Machine Learning Department
School of Computer Science
Carnegie Mellon University

Markov Chains
+
Bayesian Inference for Parameter Estimation

Matt Gormley
Lecture 14
Mar. 22, 2021

Reminders

- Project Team Formation
  - Due: Mon, Mar. 22 at 11:59pm
- Homework 3: Structured SVM
  - Out: Wed, Mar. 10
  - Due: Wed, Mar. 24 at 11:59pm

Definitions and Theoretical Justification for MCMC

MARKOV CHAINS

Markov Chains

- a **Markov chain** is a random process
  $\hookrightarrow$ gives a series of random variables

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}$$

- **first order Markov chain:**

$$p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \ldots, \mathbf{x}^{(1)}) = p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

we're focused on first order only

- **second order Markov chain:**

$$p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \ldots, \mathbf{x}^{(1)}) = p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)})$$

- **transition probabilities:**

$$R_t(\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)}) \triangleq p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

- **homogeneous Markov chain:** $R_t \triangleq R$, i.e. the transition probabilities are the same for all t

Markov Chains

Whiteboard

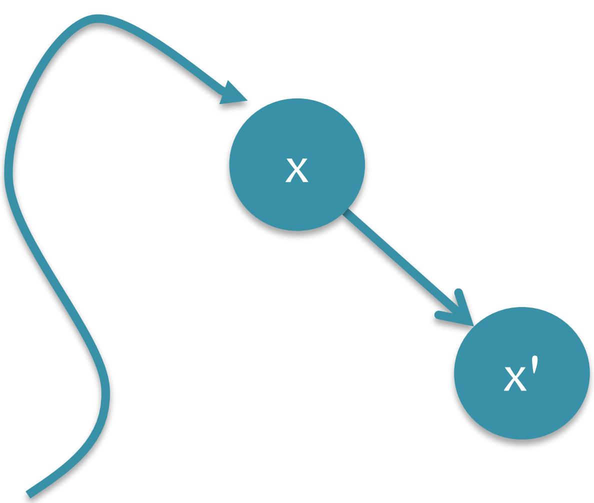
- Invariant distribution
- Equilibrium distribution
- Sufficient conditions for MCMC
- Markov chain as a WFSM

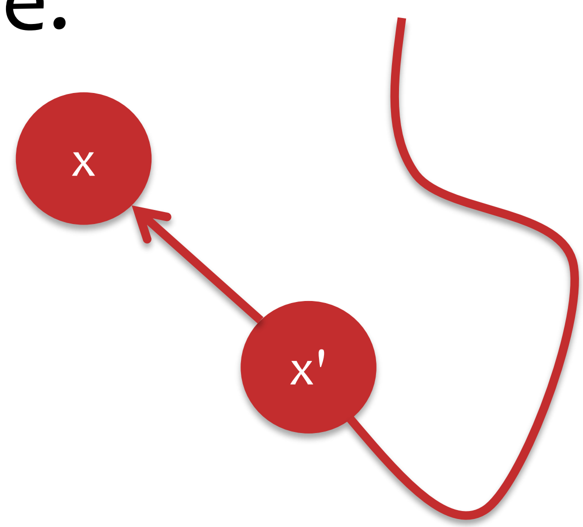
Detailed Balance

$$S(x' \leftarrow x)p(x) = S(x \leftarrow x')p(x')$$

Detailed balance means that, for each pair of states x and x',

arriving at x then x' and arriving at x' then x are equiprobable.





MCMC Summary

- **Pros**
  - Very general purpose
  - Often easy to implement
  - Good theoretical guarantees as $t \to \infty$
- **Cons**
  - Lots of tunable parameters / design choices
  - Can be quite slow to converge
  - Difficult to tell whether it's working

TOPIC MODELING

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

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:

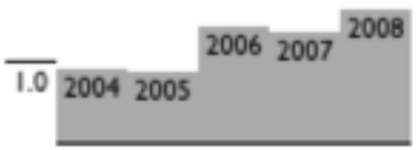
A method of (usually unsupervised) discovery of latent or hidden structure in a corpus

- 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**

Topic Modeling

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

Topic 0 [0.152]

