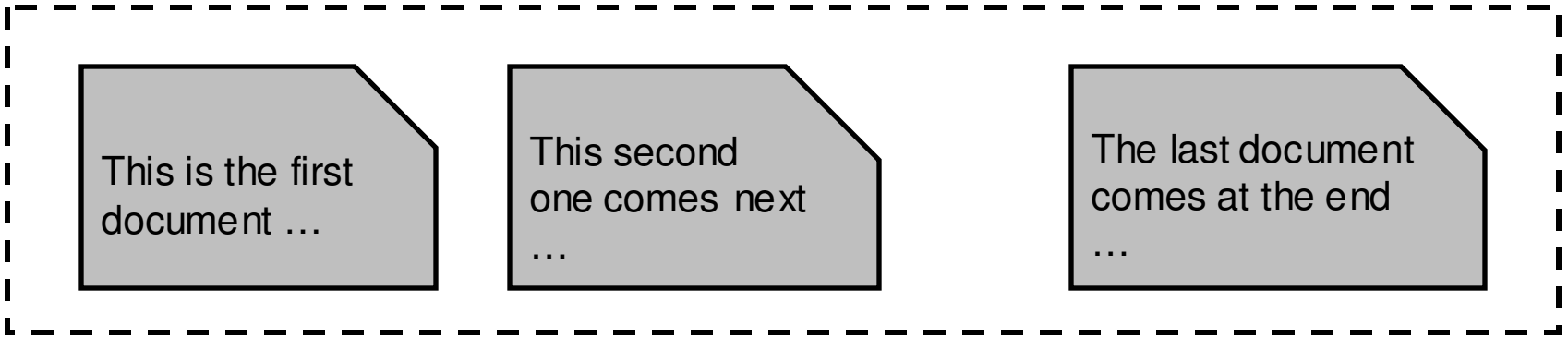
- Iterative solution
  - Estimate parameters using expected counts (M-step)
  - Infer hidden variables using estimated parameters (E-step)

- Provable convergence to local maximum
  - Coordinate ascent algorithm

Now back to Topic Models …

Modeling Textual Data

- Document Corpus:
  - Sequence of M documents
  - Document is sequence of N_i words



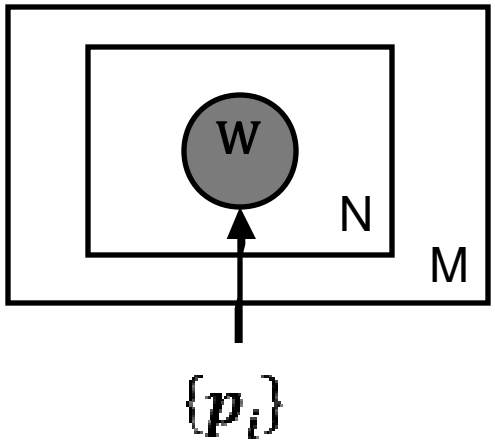
- Observed random variables:
  - $W = \{W_{11}, W_{12} \ldots W_{1N}, \ldots, W_{m1}, W_{m2} \ldots W_{mN}\}$
  - W_{ij} takes values from finite vocabulary of V words

Attempt 1: Unigram Model

- Extend coin toss model for V-dimensional discrete random variables

1. For each document i
2. For each word j
3. Sample W_{ij} from Mult($p_1 \ldots p_v$)

$$P(W) = \prod_{i=1}^{M} \prod_{j=1}^{N} P(W_{ij})$$



Unigram Model: Properties

- Single multinomial distribution over words

$$P(W_{ij}=k) = p_k \; ; \; \phi_1 = \{p_1 \ldots p_V\}$$

- Topic: Multinomial distribution over vocabulary

- Advantages:
  - Closed form learning, inference
- Problems:
  - Single topic for entire corpus
  - No distinction between documents
  - Word order does not matter, even across documents!

