Natural Language Processing
Lecture 11: Topic Modeling

Qun Liu, Valentin Malykh
Huawei Noah's Ark Lab

Spring 2022
A course delivered at KFU, Kazan

Content

Content

Generative vs. Discriminative Models

- Recall that, in Bayesian networks, there could be many different, but equivalent models of the same joint distribution

Discriminative Generative

- Although these two models are equivalent (in the sense that they imply the same independence relations), they can differ significantly when it comes to inference/prediction

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models

Discriminative Generative

- Generative models: we can think of the observations as being generated by the latent variables
  - Start sampling at the top and work downwards
  - Examples?

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models

Discriminative Generative

- Generative models: we can think of the observations as being generated by the latent variables
  - Start sampling at the top and work downwards
  - Examples: HMMs, naïve Bayes, LDA

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models

Discriminative Generative

- Discriminative models: most useful for discriminating the values of the latent variables
  - Almost always used for supervised learning
  - Examples?

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models

Discriminative Generative

- Discriminative models: most useful for discriminating the values of the latent variables
  - Almost always used for supervised learning
  - Examples: CRFs

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models

Discriminative Generative

- Suppose we are only interested in the prediction task (i.e., estimating $p(Y|X)$)
  - Discriminative model: $p(X,Y) = p(X)p(Y|X)$
  - Generative model: $p(X,Y) = p(Y)p(X|Y)$

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Models of Text Documents

- Bag-of-words models: assume that the ordering of words in a document do not matter
  - This is typically false as certain phrases can only appear together
- Unigram model: all words in a document are drawn uniformly at random from categorical distribution
- Mixture of unigrams model: for each document, we first choose a topic z and then generate words for the document from the conditional distribution $p(w|z)$
  - Topics are just probability distributions over words

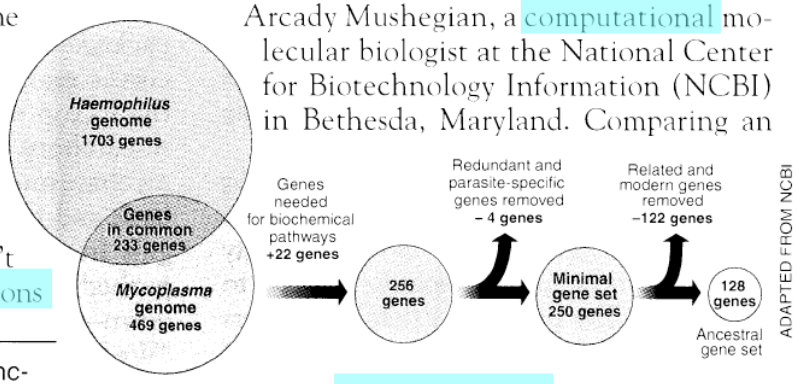
A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Topic Models

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Content

Aside: don't always sample!

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

Example: numerical solutions to (nice) 1D integrals are fast

```
octave:1> 4 * quadl(@(x) sqrt(1-x.^2), 0, 1, tolerance)
```

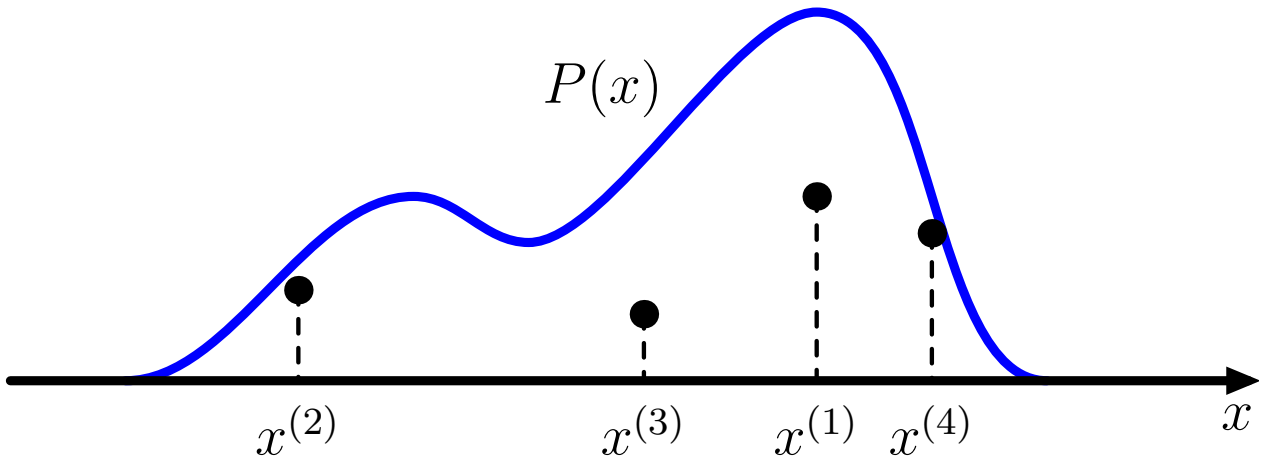
Gives π to 6 dp's in 108 evaluations, machine precision in 2598.

(NB Matlab's `quadl` fails at zero tolerance)

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Sampling from distributions

Draw points uniformly under the curve:



Probability mass to left of point $\sim$ Uniform[0,1]

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Sampling from distributions

How to convert samples from a Uniform[0,1] generator:

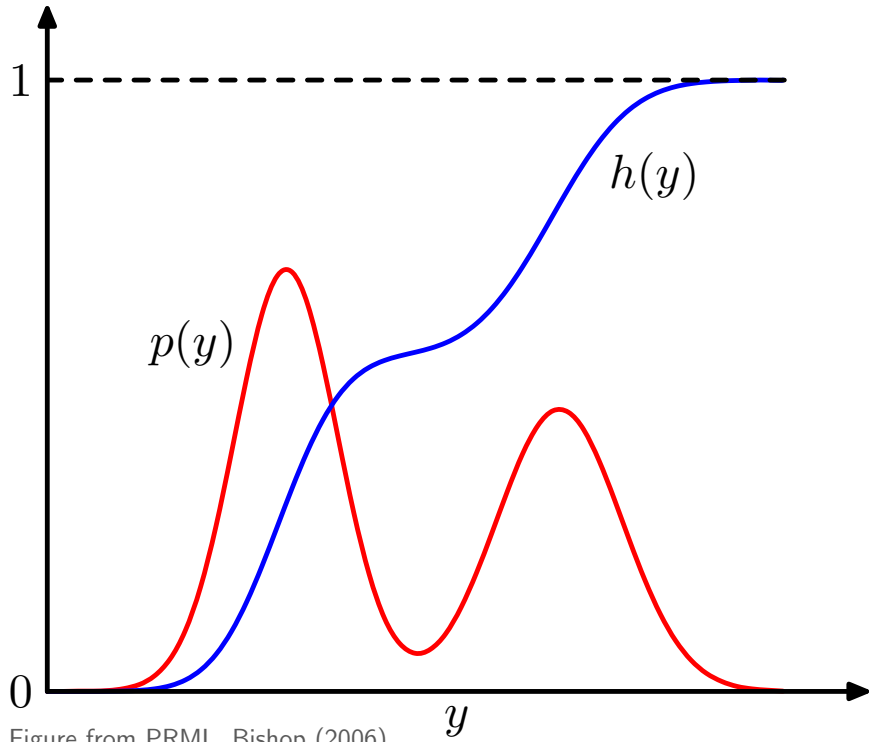


Figure from PRML, Bishop (2006)

$$h(y) = \int_{-\infty}^{y} p(y')\,\mathrm{d}y'$$

Draw mass to left of point:

$u \sim$ Uniform[0,1]

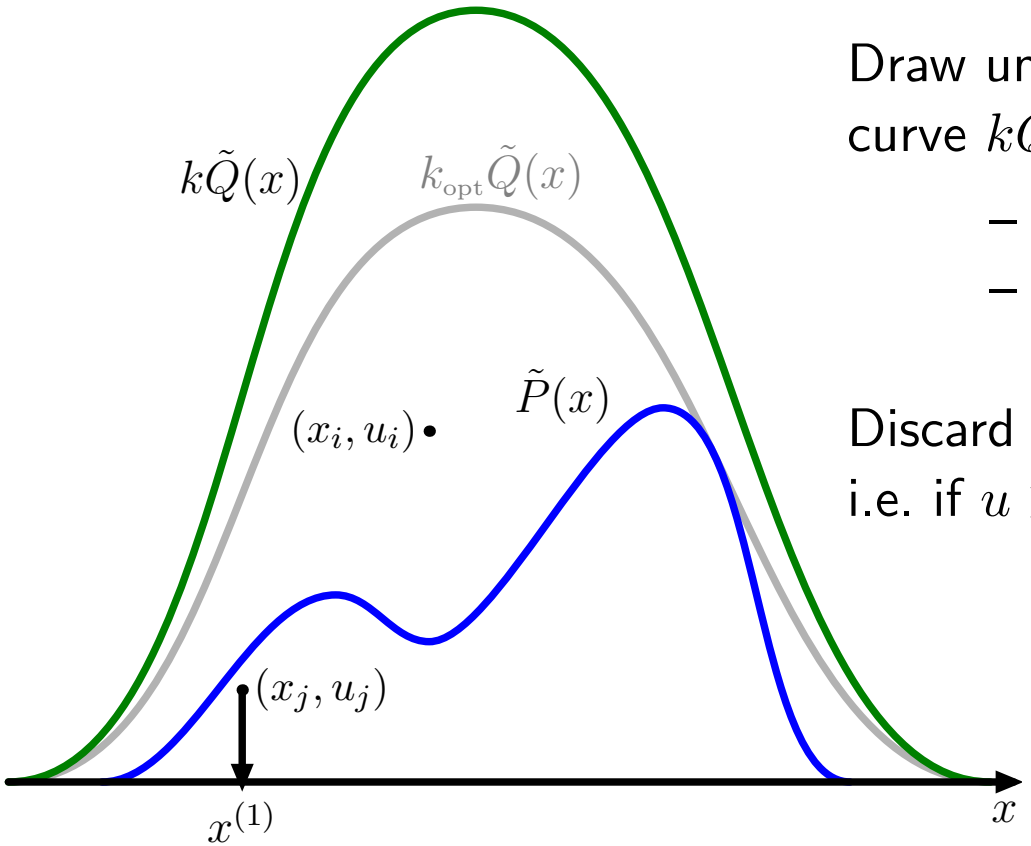
Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Rejection sampling