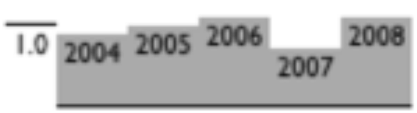


problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

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]



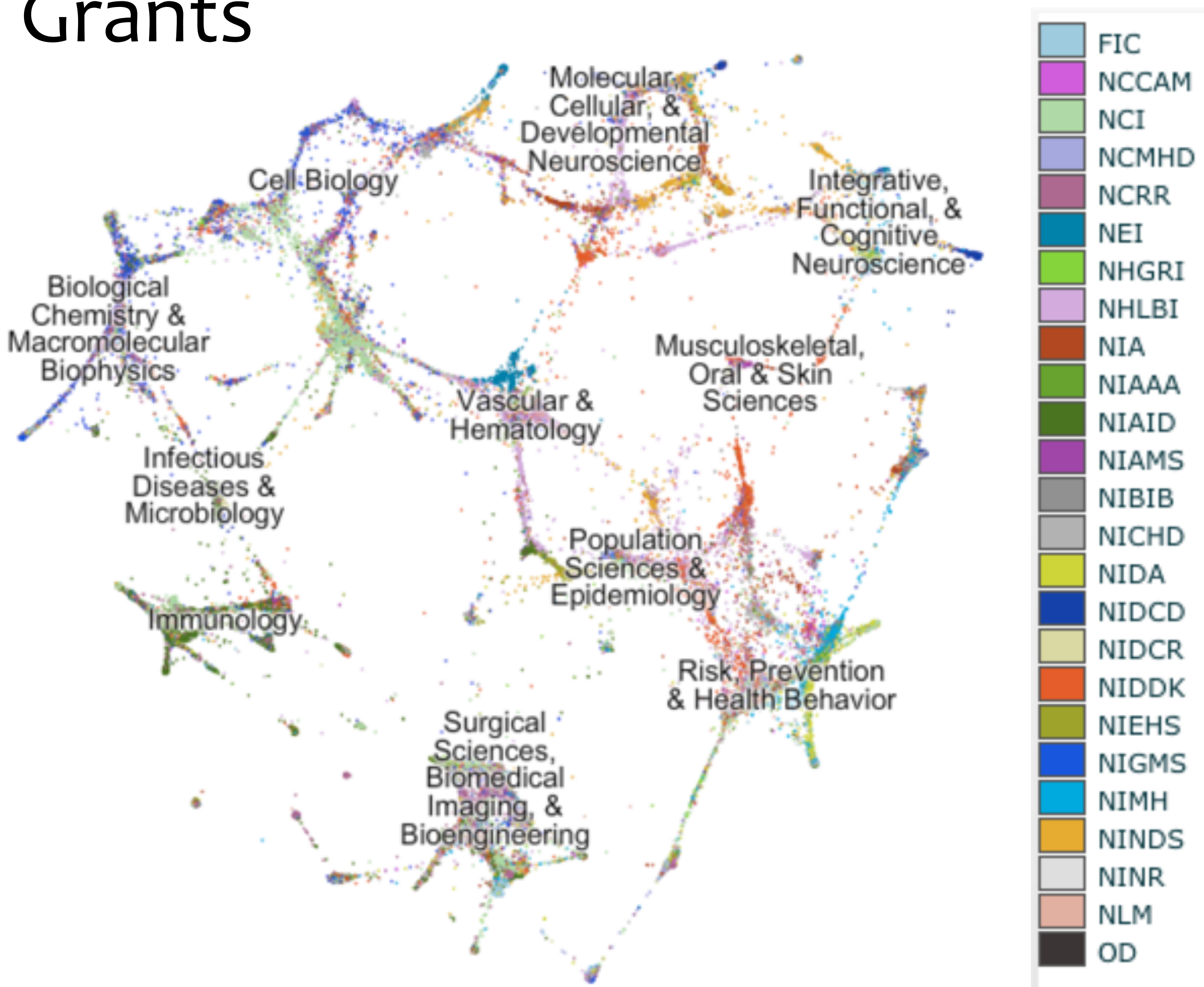
inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