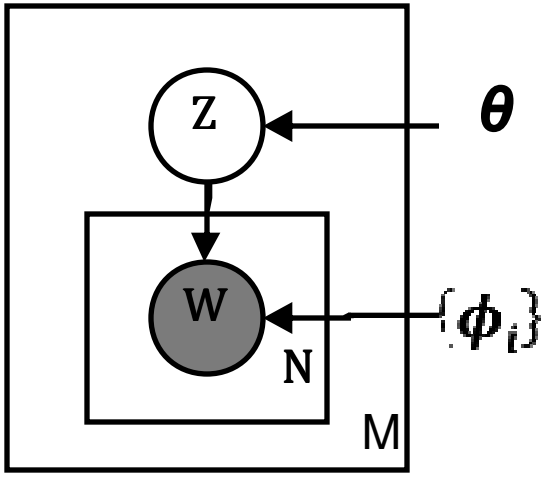
Attempt 2: Mixture of Unigrams

- Assume T multinomial distributions / topics

$$\phi = \{\phi_1 \ldots \phi_T\}, \quad \phi_t = \{p_{i1} \ldots p_{iN}\}$$

- (Hidden) Topic variable Z_i for each document
  - (T-dim) Multinomial distribution $\mathrm{Mult}(\theta)$ over Z_i

1. For each document i
2. Sample Z_i from Mult(θ)
3. For each word j
4. Sample W_{ij} from Mult(ϕ_{Z_i})



Mixture of Unigrams: Properties

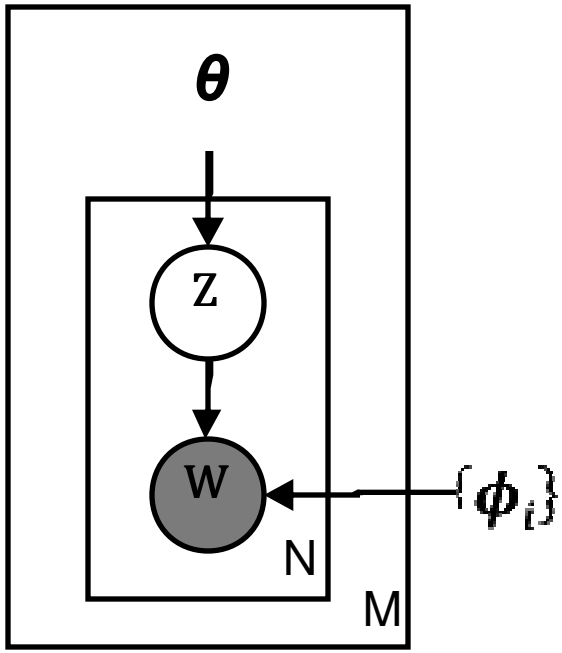
- Price:
  - Inference gets harder (but manageable)
  - EM for Parameter Estimation

- Issues:
  - Word order does not matter within documents
  - Exactly one topic for each document; too simplistic

Attempt 3: Probabilistic Latent Semantic Analysis (pLSA)

- Allow multiple topics for each document
  - Each word can come from a different topic
- Document specific "mixture" over topics: θ_i
- $\theta_{ij} = P(j^{th}\ topic\ in\ i^{th}\ doc)$

1. For each document i
2. For each word j
3. Sample Z_{ij} from $Mult(\theta_i)$
4. Sample W_{ij} from $Mult(\varphi_{Z_{ij}})$



pLSA: Properties

- Price:
  - Learning and inference get harder
  - Generic EM gives suboptimal solutions

- Issues:
  - No generative process for mixture of topics
  - No (principled) way to deal with new unseen documents
  - No. of parameters grows (linearly) with no. of docs

Latent Dirichlet Allocation (LDA)

- Model topic mixture as random variable
- Use DeFinetti Theorem for factorization