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

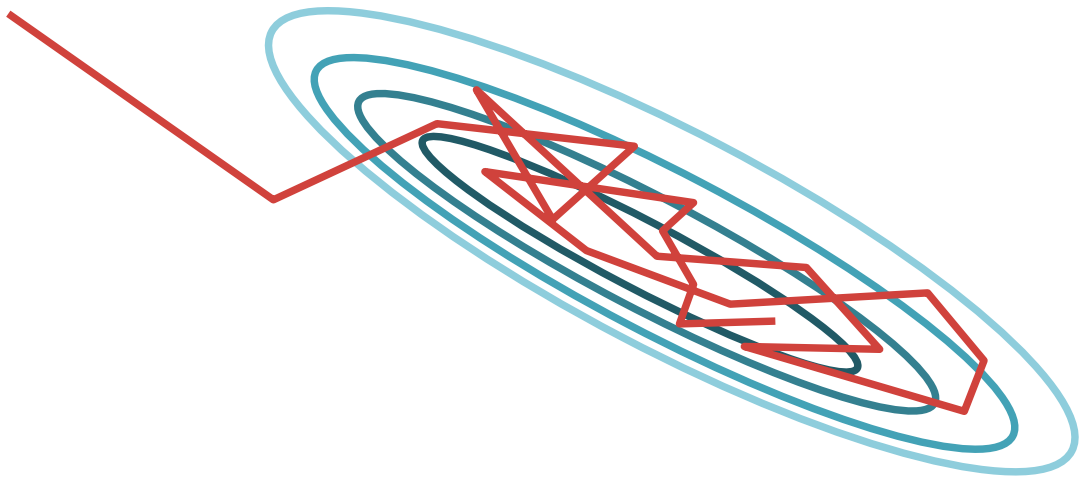
Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Content

MCMC

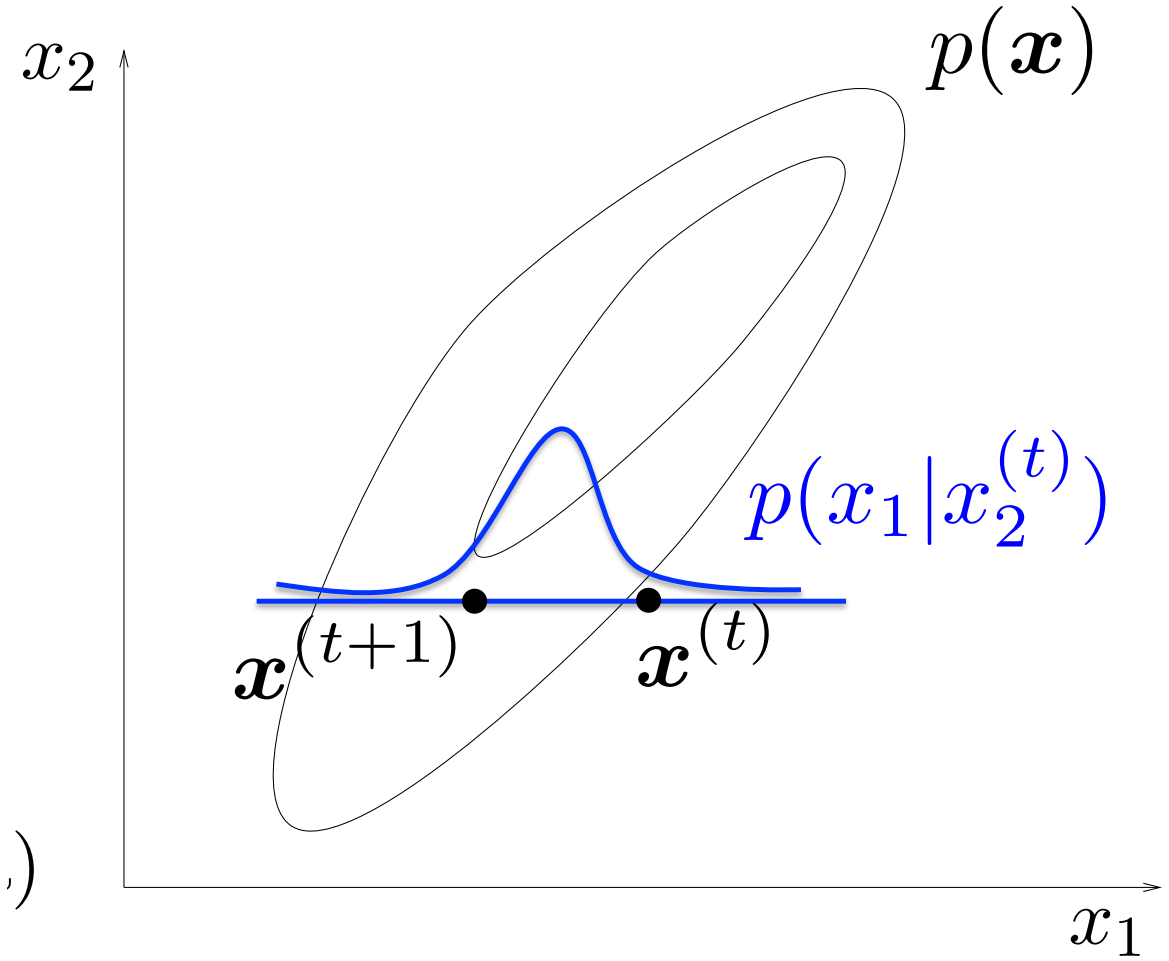
- **Goal:** Draw approximate, correlated samples from a target distribution p(x)
- **MCMC:** Performs a biased random walk to explore the distribution



11

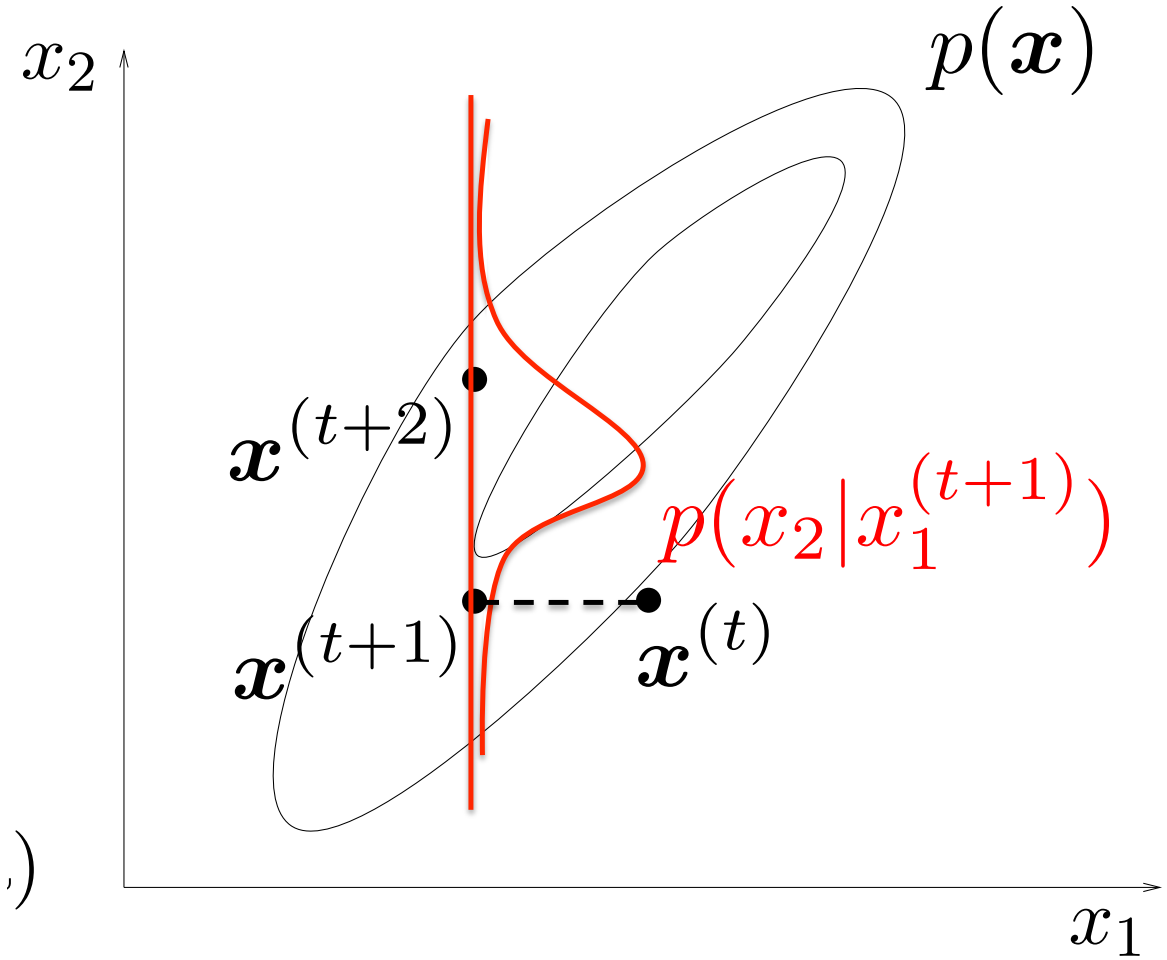
A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Gibbs Sampling