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

Topic Modeling

- Map of NIH Grants

(Talley et al., 2011)

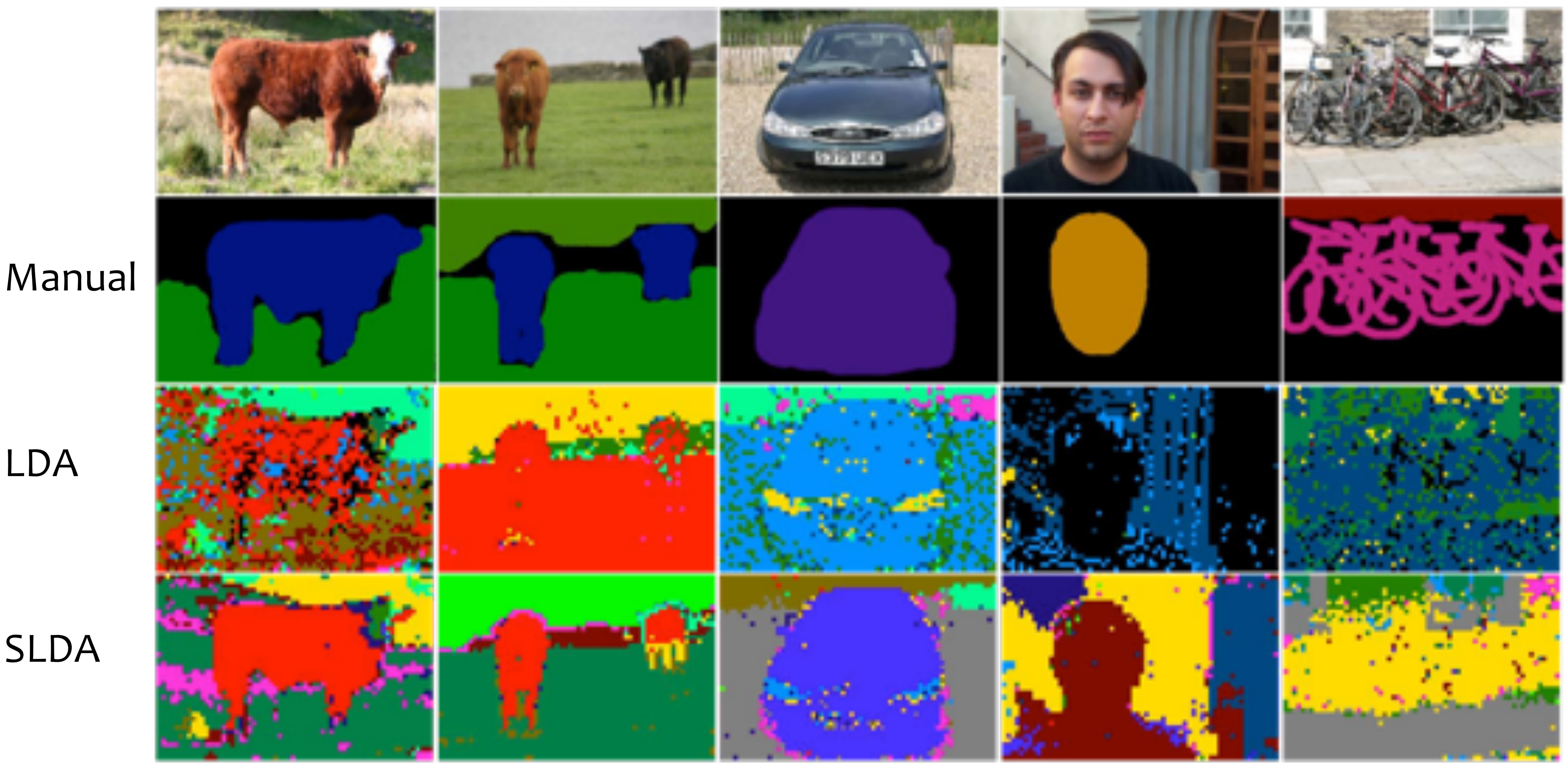


https://app.nihmaps.org/

Other Applications of Topic Models

- Spacial LDA

(Wang & Grimson, 2007)