$$P(W_i) = \int_{\theta_i} p(\theta_i) \prod_i f(W_{ij}|\theta_i)\, d\theta_i$$

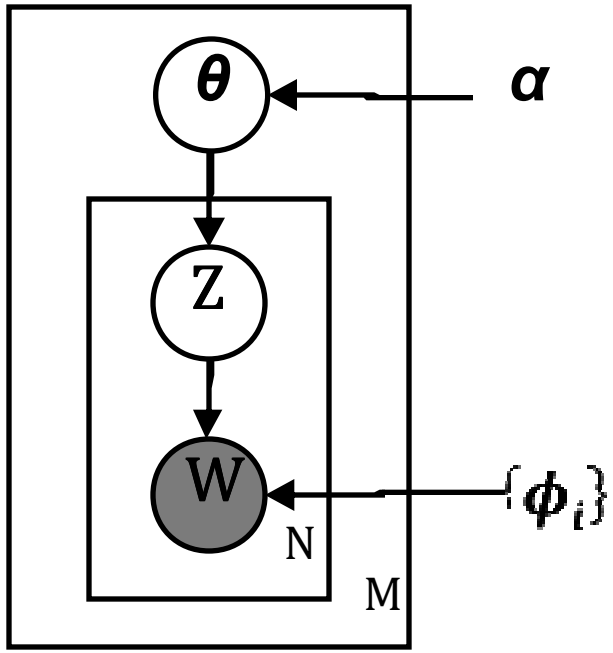
- Dirichlet Prior: $Dir(\alpha_1, \ldots \alpha_T)$

$$p(\theta_i) \propto \theta_{i1}^{\alpha_1 - 1} \theta_{i2}^{\alpha_2 - 1} \ldots \theta_{iT}^{\alpha_T - 1}$$

  - Conjugate for Multinomial
  - Able to incorporate domain knowledge

LDA: Generative Process

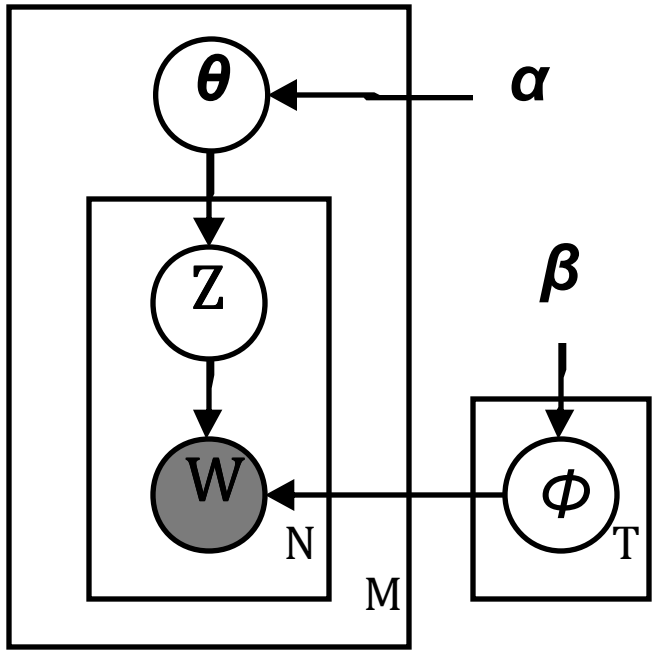
1. For each document i
2. Sample θ_i from $Dir(\alpha)$
3. For each word j
4. Sample Z_{ij} from $Mult(\theta_d)$
5. Sample W_{ij} from Mult($\phi_{Z_{ij}}$)