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

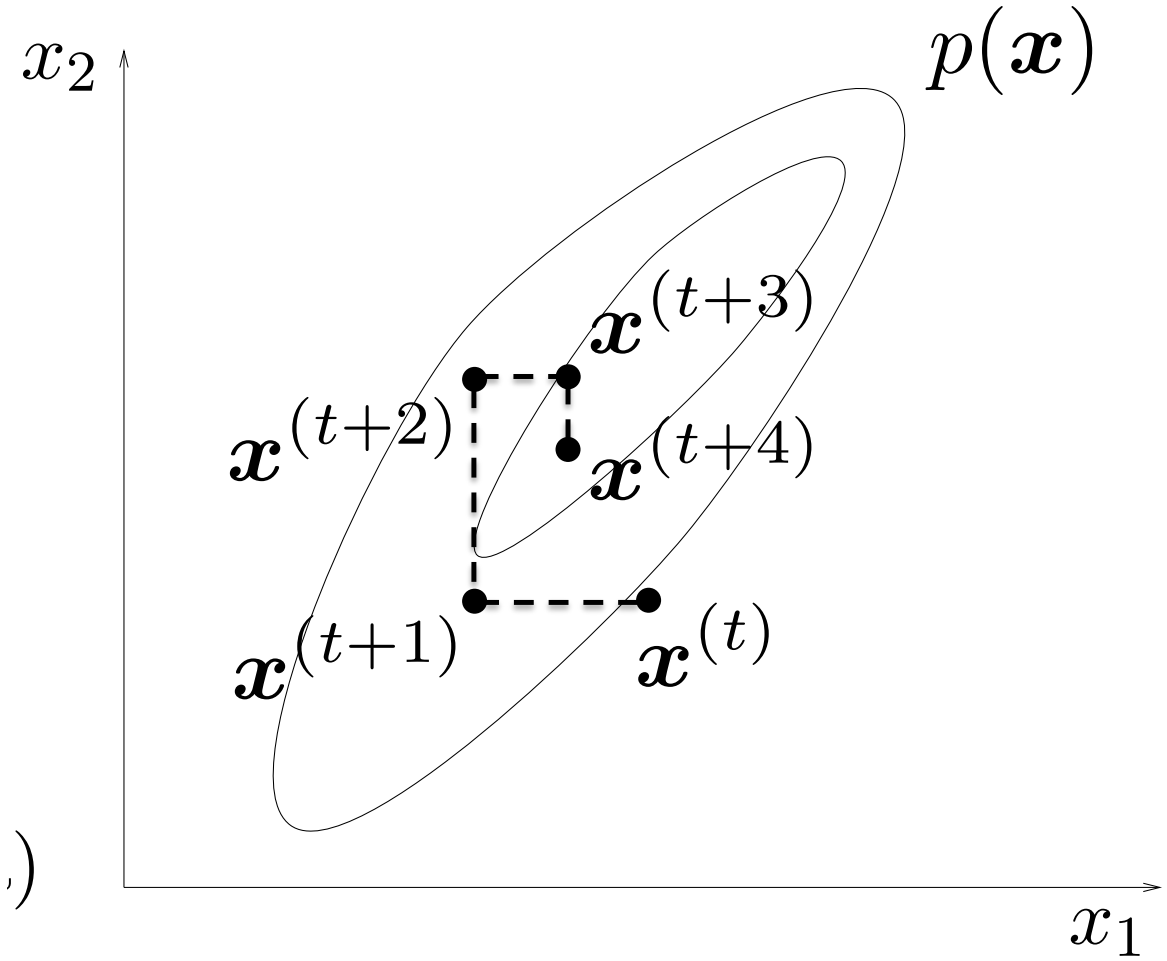
Gibbs Sampling



15

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Gibbs Sampling



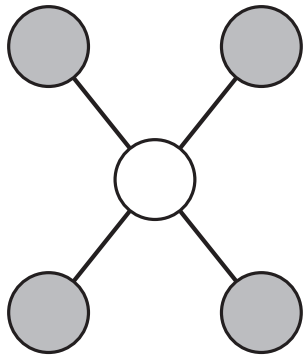
16

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

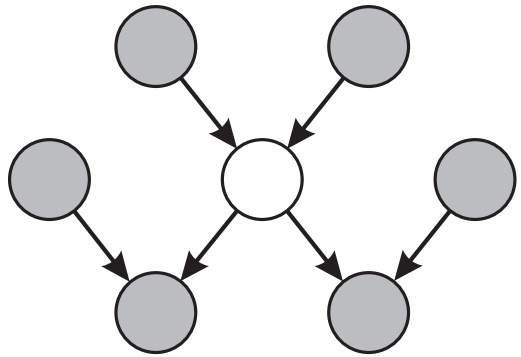
Gibbs Sampling

Full conditionals only need to condition on the Markov Blanket

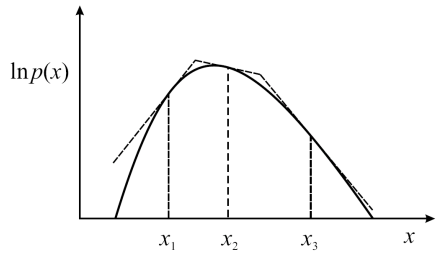
MRF



Bayes Net



- Must be "easy" to sample from conditionals
- Many conditionals are log-concave and are amenable to adaptive rejection sampling



17

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

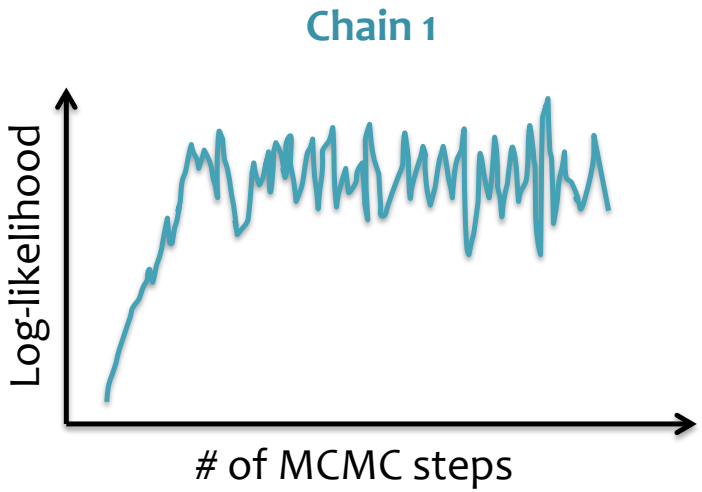
Why does Gibbs sampling work?

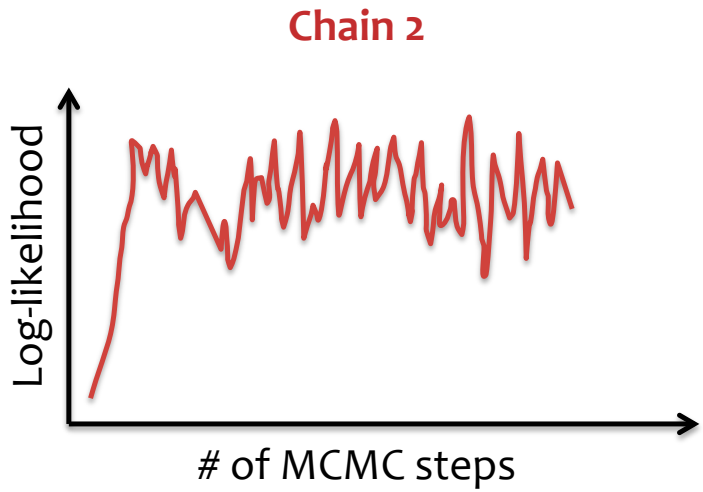
- Metropolis-Hastings
  - Markov chains
  - Stationary distribution
  - MH Algorithm
    - Constructs a Markov chain whose stationary distribution is the desired distribution
  - Proof that samples will be from desired distribution:
    - Sufficient conditions for constructing a markov chain with desired stationary distribution:
      - ergodicity
      - detailed balance (stronger, than what we need, but easier for the proof)
- Gibbs Sampling is a special case of Metropolis-Hastings
  - a special proposal distribution, which ensures the hastings ratio is always 1.0

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Practical Issues

- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods



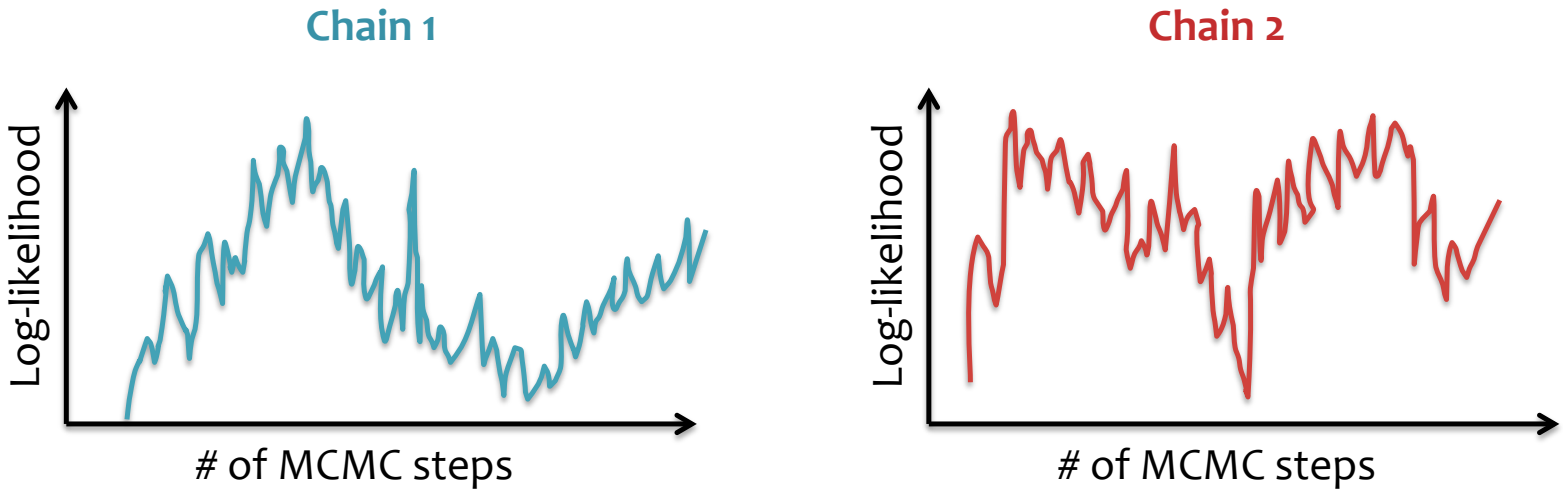


19

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Practical Issues

- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
  - Compare statistics of multiple independent chains
  - Ex: Compare log-likelihoods