Outline

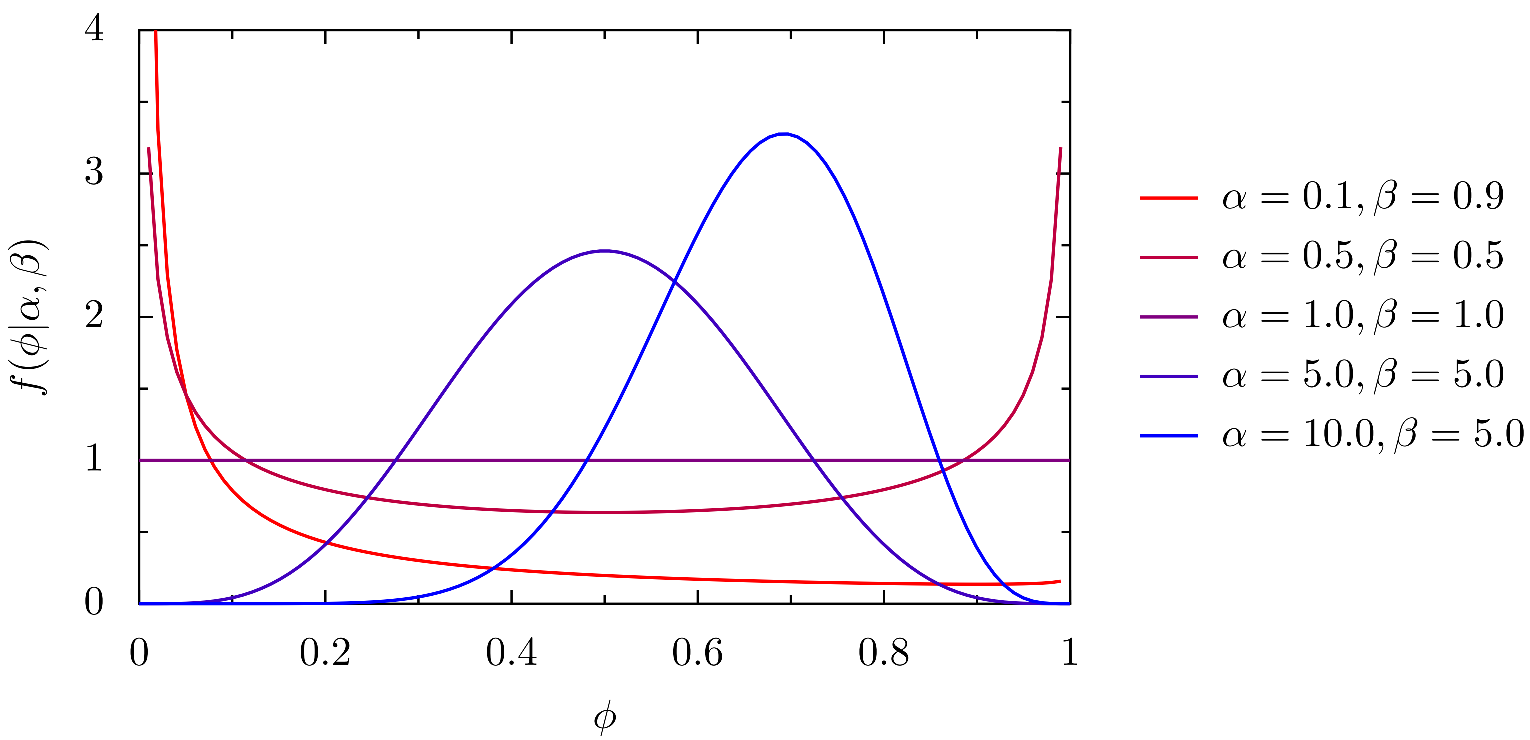
- **Applications of Topic Modeling**
- **Latent Dirichlet Allocation (LDA)**
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
- **Bayesian Inference for Parameter Estimation**
  - Exact inference
  - EM
  - Monte Carlo EM
  - Gibbs sampler
  - Collapsed Gibbs sampler
- **Extensions of LDA**
  - Correlated topic models
  - Dynamic topic models
  - Polylingual topic models
  - Supervised LDA