LDA: Smoothing Topics

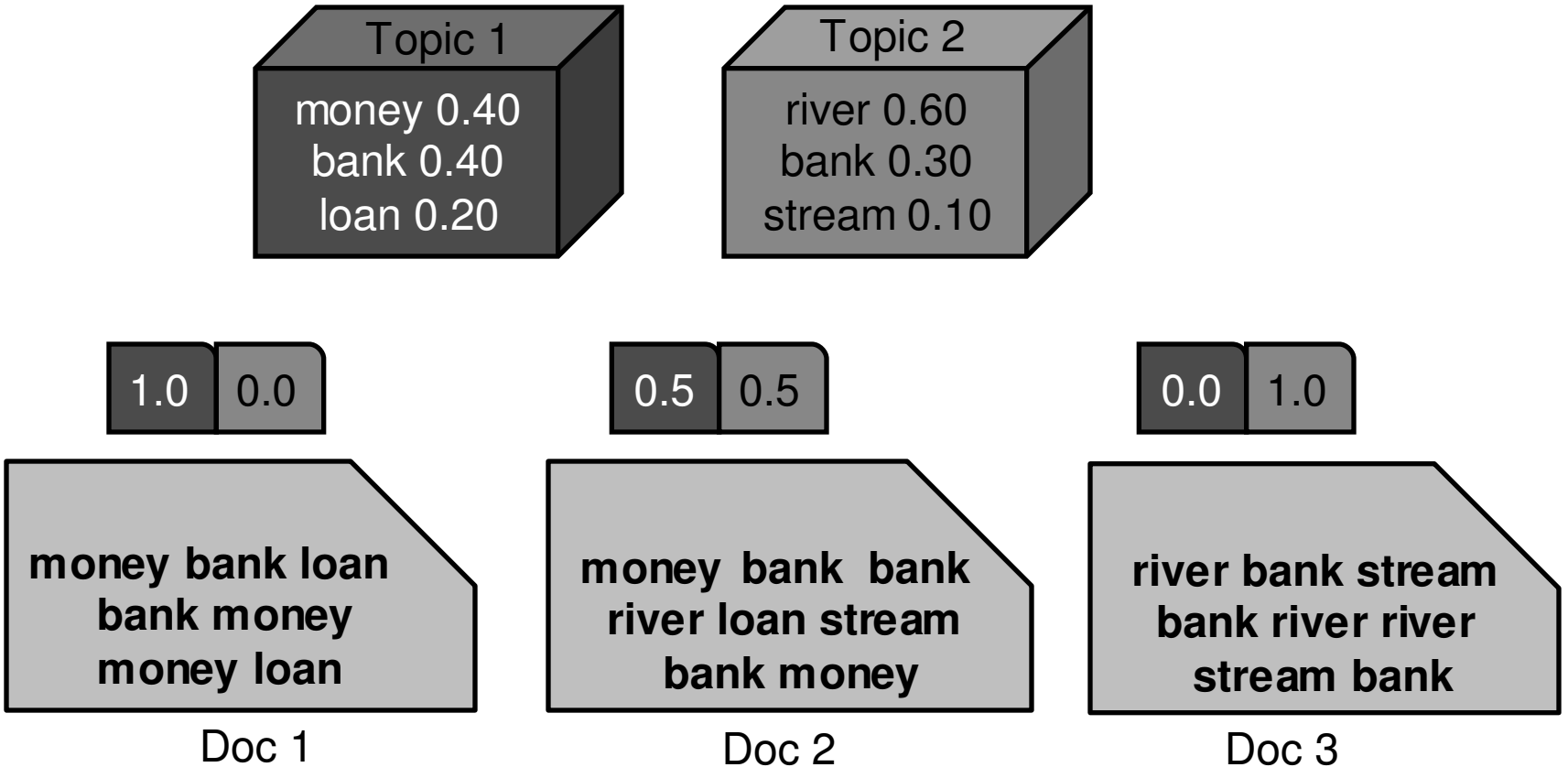
1. For each topic i
2. Sample Φ_i from $Dir(\beta)$

3. For each document I
4. Sample θ_i from $Dir(\alpha)$
5. For each word j
6. Sample Z_{ij} from $Mult(\theta_d)$
7. Sample W_{ij} from Mult($\varphi_{Z_{ij}}$)



- DeFinetti Theorem for Corpus as exchangeable collection of documents

LDA Generative Process: Example



Example from Giffiths and Steyvers 2009

LDA: Role of Dirichlet Prior

$$p(\theta_i) \propto \theta_{i1}^{\alpha_1 - 1} \theta_{i2}^{\alpha_2 - 1} \ldots \theta_{iT}^{\alpha_T - 1}$$