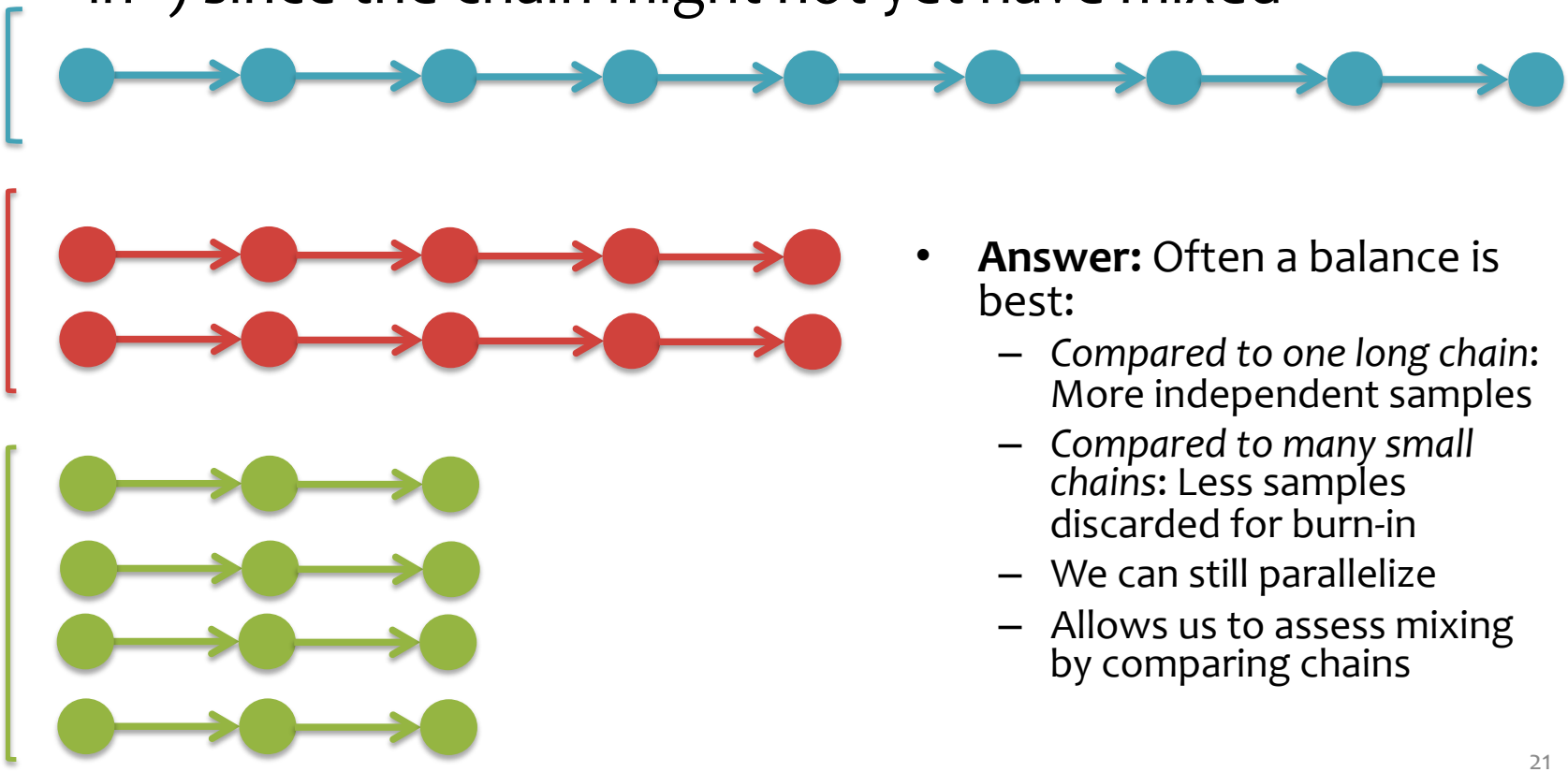


20

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Practical Issues

- **Question:** Is one long Markov chain better than many short ones?
- **Note:** typical to discard initial samples (aka. "burn-in") since the chain might not yet have mixed



- **Answer:** Often a balance is best:
  - *Compared to one long chain*: More independent samples
  - *Compared to many small chains*: Less samples discarded for burn-in
  - We can still parallelize
  - Allows us to assess mixing by comparing chains

21

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Content

Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

Topic Modeling:

A method of (usually unsupervised) discovery of latent or hidden structure in a corpus

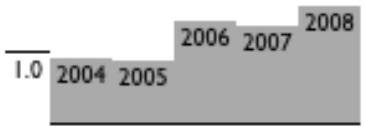
- Applied primarily to text corpora, but **techniques are more general**
- Provides a **modeling toolbox**
- Has prompted the exploration of a variety of new **inference methods** to accommodate **large-scale datasets**

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Topic Modeling

Dirichlet-multinomial regression (DMR) topic model on ICML
(Mimno & McCallum, 2008)

Topic 0 [0.152]

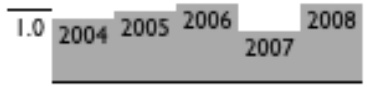


problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

Topic 54 [0.051]

decision trees, trees, tree, decision tree, decision, tree ensemble, junction tree, decision tree learners, leaf nodes, arithmetic circuits, ensembles modts, skewing, ensembles, anytime induction decision trees, trees trees, random forests, objective decision trees, tree learners, trees grove, candidate split

Topic 99 [0.066]



inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

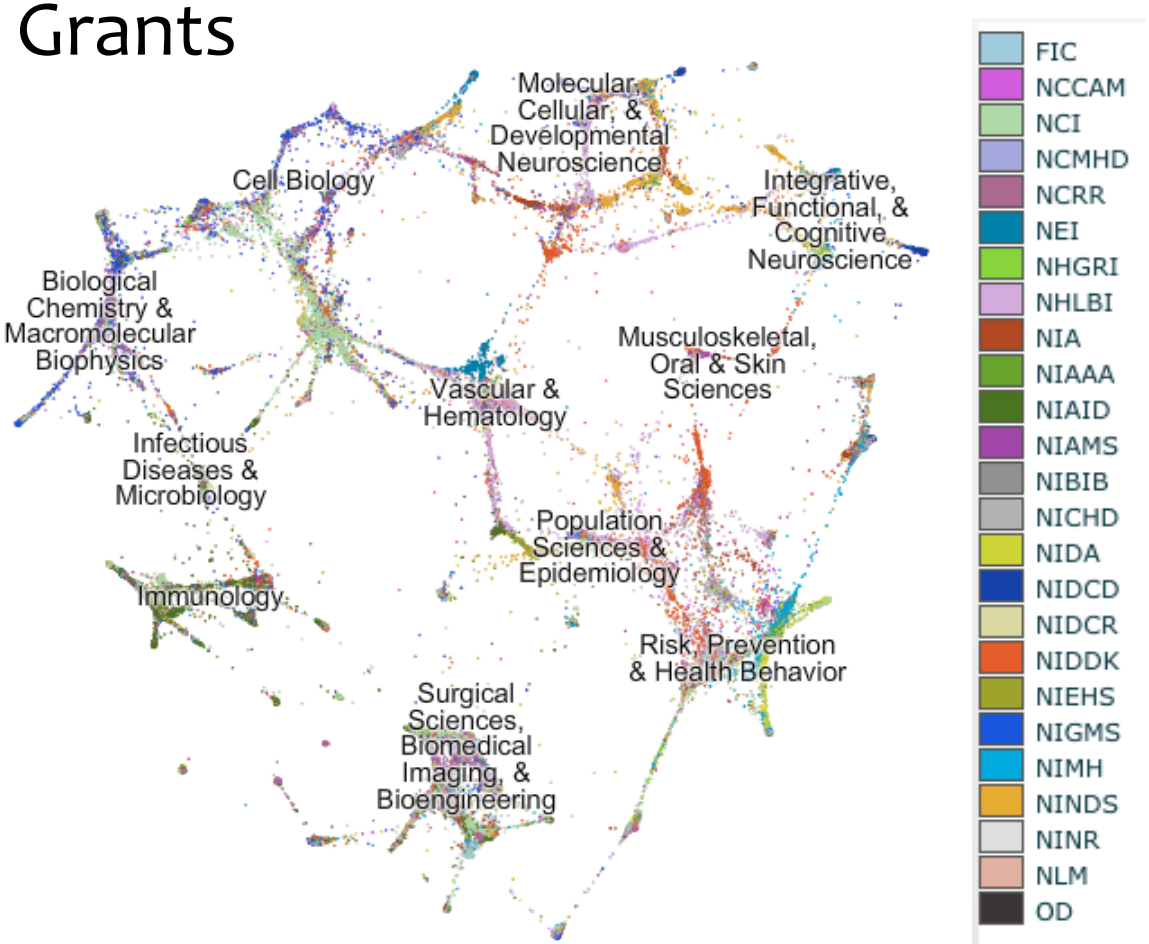
http:// www.cs.umass.edu/~mimno/icml100.html

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Topic Modeling

- Map of NIH Grants

(Talley et al., 2011)