BAYESIAN INFERENCE FOR NAÏVE BAYES

Beta-Bernoulli Model

- Beta Distribution

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



Beta-Bernoulli Model

- Generative Process

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

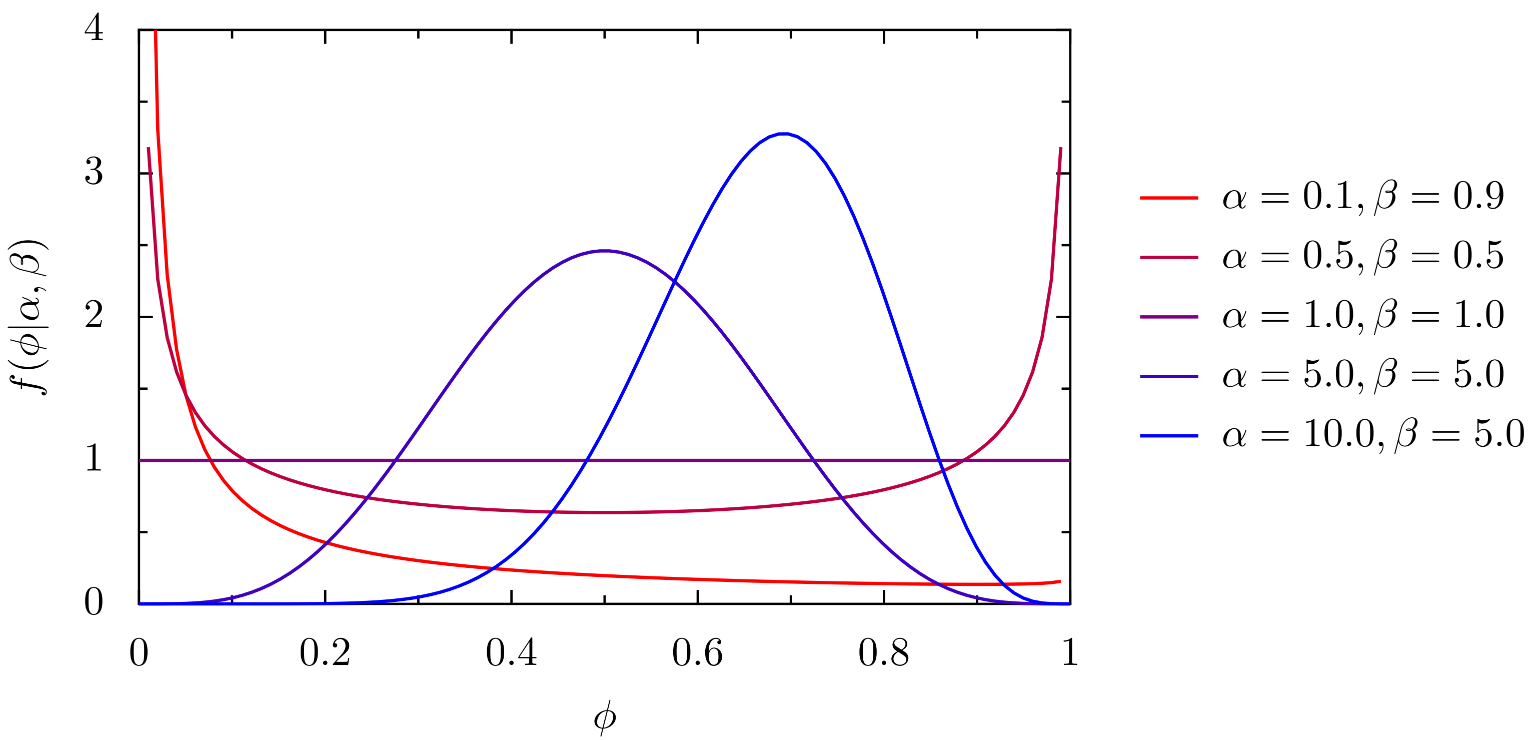
- Example corpus (heads/tails)