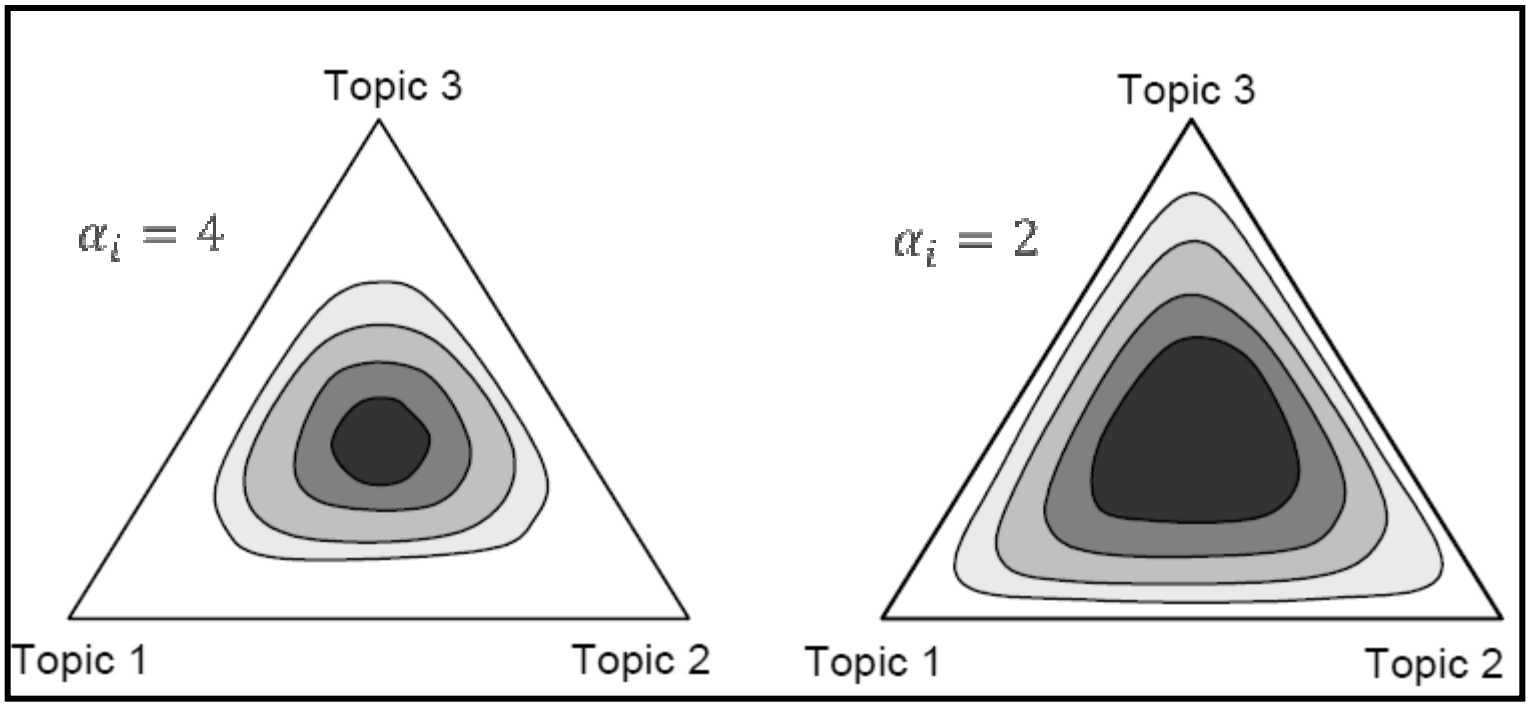
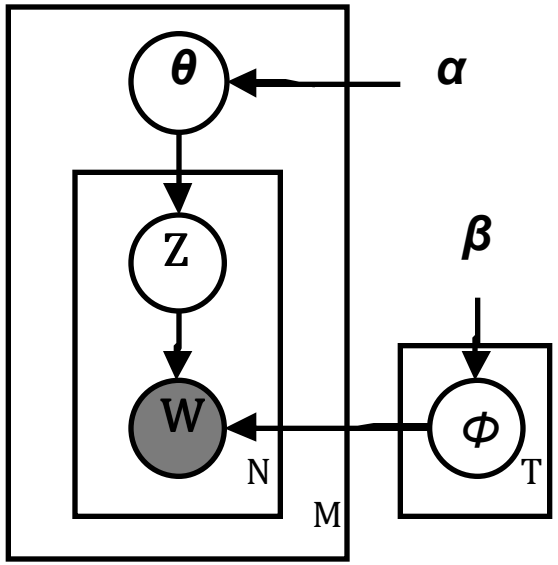