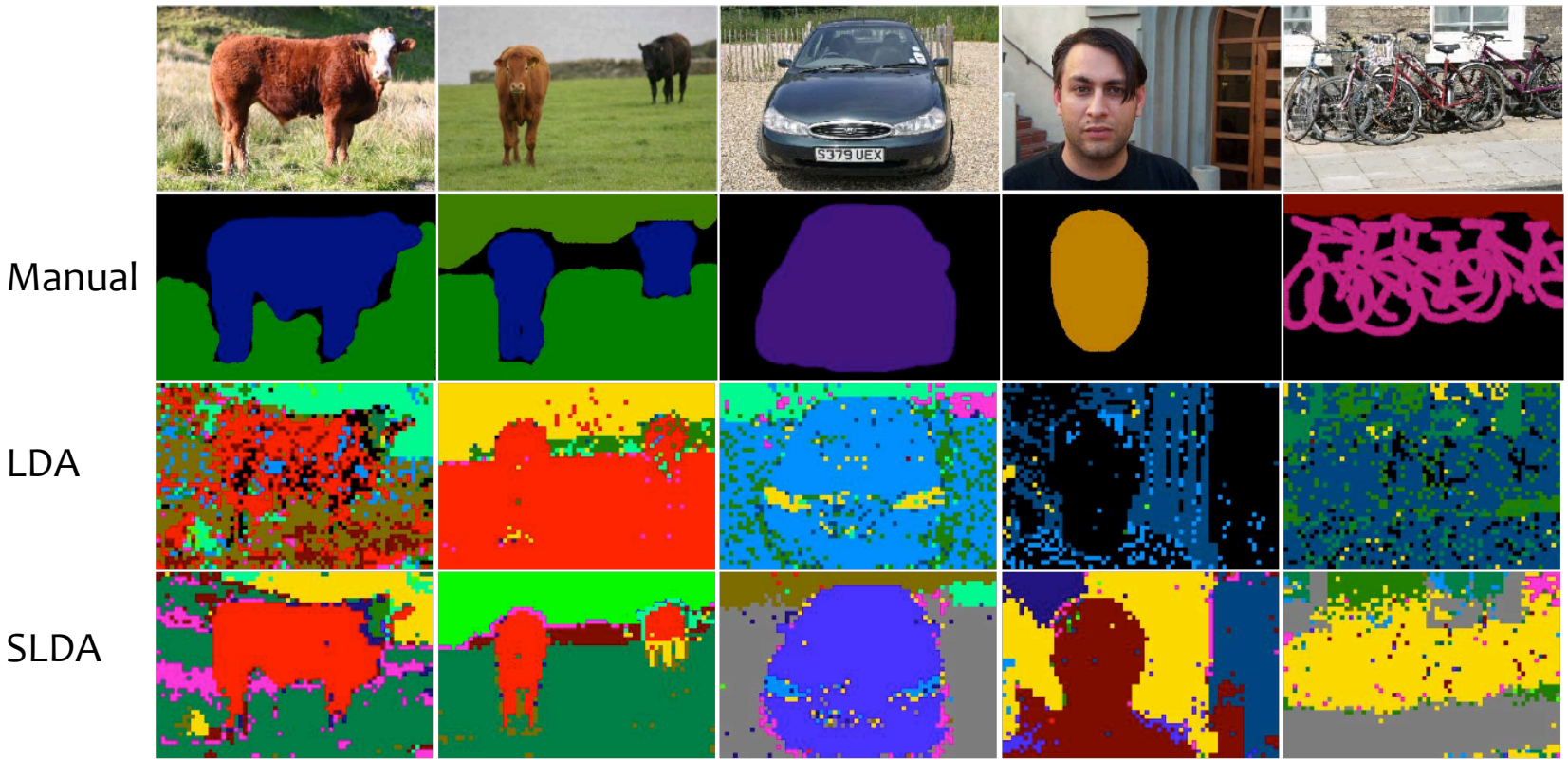
https://app.nihmaps.org/

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Other Applications of Topic Models

- Spacial LDA

(Wang & Grimson, 2007)



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Other Applications of Topic Models

- Word Sense Induction

(Brody & Lapata, 2009)

Senses of *drug* (WSJ)
1. U.S., administration, federal, against, war, dealer
2. patient, people, problem, doctor, company, abuse
3. company, million, sale, maker, stock, inc.
4. administration, food, company, approval, FDA

Senses of *drug* (BNC)
1. patient, treatment, effect, anti-inflammatory
2. alcohol, treatment, patient, therapy, addiction
3. patient, new, find, effect, choice, study
4. test, alcohol, patient, abuse, people, crime
5. trafficking, trafficker, charge, use, problem
6. abuse, against, problem, treatment, alcohol
7. people, wonder, find, prescription, drink, addict
8. company, dealer, police, enforcement, patient

- Selectional Preference

(Ritter et al., 2010)

Topic t	Arg1	Relations which assign highest probability to t	Arg2
18	The residue - The mixture - The reaction mixture - The solution - the mixture - the reaction mixture - the residue - The reaction - the solution - The filtrate - the reaction - The product - The crude product - The pellet - The organic layer - Thereto - This solution - The resulting solution - Next - The organic phase - The resulting mixture - C.)	was treated with, is treated with, was poured into, was extracted with, was purified by, was diluted with, was filtered through, is disolved in, is washed with	EtOAc - CH2Cl2 - H2O - CH.sub.2Cl.sub.2 - H.sub.2O - water - MeOH - NaHCO3 - Et2O - NHCl - CHCl.sub.3 - NHCl - dropwise - CH2Cl.sub.2 - Celite - Et.sub.2O - Cl.sub.2 - NaOH - AcOEt - CH2C12 - the mixture - saturated NaHCO3 - SiO2 - H2O - N hydrochloric acid - NHCl - preparative HPLC - to0 C

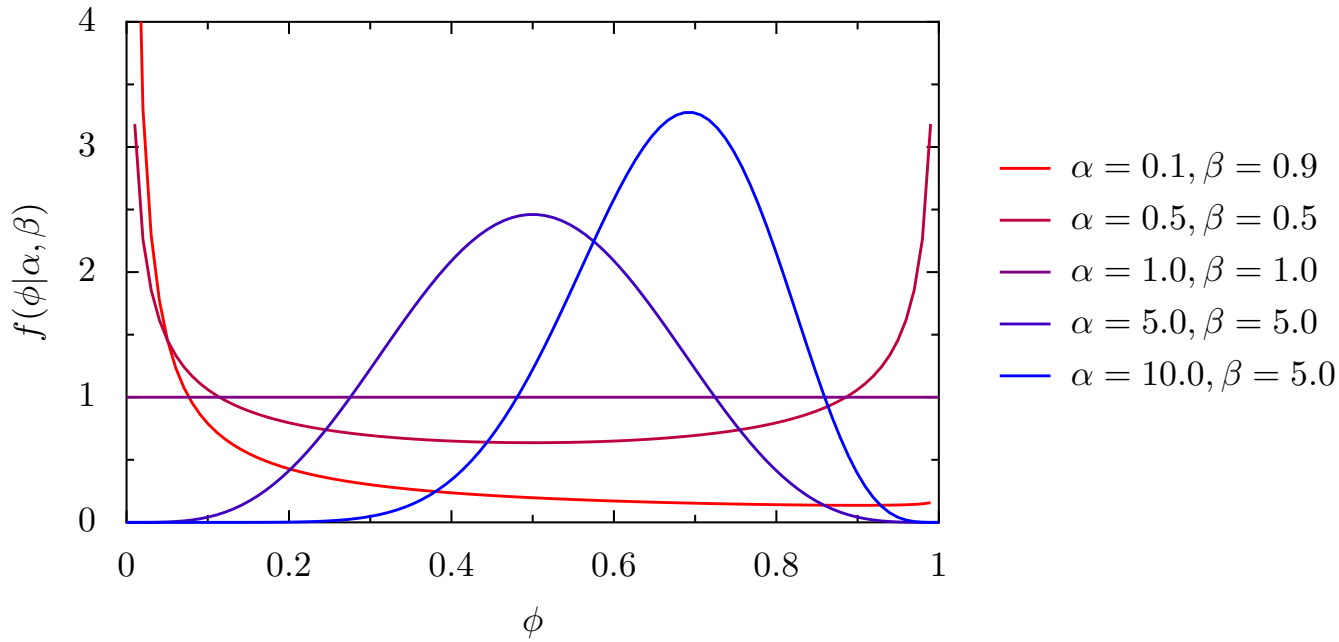
A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Content

Beta-Bernoulli Model

- Beta Distribution

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



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Beta-Bernoulli Model

- 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)