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}

Dirichlet-Multinomial Model

- Dirichlet Distribution

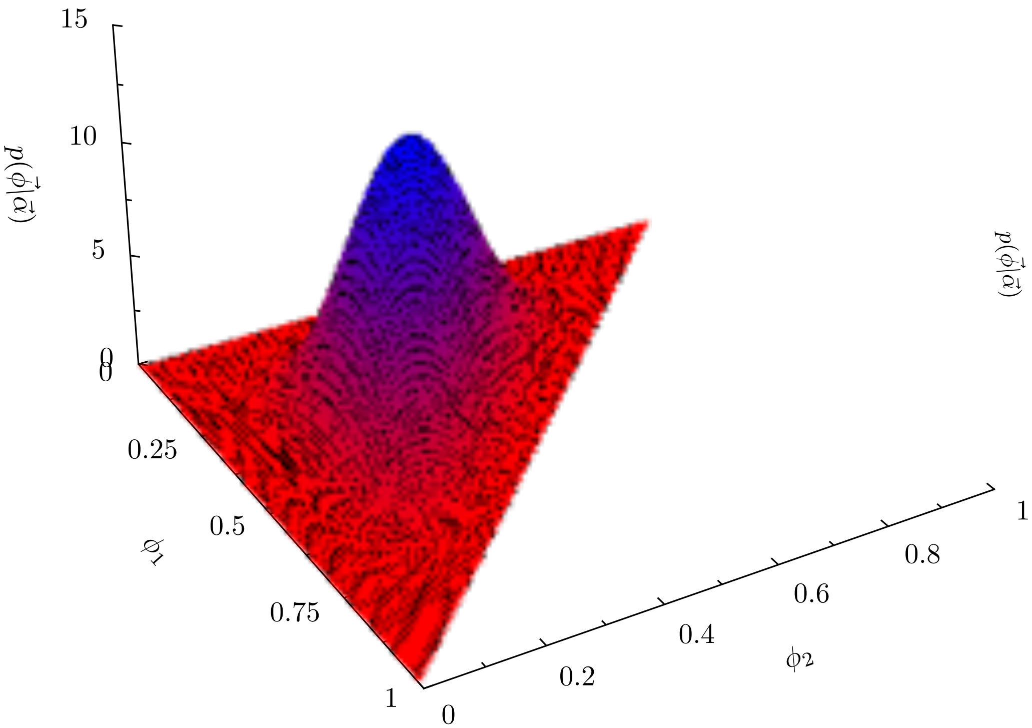
$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

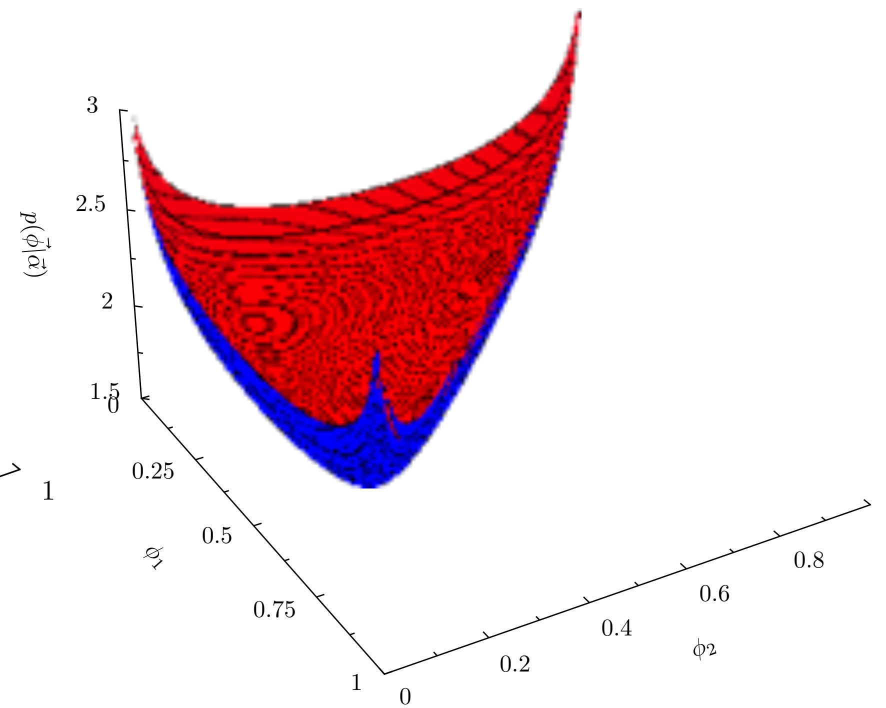


Dirichlet-Multinomial Model

- Dirichlet Distribution

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





Dirichlet-Multinomial Model

- Generative Process

$$\boldsymbol{\phi} \sim \mathrm{Dir}(\boldsymbol{\beta}) \qquad \text{[\textit{draw distribution over words}]}$$

For each word $n \in \{1, \ldots, N\}$

$$x_n \sim \mathrm{Mult}(1, \boldsymbol{\phi}) \qquad \text{[\textit{draw word}]}$$

- 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}

Dirichlet-Multinomial Model

The Dirichlet is **conjugate** to the Multinomial

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

- The posterior of ϕ is $p(\phi|X) = \frac{p(X|\phi)p(\phi)}{P(X)}$
- Define the count vector $\boldsymbol{n}$ such that n_t denotes the number of times word t appeared
- Then the posterior is also a Dirichlet distribution: $p(\phi|X) \sim \text{Dir}(\boldsymbol{\beta} + \boldsymbol{n})$

$$p(\vec{\phi}|\vec{x}, \vec{\beta}) \propto p(\vec{x}|\vec{\phi})\, p(\vec{\phi})$$

$$= \left[\prod_{i=1}^{N} p(x^{(i)}|\vec{\phi})\right] p(\vec{\phi})$$

$$\propto \left[\prod_{k=1}^{K} \phi_k^{\beta_k - 1}\right] \left[\prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{1}(x^{(i)} = k)}\right]$$

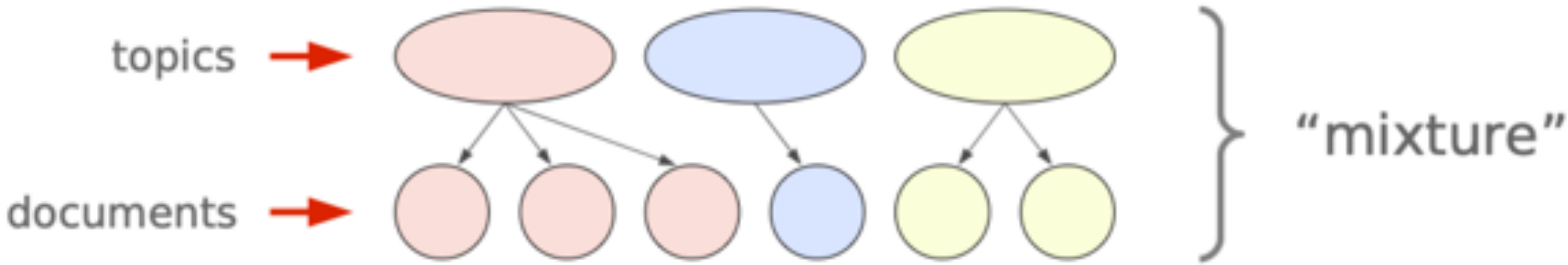
$$= \prod_{k=1}^{K} \phi_k^{\left[\beta_k - 1 + \sum_{i=1}^{N} \mathbb{1}(x^{(i)} = k)\right]}$$

$$\Rightarrow p(\vec{\phi}|\vec{x}, \vec{\beta}) \sim \text{Dirichlet}(\vec{\beta} + \vec{n})$$

where $n_k = \#$ times $x^{(i)} = k$

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

Figure from Wallach, JHU 2011, slides

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

Bayesian Inference for Naïve Bayes

***Whiteboard*:**

- Naïve Bayes is not Bayesian
- What if we observed both words and topics?
- Dirichlet-Multinomial in the fully observed setting is just Naïve Bayes
- Three ways of estimating parameters:
  1. MLE for Naïve Bayes
  2. MAP estimation for Naïve Bayes
  3. Bayesian parameter estimation for Naïve Bayes