Figure from Giffiths and Steyvers 2009

LDA: Factorization

$$P(W, Z, \theta, \phi; \alpha, \beta)$$
$$= \left(\prod_{t=1}^{T} P(\phi_t; \beta) \right) \prod_{d=1}^{M} \left(P(\theta_d; \alpha) \prod_{j=1}^{N} P(Z_{dj}|\theta_d) P(W_{dj}|Z_{dj}, \phi) \right)$$



LDA: Inference and Learning

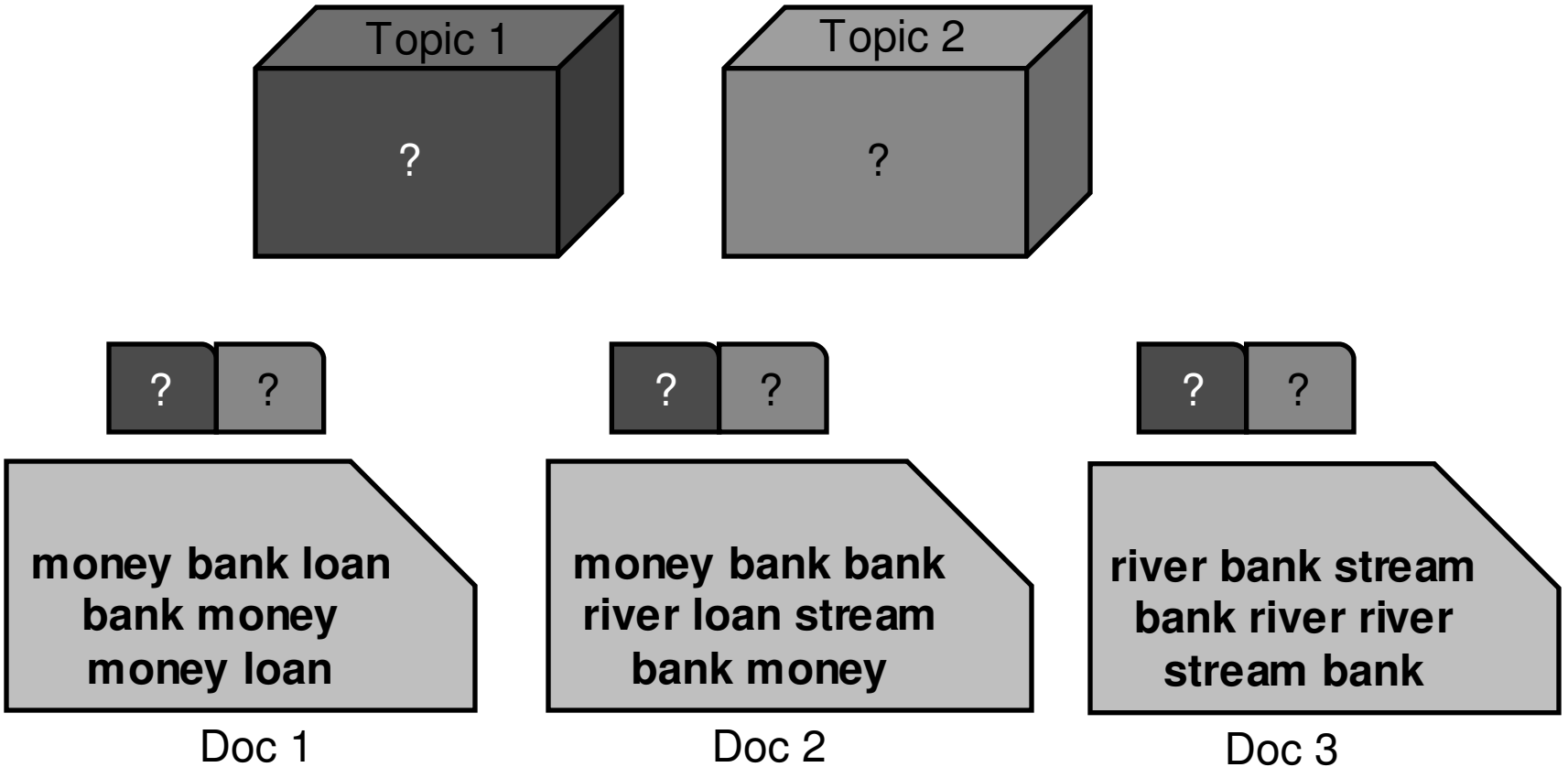


Figure from Giffiths and Steyvers 2009

LDA: Inference

- Topics known; determine
  1. Topic mixture for document, and
  2. Topics of each word

$$P(\theta, Z|W, \phi, ; \alpha, \beta) = P(\theta, Z, W, \phi; \alpha, \beta) \Big/ P(W, \phi; \alpha, \beta)$$

$$P(W, \phi; \alpha, \beta) = \int_{\theta} \sum_{Z} P(\theta, Z, \phi, W; \alpha, \beta)$$

- Intractable due to coupling between θ and ϕ

LDA: Learning

- Determine T topics $\{\phi_i\}$

$$P(\phi|W;\alpha,\beta) = P(\theta,Z,W,\phi;\alpha,\beta) \Big/ \int_{\phi} P(W,\phi;\alpha,\beta)$$

- Use EM algorithm
  - Involves inference as a sub-task
- More intractable!

LDA: Approximate Inference

- Exact inference is intractable
- Approximate Inference
  - Variational Methods
  - Sampling Methods
  - Others
- Approximate Learning
  - Perform approximate inference in E-step of EM

LDA: Conditional Distribution

- Conditional distribution of topic for a specific word in a document, given topic assignments to all other words in all documents

$$P(Z_i = j \mid Z_{-i}, W_i, D_i) \propto \frac{n^{WT}_{W_i j} + \beta}{\sum_w n^{WT}_{wj} + W\beta} \times \frac{n^{DT}_{D_i j} + \alpha}{\sum_t n^{WT}_{D_i t} + T\alpha}$$

- Straight forward to compute by maintaining counts
- Suggests simple iterative algorithm …

LDA: Collapsed Gibbs Sampling

```
1.  (Randomly) Initialize topics
2.  Initialize <W,T> and <D,T>counts
3.  Repeat until convergence
4.    For each word of each document
5.    Sample new topic from cond distribution
6.    Update <W,T> and <D,T> counts
```

- Reject initial samples (Burn in period)
- Sample topics to estimate posterior distribution

Does this work?

- Monte Carlo Integration
  - Sample average approximates Expectation
- Markov Chain Monte Carlo
  - Sample from Markov Chain with target distribution as stationary distribution
- Gibbs Sampling
  - Sample each latent variable from its conditional distribution
  - Ergodic Markov Chain
- Caveats
  - Convergence
  - Uncorrelated samples

Estimating Topics

- Estimate topics ϕ_t (and θ) from the topic assignments of words

$$\phi_t^j = \frac{n_{ij}^{WT} + \beta}{\sum_k n_{kj}^{WT} + W\beta}$$

$$\theta_j^d = \frac{n_{dj}^{DT} + \alpha}{\sum_k n_{dk}^{DT} + T\alpha}$$

- Conditional expectation of ϕ_t (and θ) given a sample (topic assignment to words)

Implementation Issues (1)

- Detecting Convergence
  - Techniques exist; but hard in general
  - Fixed (large) number of iterations
- One long chain vs Many short chains
  - Latter preferred
- Initialization
  - Critical in determining time to convergence
  - Random; Sample from conditional given previous

Implementation Issues (2)

- Determining number of topics
  - Bayesian model selection
  - Best generalization; perplexity measure
  - Non-parametric techniques

- Setting (Hyper-) Parameters
  - Possible to estimate; slow in general
  - β : Controls granularity and sparsity of topics;
  - α : Controls of topic sparsity of documents
  - β=0.1, α=50/T typical

Implementation Issues (3)

- Choosing vocabulary
  - Typically not useful to include all words
  - Top K words by frequency or TF-IDF scores

- Speeding up and Parallelization
  - Lot of recent work

Analyzing Topics

- Visualizing a topic ϕ_t: What is this topic about?
  - Top k words in decreasing order of probability ϕ_t^j
  - Alternatively, $\phi_t^j \log\left(\frac{\phi_t^j}{\left(\prod_{t=1}^{T} \phi_t^j\right)^{1/T}}\right)$
- Visualizing a document: What is this document about?
  - Top k topics in decreasing order of probability θ_d^j
  - Posterior topic assignments Z_i of individual words