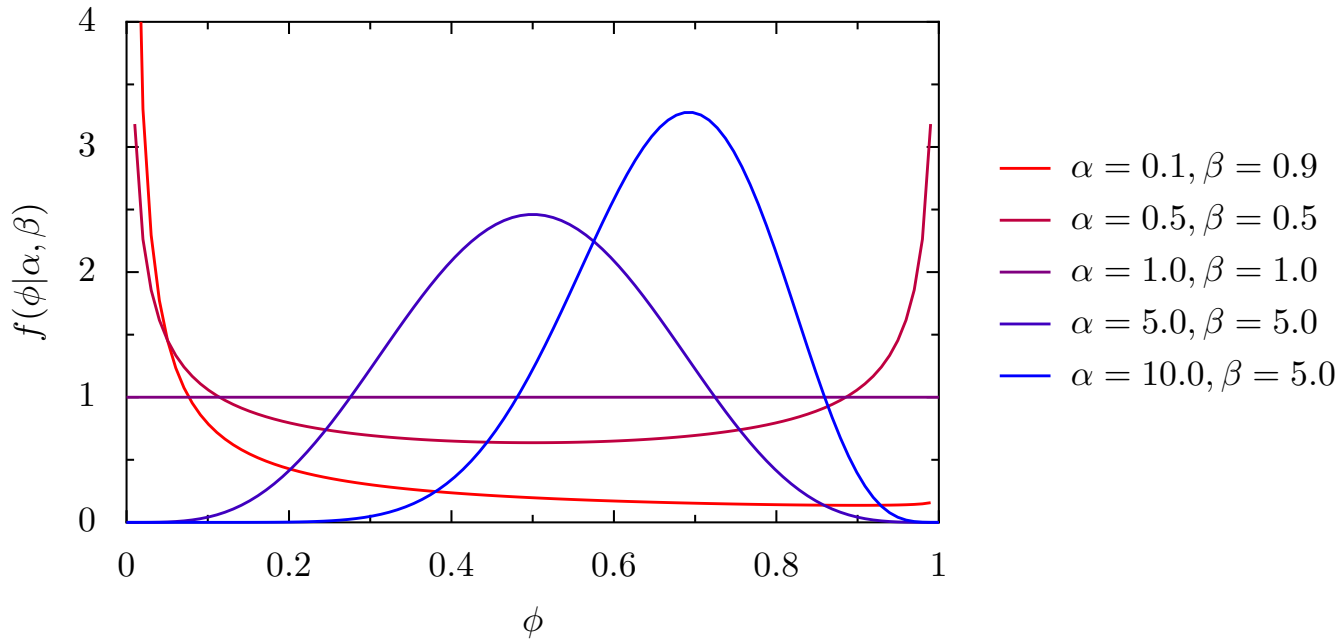
H	T	T	H	H	T	T	H	H	H
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Model

- Dirichlet Distribution

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

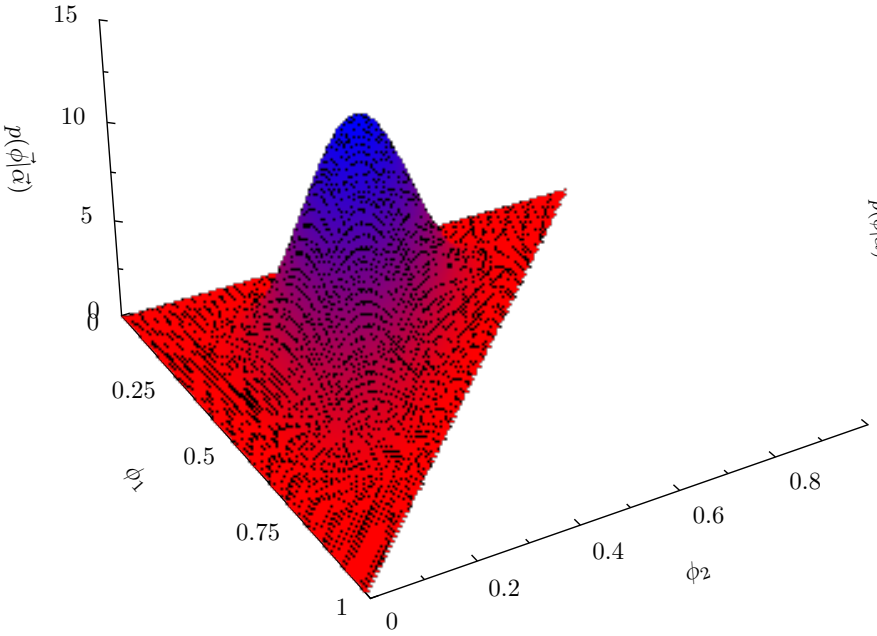


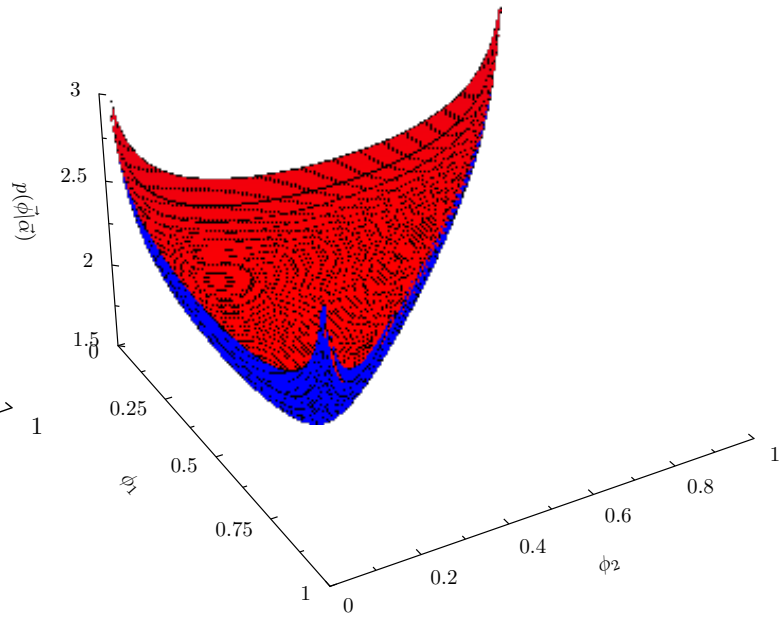
A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Model

- Dirichlet Distribution

$$p(\vec{\phi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$





A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Model

- Generative Process

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

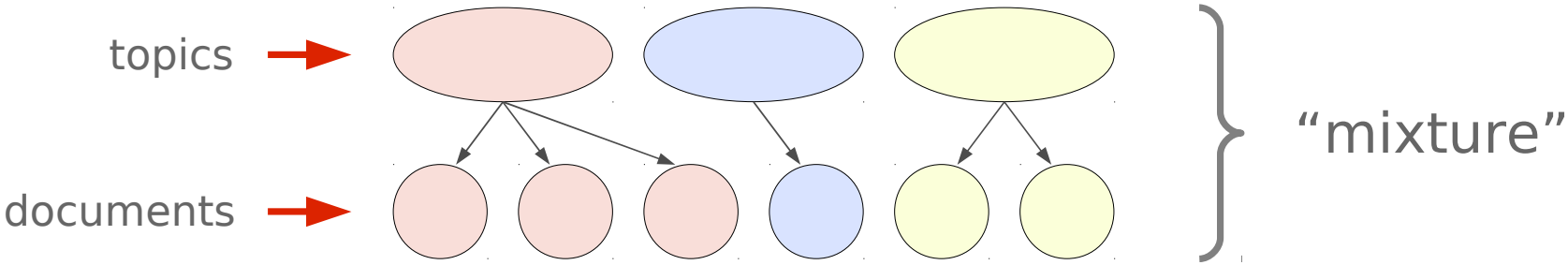
- Example corpus

the	he	is	the	and	the	she	she	is	is
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Mixture Model

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Mixture Model

- Generative Process

For each topic $k \in \{1, \ldots, K\}$:
  $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})$ [*draw distribution over words*]
$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ [*draw distribution over topics*]
For each document $m \in \{1, \ldots, M\}$
  $z_m \sim \text{Mult}(1, \boldsymbol{\theta})$ [*draw topic assignment*]
  For each word $n \in \{1, \ldots, N_m\}$
    $x_{mn} \sim \text{Mult}(1, \boldsymbol{\phi}_{z_m})$ [*draw word*]

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

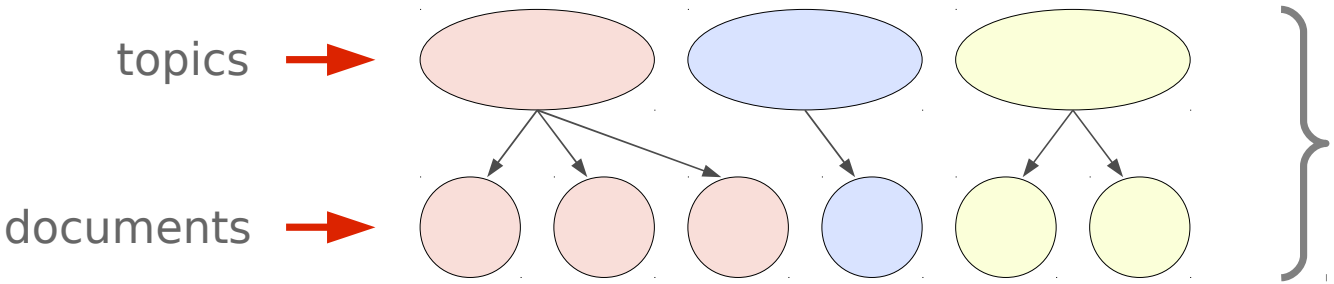
Document 2

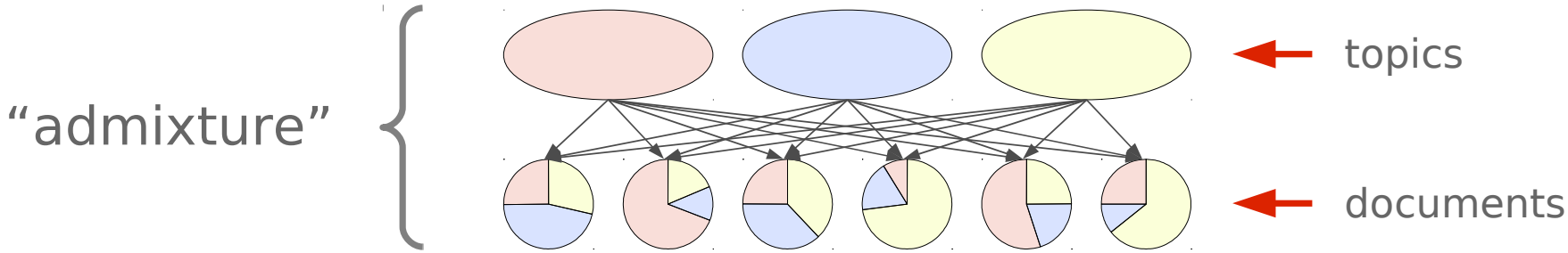
she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Mixture vs. Admixture (LDA)