Application of Topics

- Finding similar documents
  - For two docs d_1 and d_2, compare topic mixtures θ_{d_1} , θ_{d_2}
  - KL Divergence, Hellinger Distance
- Finding similar words
  - Membership in same topic(s)
- Classification
  - Instead of word vector, use the topic mixture θ_d as the representation of document
  - Significant improvement with limited training data [Blei 03]

Application of Topics

- Semantic search

$$P(d_i|q) \propto P(q|d_i)$$

$$= \prod_{w \in q} P(w|\, d_i) \;= \prod_{w \in q} \sum_{t} P(w \mid z) P(z \mid \theta_{d_i})$$

  - Consider possible topics for each word in the query
  - Check proportion of that topic in doc

  - Document can be relevant without any of the query words appearing in it

LDA Shortcomings and Extensions

- Words in document may not be exchangeable
  - HMM LDA [Griffiths 05]
- Documents in corpus may not be exchangeable
  - Dynamic Topic Models [Blei 06]
- Topics in a document may not be independent
  - Correlated Topic Models [Blei 06]
  - Hierarchical Topic Models [Blei 03]

LDA Shortcomings and Extensions

- Topics may not be coherent [Chang et 09]
  - Regularized Topic Models [Newman 11]
- Topics not optimized for any specific task
  - Supervised LDA [Blei 07], DiscLDA[Lacoste-Julien 08]
- Number of topics hard to guess
  - Non parametric models
  - Hierarchical Dirichlet Processes [Teh 06]

LDA Shortcomings and Extensions

- Data may be continuous
  - LDA extends naturally with appropriate distributions
  - Gaussian topics for spatial data [Sudderth 06]

Resources: Implementation

- LDA-C: Princeton
  - www.cs.princeton.edu/~blei/lda-c
- Mallet: UMass
  - Java; Gibbs Sampling
  - mallet.cs.umass.edu
- R Code: Princeton
  - www.cs.princeton.edu/~blei/topicmodeling.html
- Matlab TM Toolbox: UCI
  - psiexp.ss.uci.edu/research/programs_data/toolbox.htm
- Multi-threaded LDA: Nallapati
  - sites.google.com/site/rameshnallapati/software
- pLDA: Google
  - code.google.com/p/plda
- Online LDA: Princeton
  - www.cs.princeton.edu/~blei/topicmodeling.html
- Hadoop LDA: Yahoo!
  - github.com/shravanmn/Yahoo_LDA

Other Online Resources

- Topic Models mailing list
  - lists.cs.princeton.edu/mailman/listinfo/topic-models
- David Blei's Topic Models page
  - www.cs.princeton.edu/~blei/topicmodeling.html
- David Mimno's Bibliography on Topic Models
  - www.cs.princeton.edu/~mimno/topics.html

References

- *Latent Dirichlet Allocation*, Blei et al., JMLR,(2003).
- *Finding scientific topics*, Griffiths & Steyvers, PNAS , (2004).

- *Topic Models*, Blei & Lafferty, (2009)
- *Probabilistic Topic Models*, Steyvers & Griffiths, (2007)
- *Tutorial on Topic Models*, Blei, SIGKDD, (2011)

- *Probabilistic Inference Using Markov Chain Monte Carlo Methods,* Neal (1993)
- *Graphical Models Course Homepage* (2011), www.csa.iisc.ernet.in/~indrajit/HTML/269S11.html

(Additional) References

- *Indexing by latent semantic analysis*, Deerwester et al, JASIST (1990)

- *Exploiting Coherence for the simultaneous discovery of latent facets and associated sentiments*, Himabindu et al, SDM (2011)
- *Learning Dirichlet Processes from Partially Observed Groups*, Dubey et al, ICDM (2011)

- PhD Thesis, Eric Sudderth, MIT (2006)

- *Improving Topic Coherence with Regularized Topic Models*, Newman et al, NIPS (2011)

Questions?

indrajit@csa.iisc.ernet.in