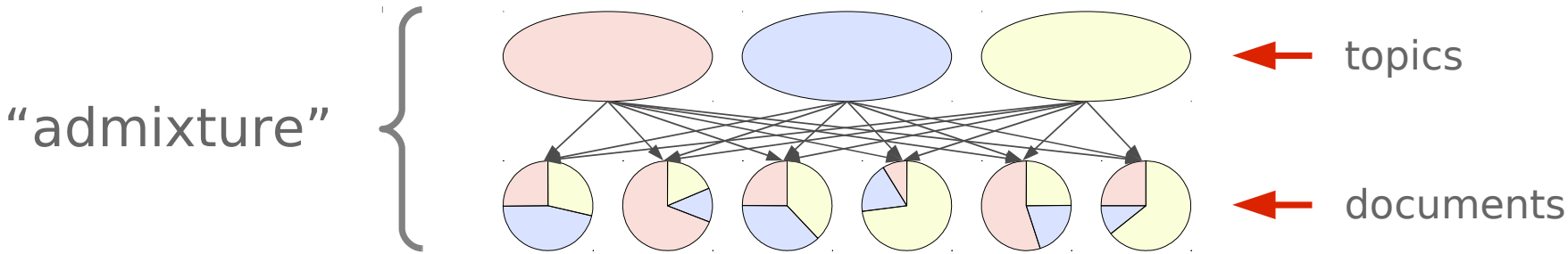


Diagrams from Wallach, JHU 2011, slides

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Latent Dirichlet Allocation

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Latent Dirichlet Allocation

- Generative Process

For each topic $k \in \{1, \ldots, K\}$:
  $\phi_k \sim \text{Dir}(\beta)$ [*draw distribution over words*]
For each document $m \in \{1, \ldots, M\}$
  $\theta_m \sim \text{Dir}(\alpha)$ [*draw distribution over topics*]
  For each word $n \in \{1, \ldots, N_m\}$
    $z_{mn} \sim \text{Mult}(1, \theta_m)$ [*draw topic assignment*]
    $x_{mn} \sim \phi_{z_{mi}}$ [*draw word*]

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

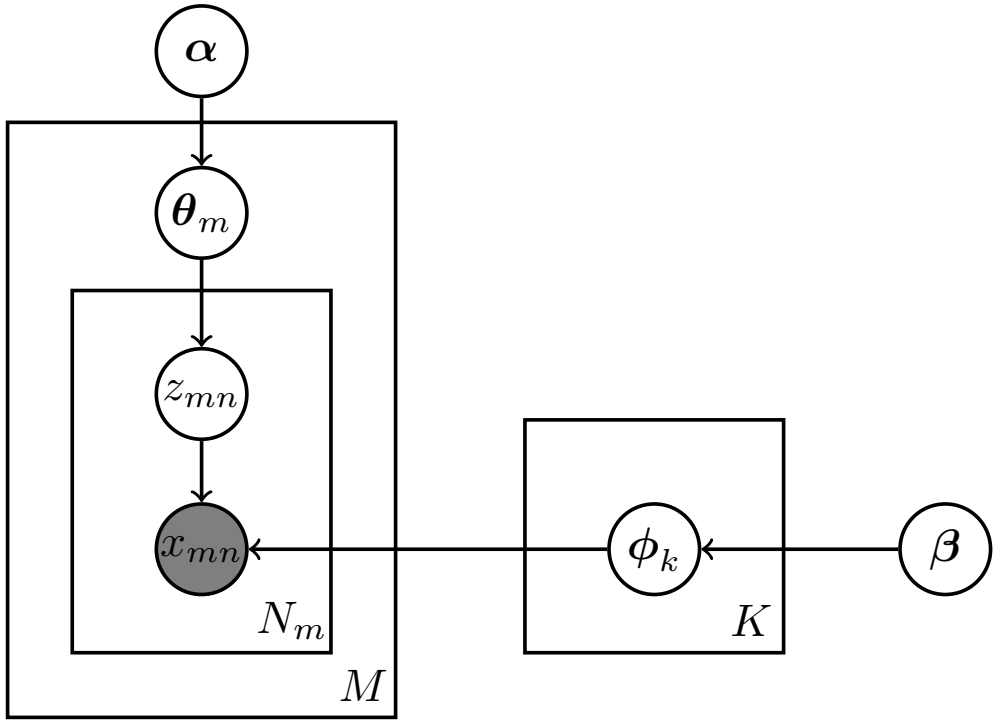
she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Latent Dirichlet Allocation

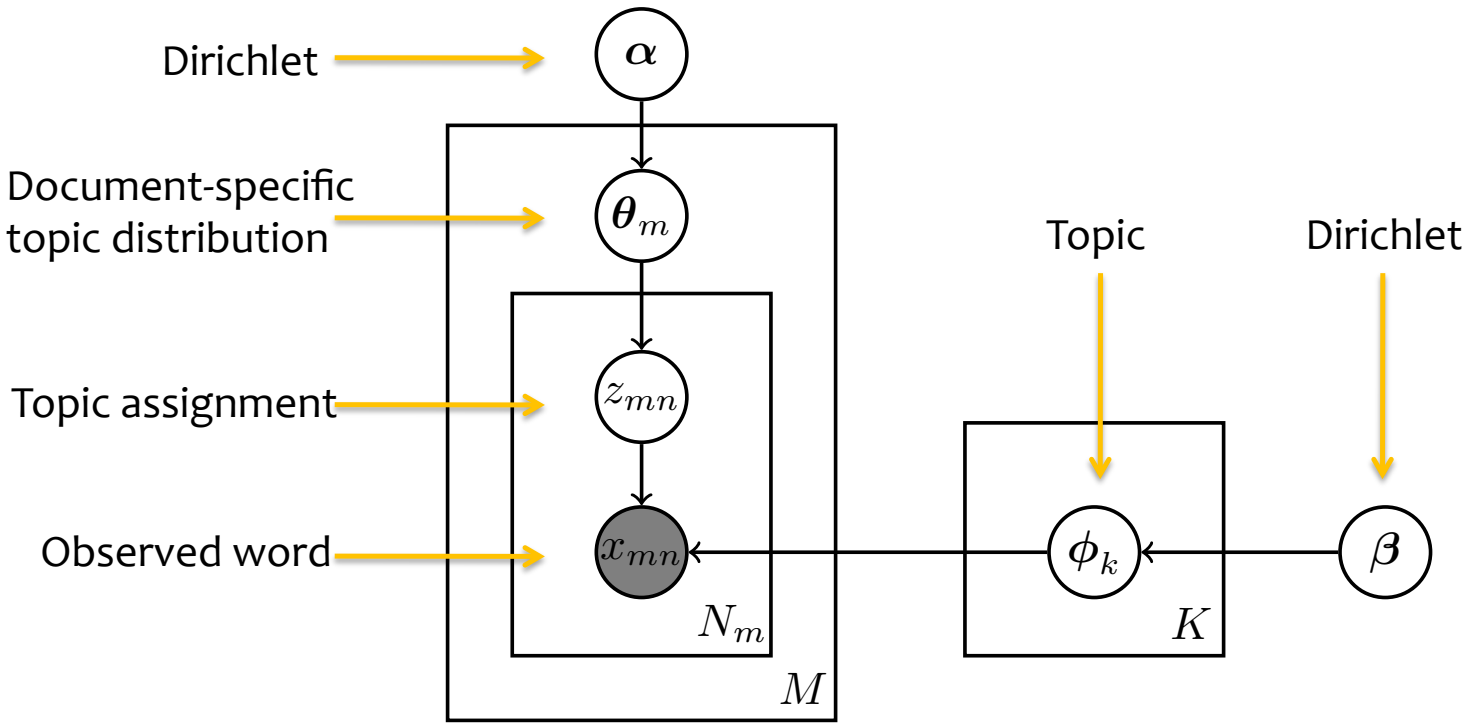
- Plate Diagram



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

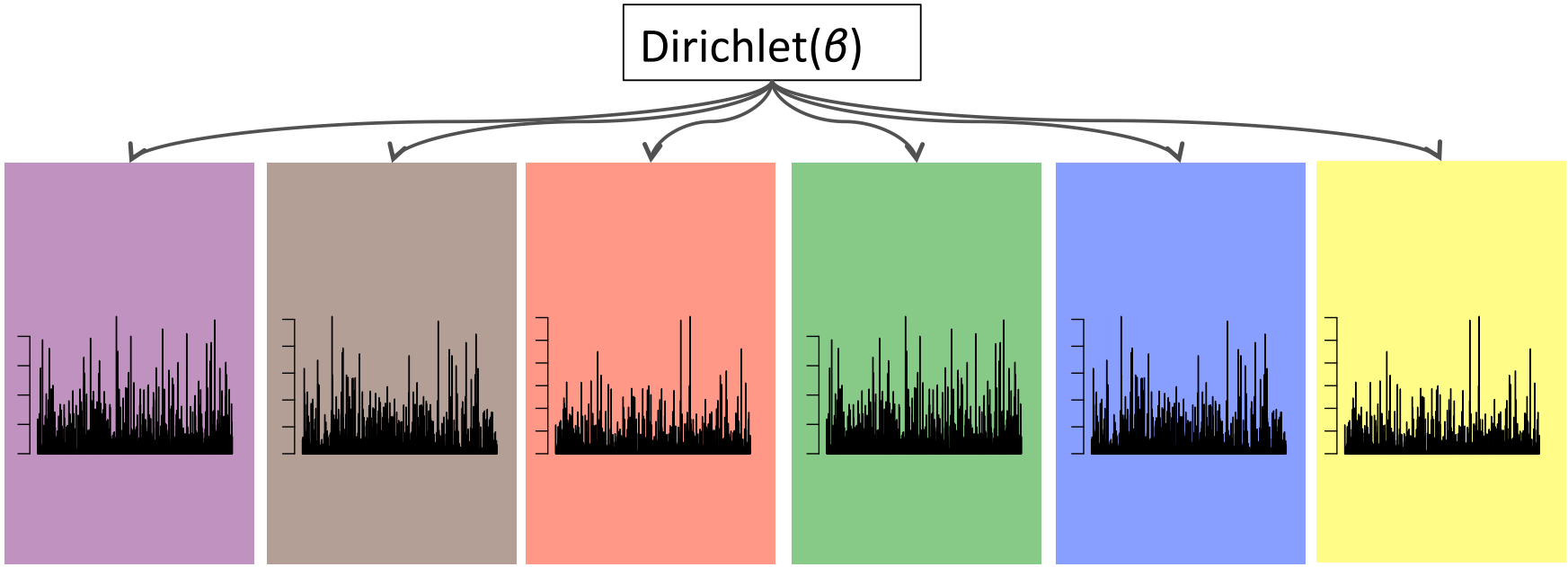
Latent Dirichlet Allocation

- Plate Diagram



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

LDA for Topic Modeling

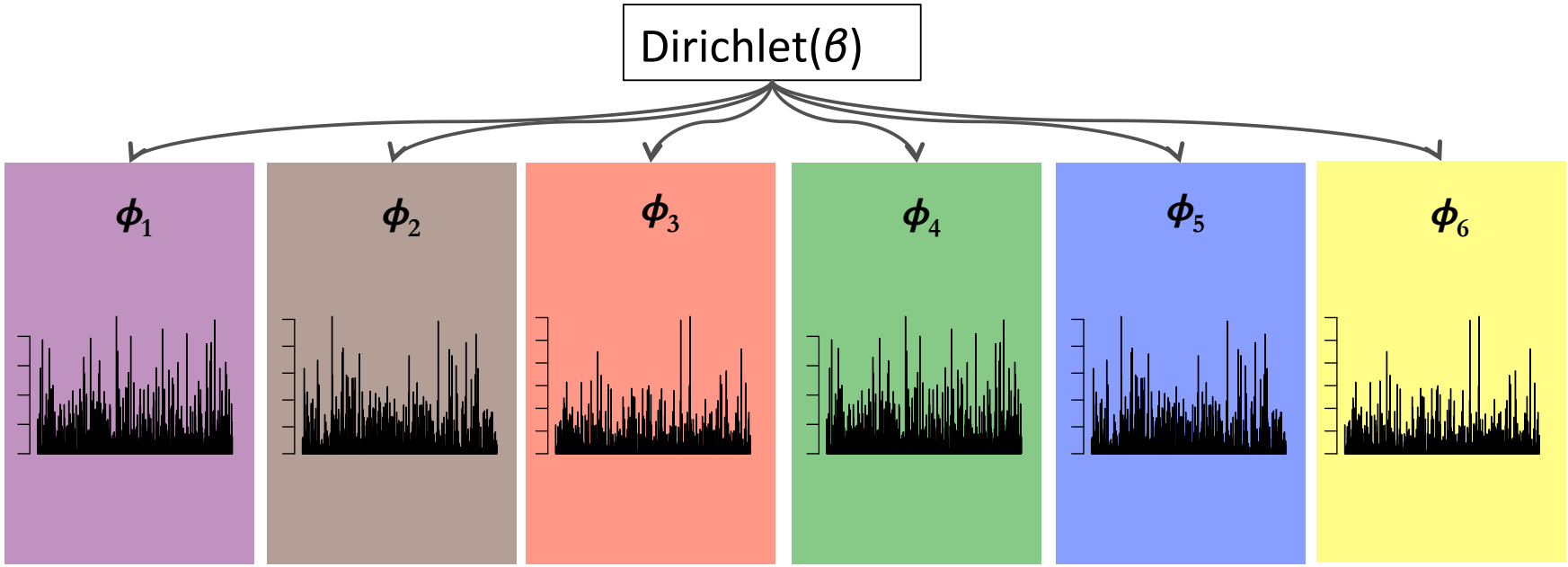


- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_k$

42

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

LDA for Topic Modeling

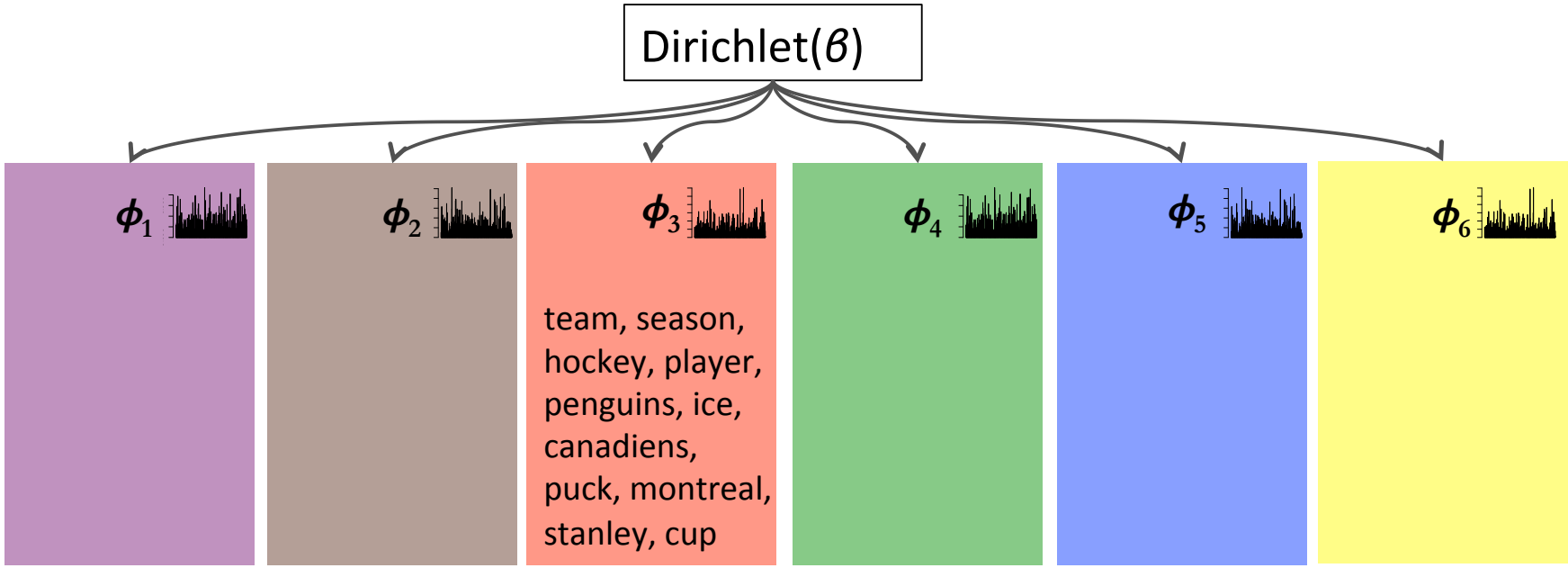


- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_k$

43

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

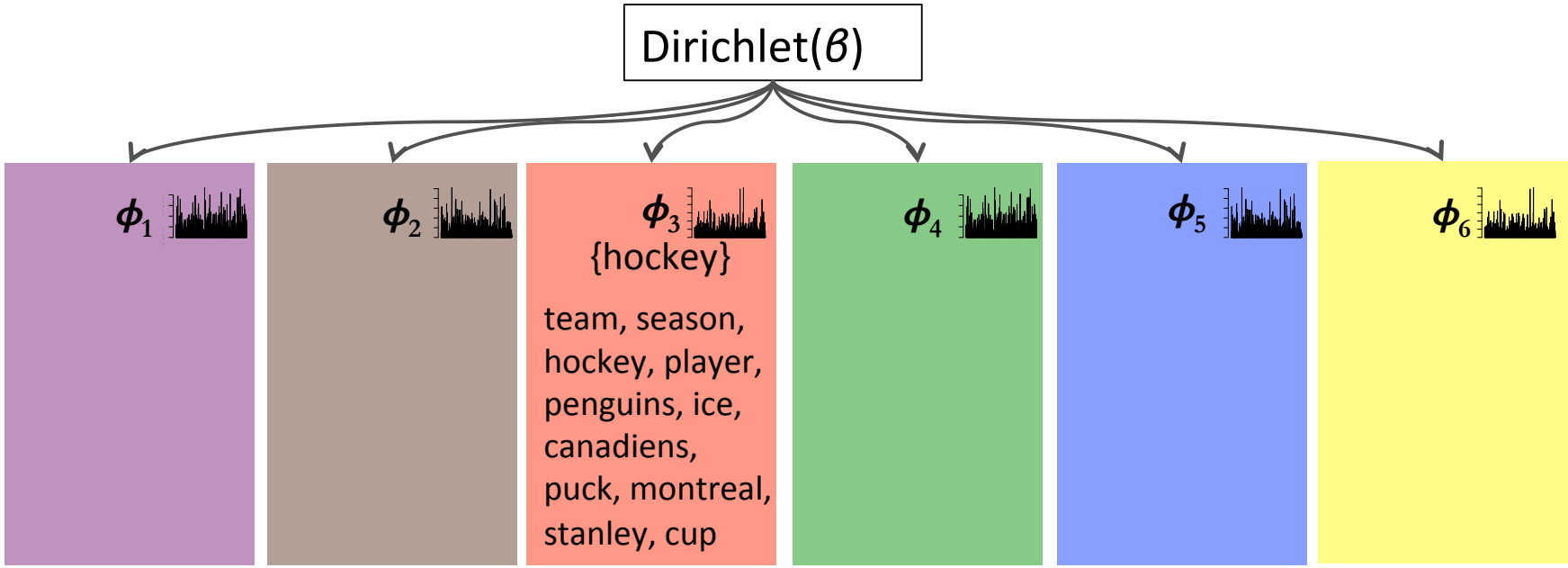
LDA for Topic Modeling



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

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

LDA for Topic Modeling



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

45

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling