Graphical Models

Lecture 21:

Topic Models & Dirichlet Processes

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Yee Whye The, Tom Griffiths and Erik Sudderth for some slide materials.

Background

Dirichlet Distribution

Dirichlet Distribution

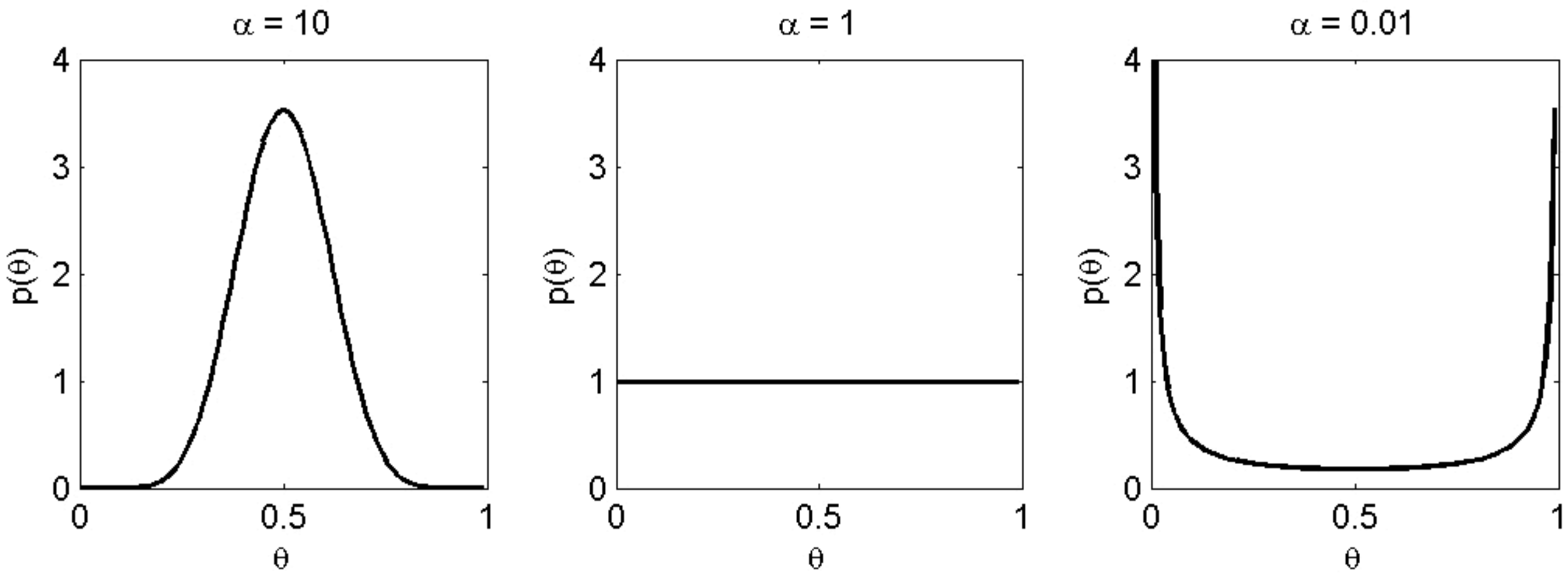
A "dice factory"

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

- This distribution is defined over a "(k-1)-simplex"
  (k non-negative arguments which sum to one).
- The Dirichlet is the conjugate prior to the multinomial.
  (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet!)
- The Dirichlet parameter α_i can be thought of as a prior count of the i^{th} class.
  Question: *How likely is multinomial* θ*?*
  Answer: *What probability would it give to the counts* α_i.

Dirichlet Distribution

- Multivariate equivalent of Beta distribution (a "coin factory")
- Parameters α determine form of the prior



Small α, most mass concentrated on a few outcomes.
Important for later!

Latent Dirichlet Allocation

A tool for discovering interpretable "topics" from large collections of documents

Analysis of PNAS abstracts

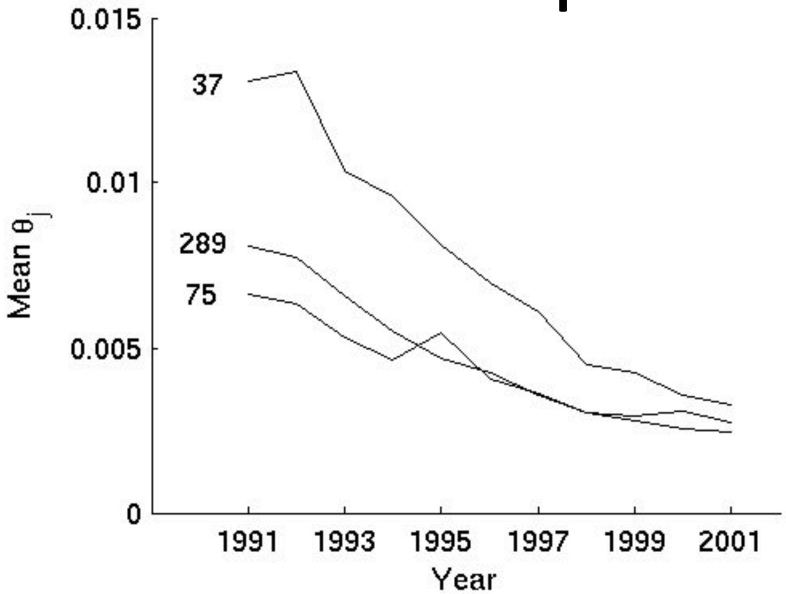
- Test topic models with a real database of scientific papers from PNAS

- All 28,154 abstracts from 1991-2001
- All words occurring in at least five abstracts, not on "stop" list (20,551)
- Total of 3,026,970 tokens in corpus

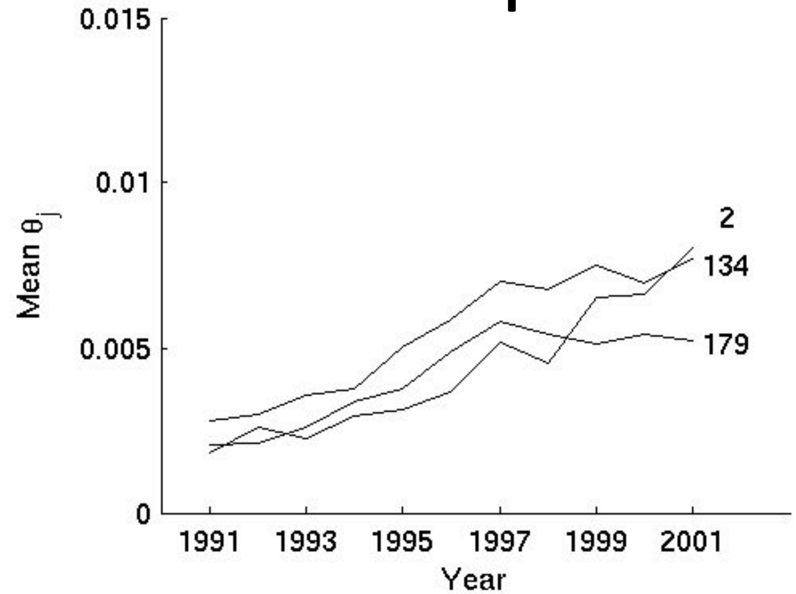
A selection of topics

FORCE	HIV	MUSCLE	STRUCTURE	NEURONS	TUMOR
SURFACE	VIRUS	CARDIAC	ANGSTROM	BRAIN	CANCER
MOLECULES	INFECTED	HEART	CRYSTAL	CORTEX	TUMORS
SOLUTION	IMMUNODEFICIENCY	SKELETAL	RESIDUES	CORTICAL	HUMAN
SURFACES	CD4	MYOCYTES	STRUCTURES	OLFACTORY	CELLS
MICROSCOPY	INFECTION	VENTRICULAR	STRUCTURAL	NUCLEUS	BREAST
WATER	HUMAN	MUSCLES	RESOLUTION	NEURONAL	MELANOMA
FORCES	VIRAL	SMOOTH	HELIX	LAYER	GROWTH
PARTICLES	TAT	HYPERTROPHY	THREE	RAT	CARCINOMA
STRENGTH	GP120	DYSTROPHIN	HELICES	NUCLEI	PROSTATE
POLYMER	REPLICATION	HEARTS	DETERMINED	CEREBELLUM	NORMAL
IONIC	TYPE	CONTRACTION	RAY	CEREBELLAR	CELL
ATOMIC	ENVELOPE	FIBERS	CONFORMATION	LATERAL	METASTATIC
AQUEOUS	AIDS	FUNCTION	HELICAL	CEREBRAL	MALIGNANT
MOLECULAR	REV	TISSUE	HYDROPHOBIC	LAYERS	LUNG
PROPERTIES	BLOOD	RAT	SIDE	GRANULE	CANCERS
LIQUID	CCR5	MYOCARDIAL	DIMENSIONAL	LABELED	MICE
SOLUTIONS	INDIVIDUALS	ISOLATED	INTERACTIONS	HIPPOCAMPUS	NUDE
BEADS	ENV	MYOD	MOLECULE	AREAS	PRIMARY
MECHANICAL	PERIPHERAL	FAILURE	SURFACE	THALAMIC	OVARIAN

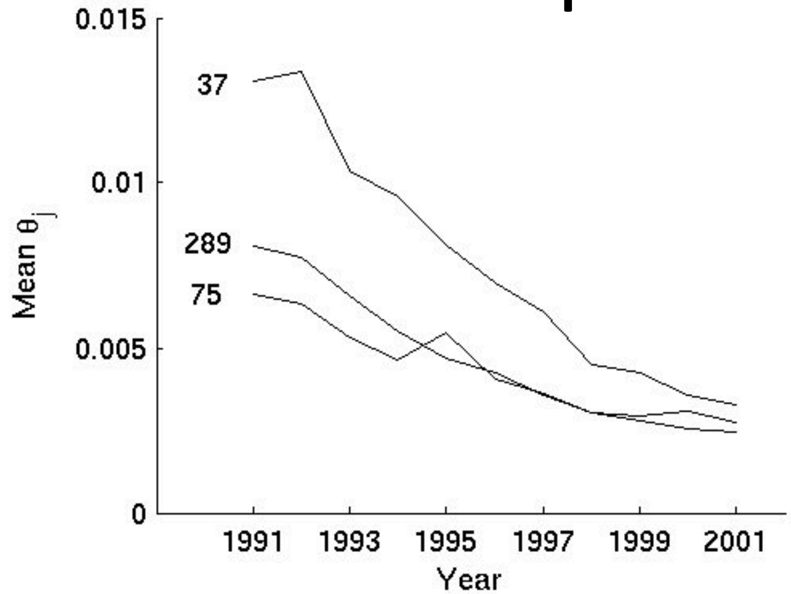
Cold topics



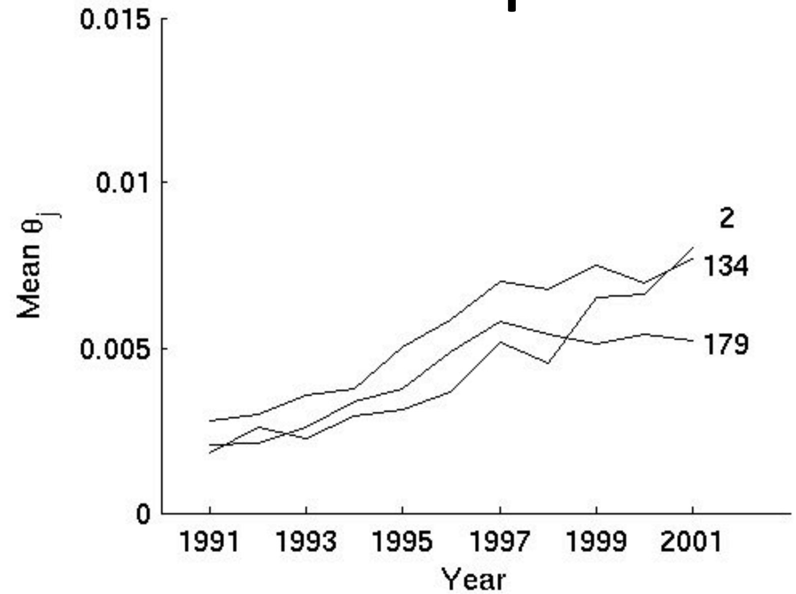
Hot topics



Cold topics

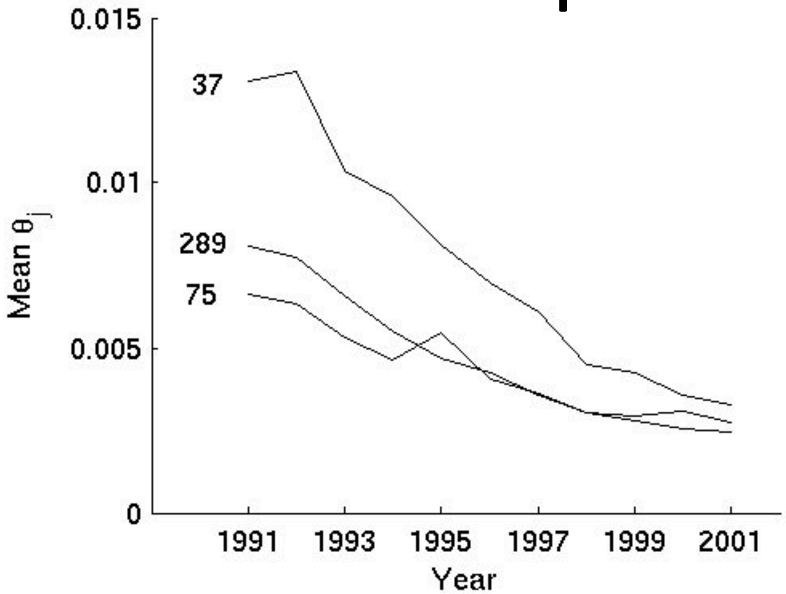


Hot topics

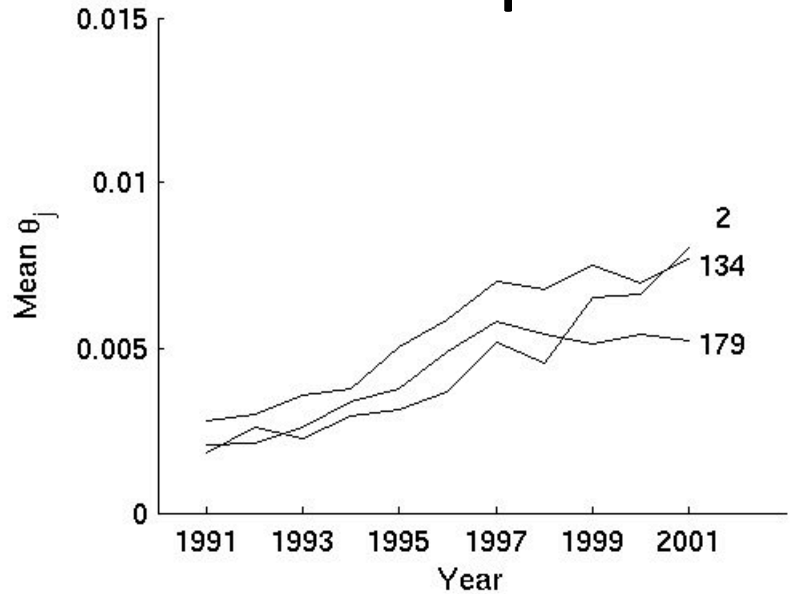


2	**134**	**179**
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED
TERRESTRIAL	LACKING	INDUCTION
COMMUNITY	ANIMALS	CERAMIDE
ABUNDANCE	REDUCED	EXPRESSION

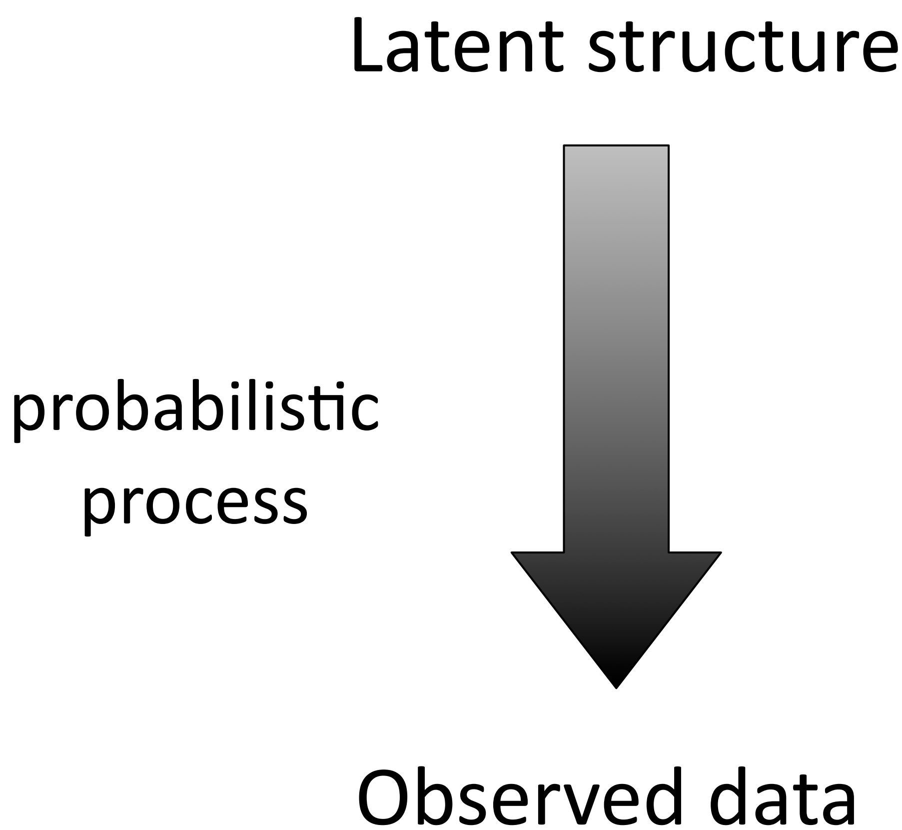
Cold topics



Hot topics

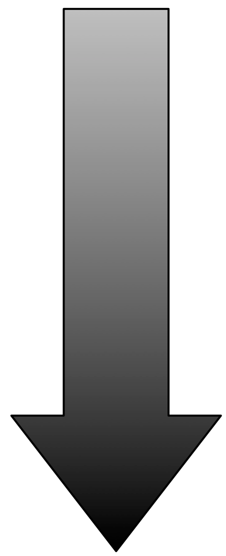


37	**289**	**75**	**2**	**134**	**179**
CDNA	KDA	ANTIBODY	SPECIES	MICE	APOPTOSIS
AMINO	PROTEIN	ANTIBODIES	GLOBAL	DEFICIENT	DEATH
SEQUENCE	PURIFIED	MONOCLONAL	CLIMATE	NORMAL	CELL
ACID	MOLECULAR	ANTIGEN	CO2	GENE	INDUCED
PROTEIN	MASS	IGG	WATER	NULL	BCL
ISOLATED	CHROMATOGRAPHY	MAB	ENVIRONMENTAL	MOUSE	CELLS
ENCODING	POLYPEPTIDE	SPECIFIC	YEARS	TYPE	APOPTOTIC
CLONED	GEL	EPITOPE	MARINE	HOMOZYGOUS	CASPASE
ACIDS	SDS	HUMAN	CARBON	ROLE	FAS
IDENTITY	BAND	MABS	DIVERSITY	KNOCKOUT	SURVIVAL
CLONE	APPARENT	RECOGNIZED	OCEAN	DEVELOPMENT	PROGRAMMED
EXPRESSED	LABELED	SERA	EXTINCTION	GENERATED	MEDIATED
ENCODES	IDENTIFIED	EPITOPES	TERRESTRIAL	LACKING	INDUCTION
RAT	FRACTION	DIRECTED	COMMUNITY	ANIMALS	CERAMIDE
HOMOLOGY	DETECTED	NEUTRALIZING	ABUNDANCE	REDUCED	EXPRESSION

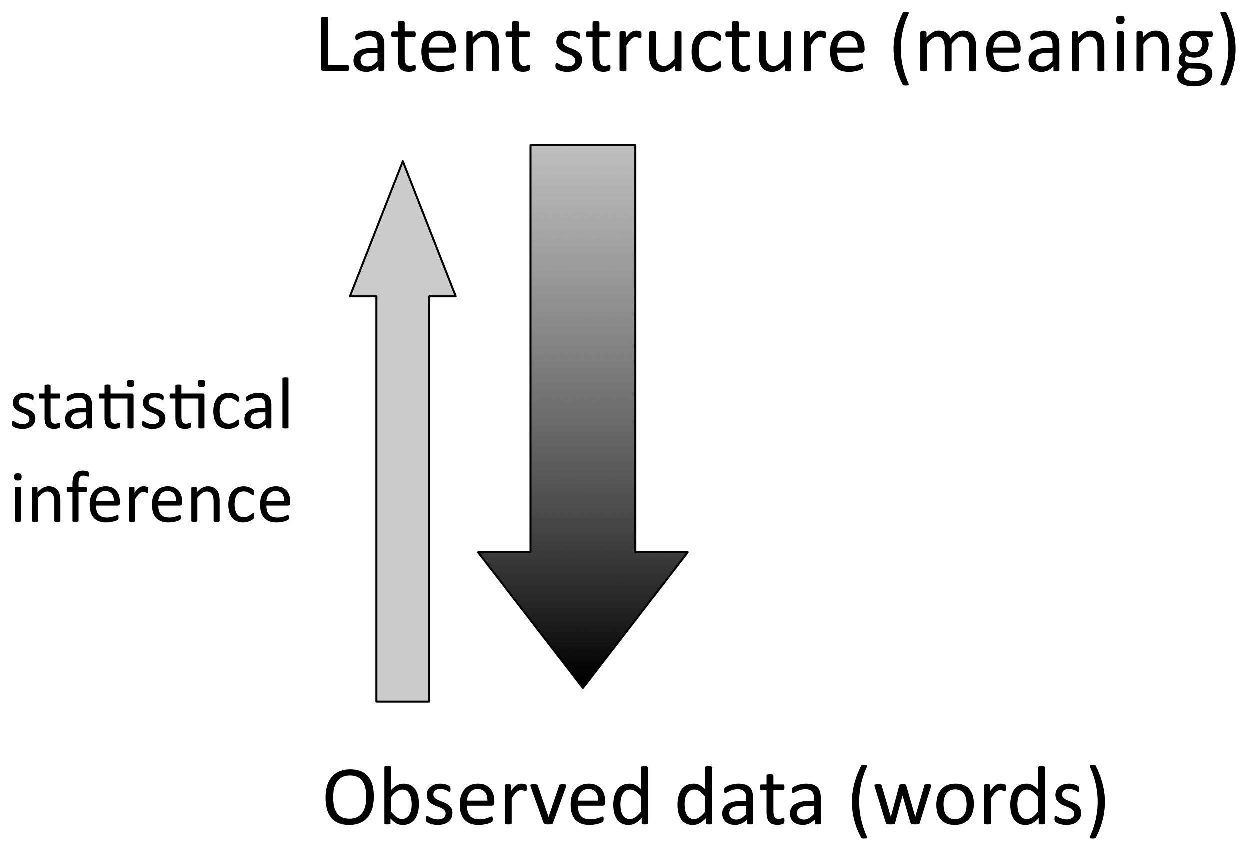

Latent structure
probabilistic process
Observed data

Latent structure (meaning)

probabilistic
process

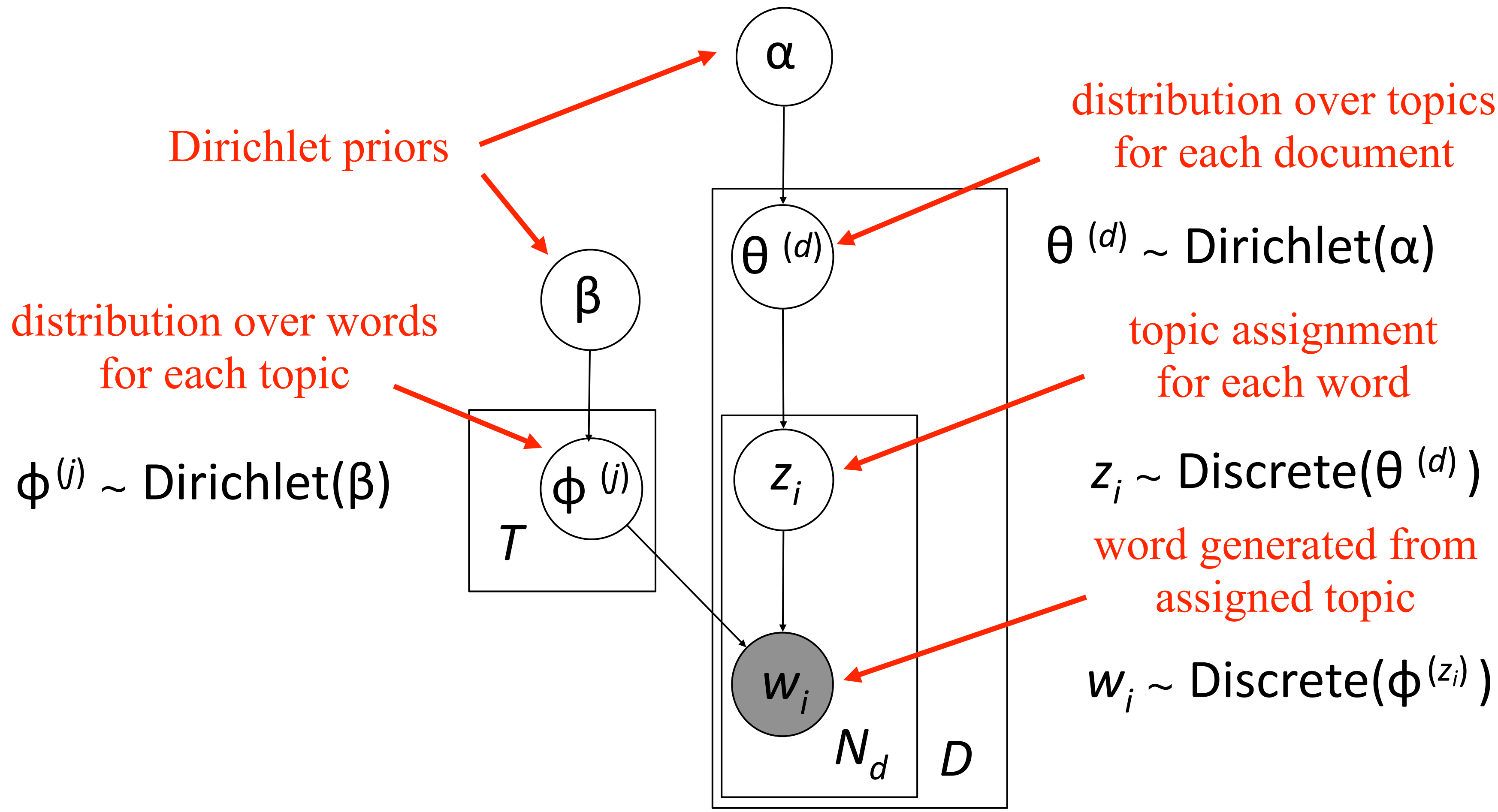


Observed data (words)


Latent structure (meaning)
statistical
inference
Observed data (words)

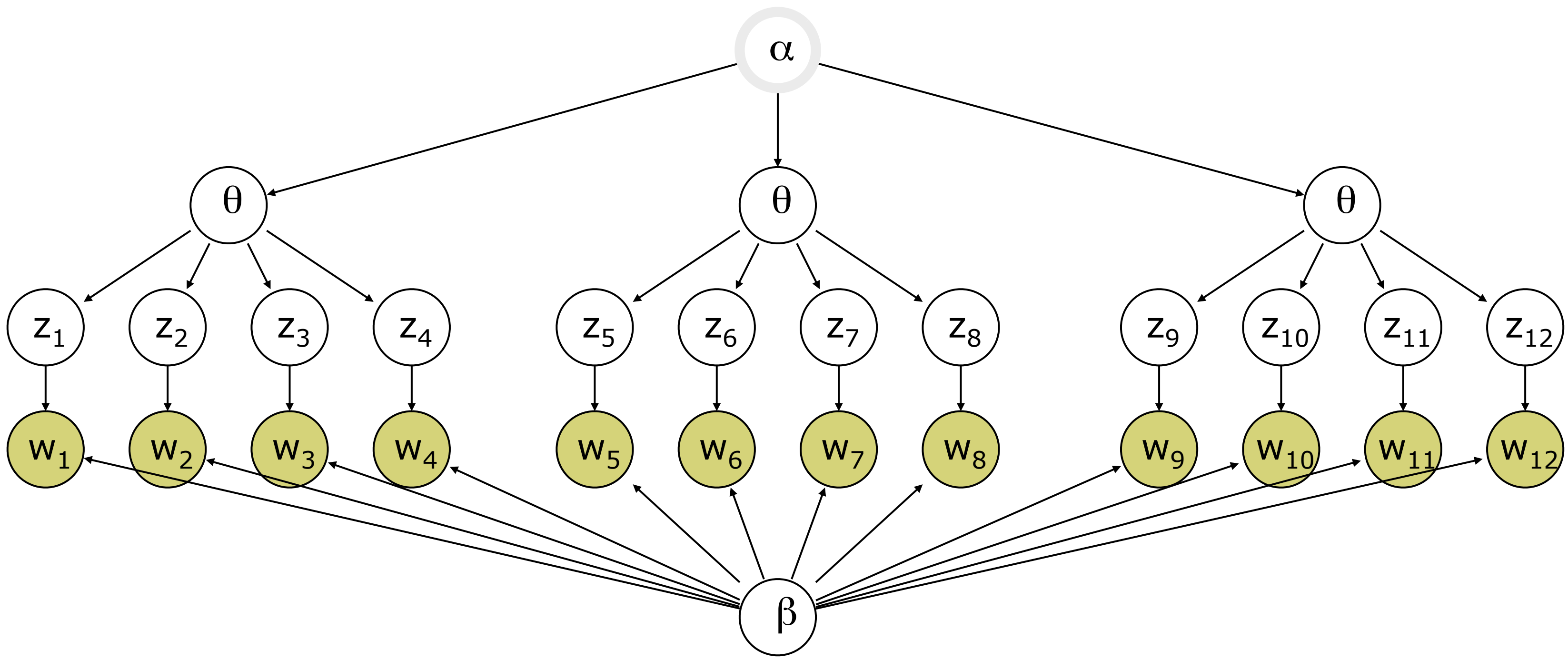
Latent Dirichlet allocation

(Blei, Ng, & Jordan, 2001; 2003)



Inference in LDA

High tree width = intractable exact inference



Approximate inference:

- Variational Mean Field
- Gibbs Sampling
- Collapsed Gibbs Sampling

The collapsed Gibbs sampler

- Using conjugacy of Dirichlet and multinomial distributions, integrate out continuous parameters

$$P(\mathbf{z}) = \int_{\Delta_T^D} P(\mathbf{z} \mid \Theta) p(\Theta) d\Theta \quad = \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(\alpha)^T} \frac{\Gamma(T\alpha)}{\Gamma(\sum_j n_j^{(d)} + \alpha)}$$

$$P(\mathbf{w} \mid \mathbf{z}) = \int_{\Delta_W^T} P(\mathbf{w} \mid \mathbf{z}, \Phi) p(\Phi) d\Phi = \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_w^{(j)} + \beta)}{\Gamma(\beta)^W} \frac{\Gamma(W\beta)}{\Gamma(\sum_w n_w^{(j)} + \beta)}$$

- Defines a distribution on discrete ensembles **z**

$$P(\mathbf{z} \mid \mathbf{w}) = \frac{P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}$$

The collapsed Gibbs sampler

- Sample each z_i conditioned on $\mathbf{z}_{-i}$

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_{j}^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha}$$

- Notation:
  - i indexes over words w and their topic assignments z
  - j indexes over topics
  - $n_{wi}^{(zi)}$ is the number of times word type i occurs in topic z_i
  - $n_j^{(di)}$ is the number of tokens in document d_i assigned to topic j.
  - $n_.^{(zi)}$ is the total number tokens in topic z_i (the "." is wildcard)
  - $n_.^{(di)}$ is the total number of tokens in document d_i

The collapsed Gibbs sampler

- Sample each z_i conditioned on $\mathbf{z}_{-i}$

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha}$$

- This is nicer than your average Gibbs sampler:
  - memory: counts can be cached in two sparse matrices
  - optimization: no special functions, simple arithmetic
  - the distributions on Φ and Θ are analytic given **z** and **w**, and can later be found for each sample

Gibbs sampling in LDA

			iteration
			1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

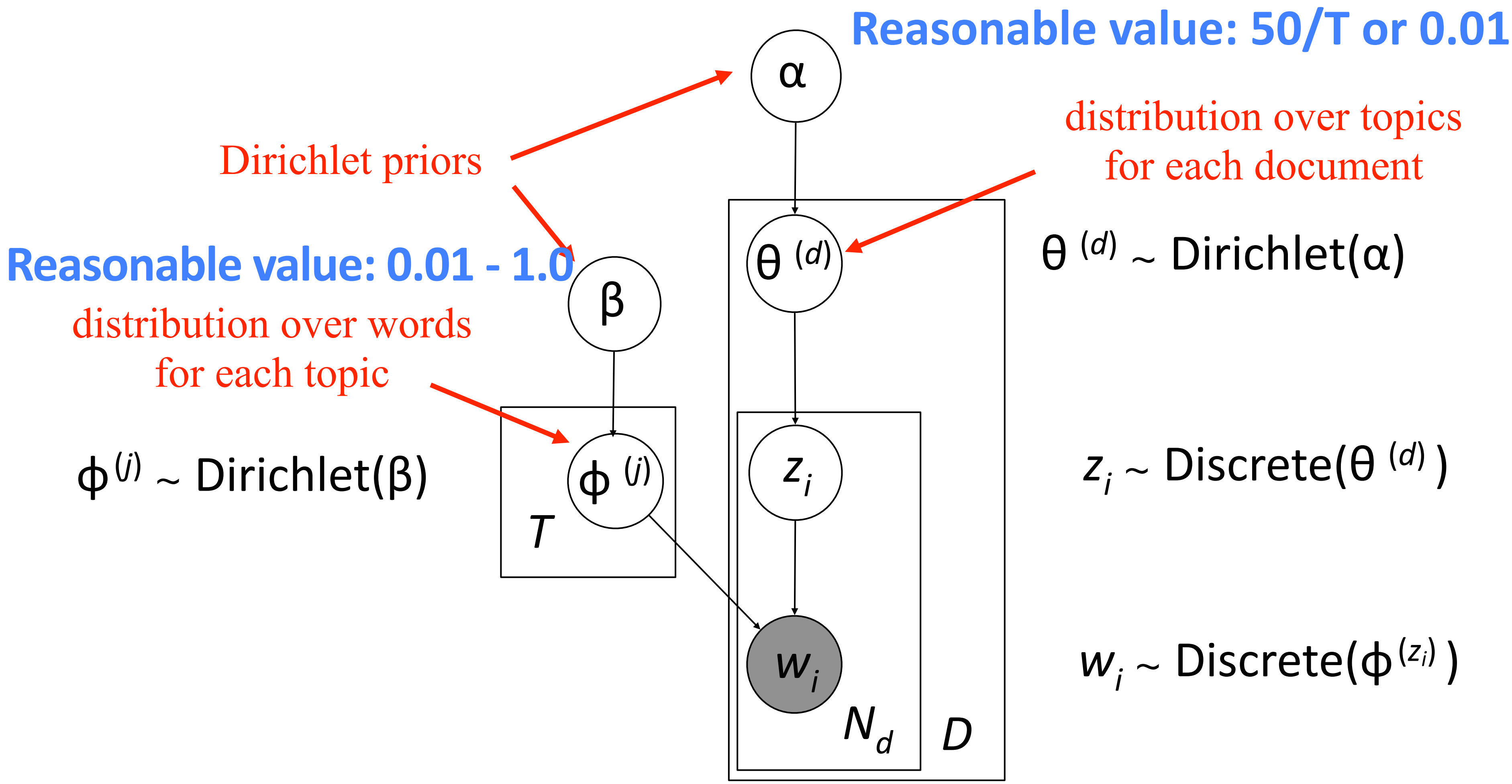
Gibbs sampling in LDA

			iteration 1	2	…	1000
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Latent Dirichlet allocation

(Blei, Ng, & Jordan, 2001; 2003)



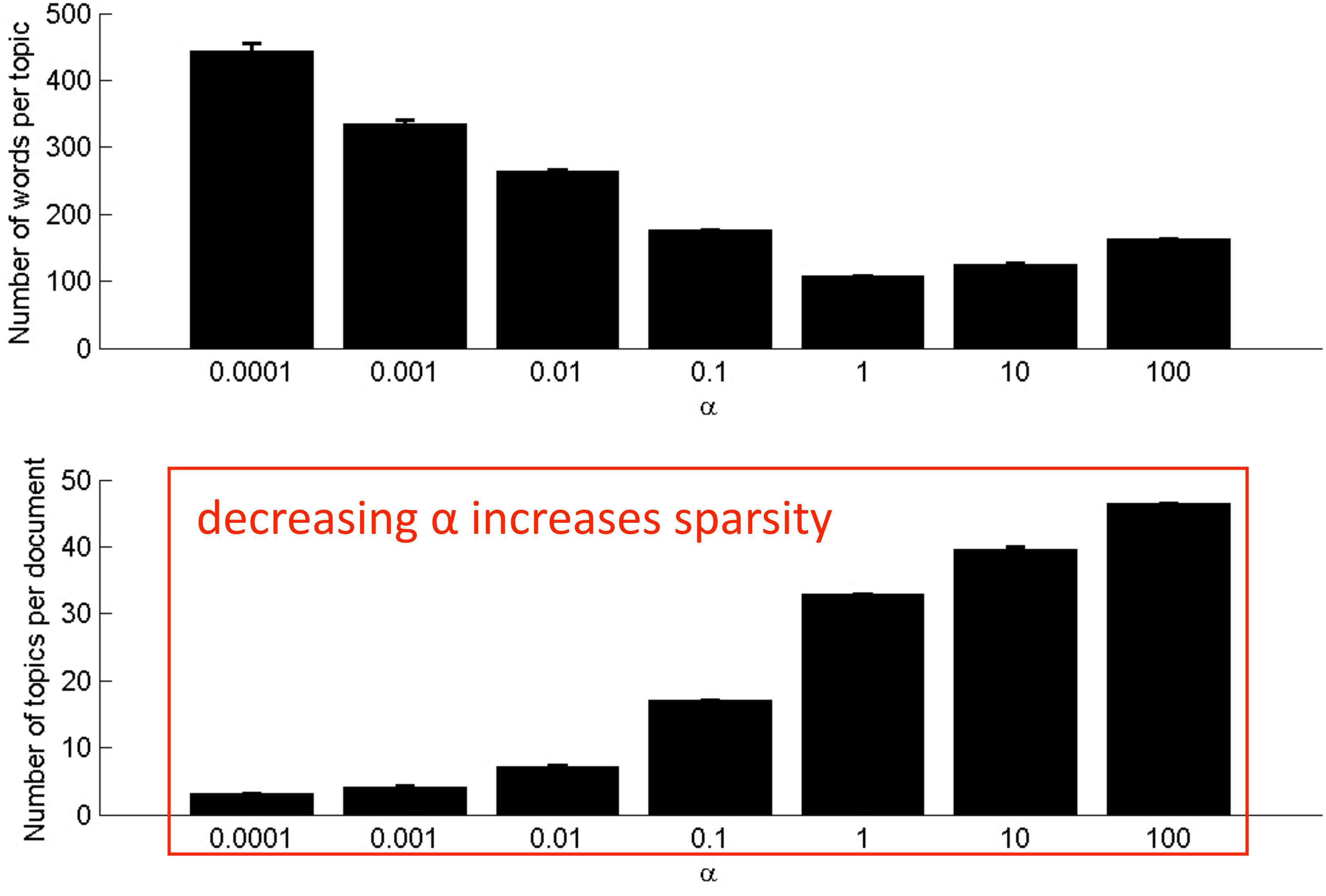
$\theta^{(d)} \sim \text{Dirichlet}(\alpha)$

$\phi^{(j)} \sim \text{Dirichlet}(\beta)$

$z_i \sim \text{Discrete}(\theta^{(d)})$

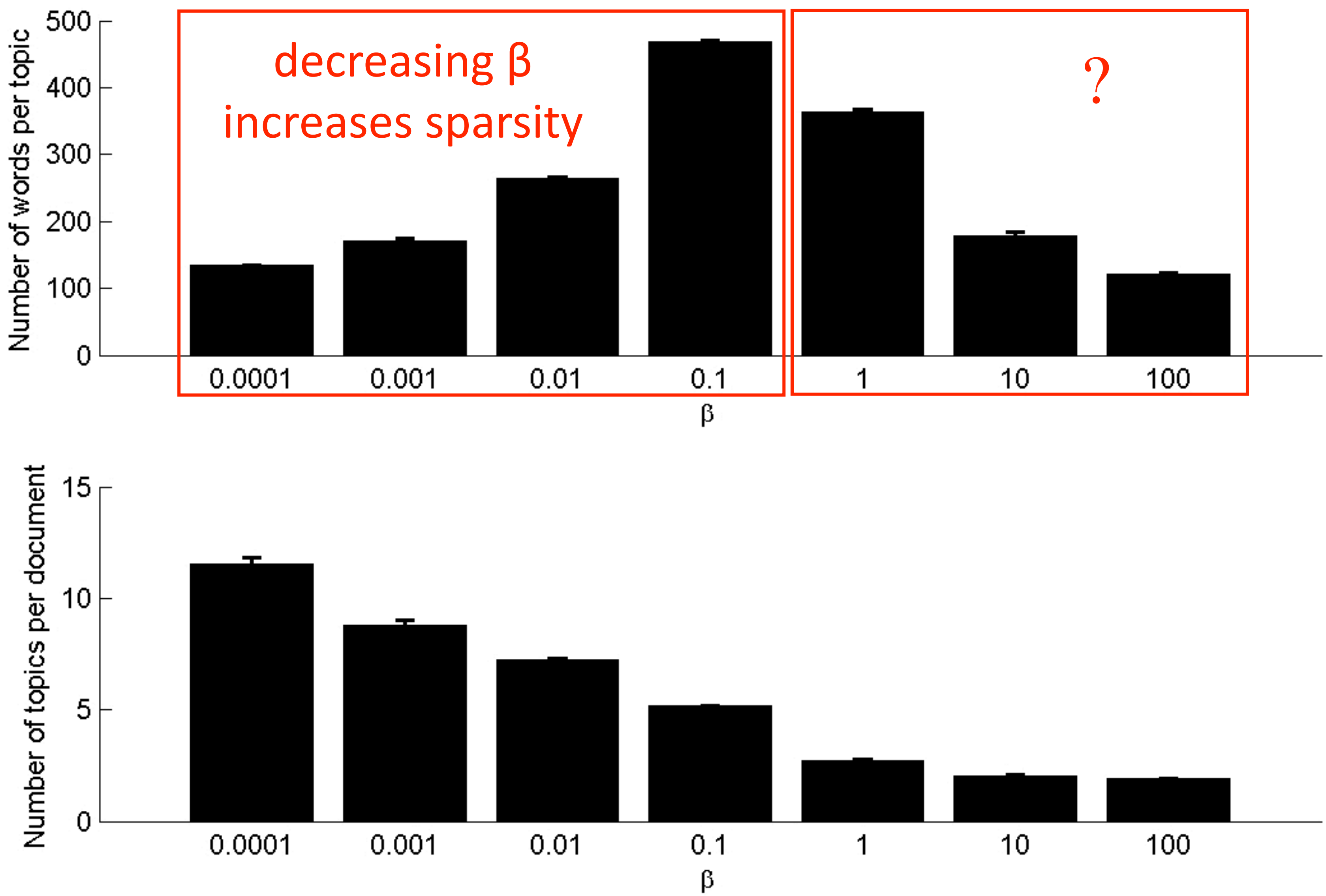
$w_i \sim \text{Discrete}(\phi^{(z_i)})$

Varying α



Number of words per topic
500
400
300
200
100
0
0.0001
0.001
0.01
0.1
1
10
100
α
Number of topics per document
50
40
30
20
10
0
decreasing α increases sparsity
0.0001
0.001
0.01
0.1
1
10
100
α

Varying β



decreasing β
increases sparsity
?
Number of words per topic
500
400
300
200
100
0
0.0001
0.001
0.01
0.1
1
10
100
β
Number of topics per document
15
10
5
0
0.0001
0.001
0.01
0.1
1
10
100
β

Selecting the number of topics

- An example of BN model structure learning
- How to do it?
  - *Earlier lecture*:
    data likelihood + prior preferring smaller structure;
    then try lots of possible structures
  - *Today*:
    define infinite structure with a
    prior enforcing that most of it is rarely used

    Bake together parameter estimation and structure considerations
- Non-parametric models
  - have parameters
  - number of parameters instantiated grows with data

Dirichlet Processes

Can be confusing because
there are different ways to see it.

I'll describe four.

Dirichlet Process as Noisy Copier

- **Motivation**:
  - You have a distribution.
  - You'd like to have a copy
    that can be a little different from the original

- G ~ DP(α, H)

perturbed copy

copy fidelity parameter (higher = closer)

original distribution

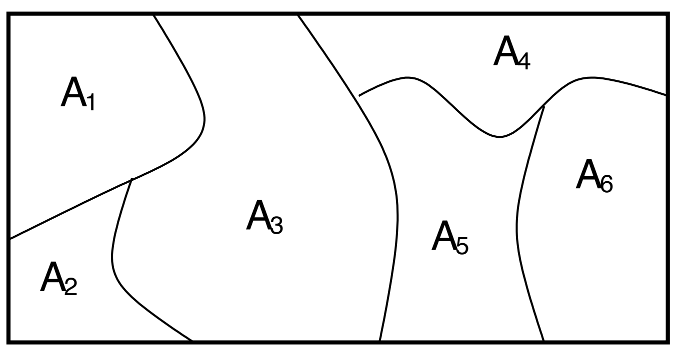
Dirichlet Process Definition

- A **Dirichlet Process** (DP) is a distribution over prob distributions.
  (More correctly, instead of "prob distribution" we should say "probability measure" which works on continuous domains.)
- A DP has two parameters:
  - **Base distribution** H (which is the mean of the DP).
  - **Strength parameter** α (which is like an inverse-variance of the DP).
- We write:

like a game: opponent gets to pick the partitioning; DP generates a bunch of Gs.

$G \sim DP(\alpha,H)$ if for any partition $(A1,\ldots,An)$ of $\boldsymbol{X}$:

$$(G(A1), \ldots, G(An)) \sim Dirichlet(\alpha H(A1), \ldots, \alpha H(An))$$



Summary: A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.

Dirichlet Process Definition

- Sounds magical! Is this even possible?
  Yes.
- Fact: Samples from a DP are always discrete distributions.
- Intuition:
  - Make an infinite number of samples from original H, but re-weight them according to draw from Dirichlet with uniform mean and concentration related to α.
  - With small α, most mass will be on just a few samples. (Will probably never even see the other samples.)

Constructing a DP draw

Blackwell-MacQueen Urn Scheme

- Imagine picking balls of different colors from an urn. Start with no balls in the urn.

- For the *n*th draw, 1...∞:
  - with probability ∝ α, draw $\theta_n \sim H$, and add a ball of that color into the urn.
  - With probability ∝ *n − 1*, pick a ball at random from the urn, record θ_n to be its color, return the ball into the urn and place a second ball of same color into urn.

Note: For large α, mostly just draw from H. For small α, often copy an old value, perturbing G away from H.

Blackwell-MacQueen urn scheme is like a "representer" for the DP—a finite projection of an infinite object.

NEXT: We'd like to know G(x) for each different color x. Need to gather all balls of same color and count them...

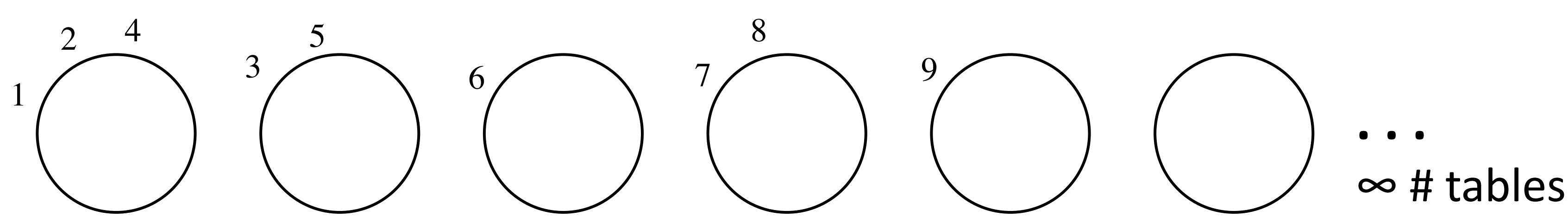
Alternative view of the same construction

Chinese Restaurant Process

Use de Finetti's Theorem about exchangeability
to gather together balls of the same color ...into "restaurant tables"

customer = urn scheme draw
table = ball color = θ_i

- Generating from the CRP:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $n_k/(\alpha+n-1)$, n_k = # people @ table k
    - A new table K+1 with probability $\alpha/(\alpha+n-1)$



customers at a table = (re)-weighting of that table's value.
Most mass focussed on early tables

NEXT: We'd like to know G(x) for each different color x without having to simulate an infinite # customers...

Alternative view of the same construction

Stick Breaking Construction

Answers: “What are the table-weights when there are an infinite number of customers?”

What do draws G ~ DP(a,H) look like?

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

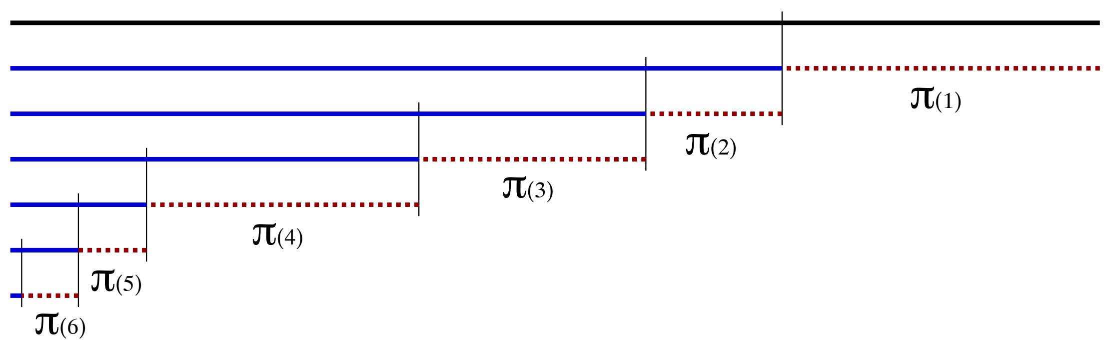
$\delta_{\theta k}$ = point mass on θ_k

where

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

$$\beta_k \sim \text{Beta}(1, \alpha)$$

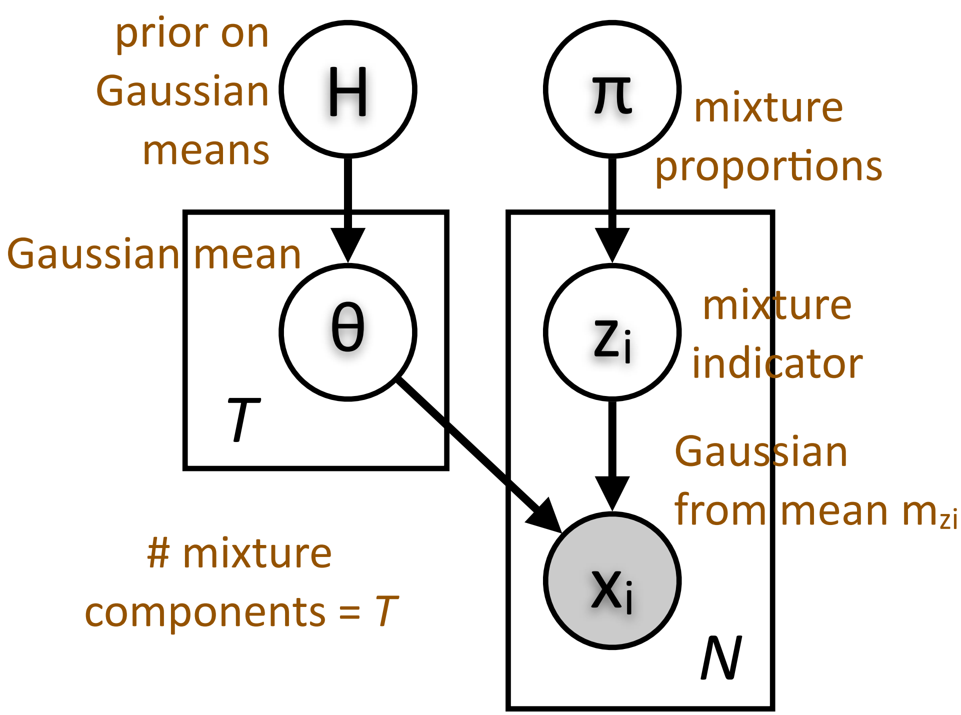
$$\theta_k^* \sim H$$



What does all this have to do with non-parametric infinite mixtures?!

Finite Mixture Model

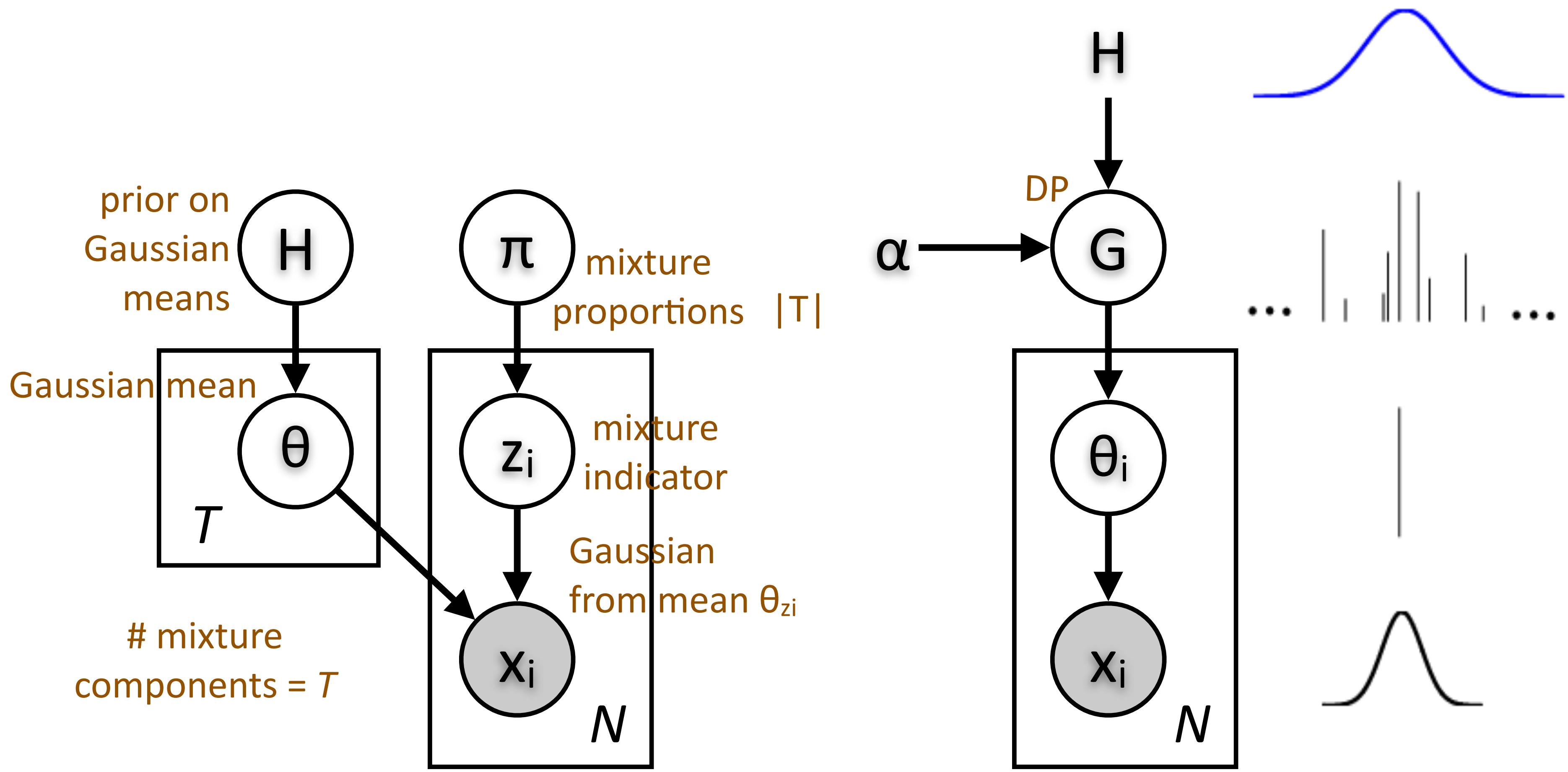
E.g. Mixture of Gaussians



Finite Mixture Model

E.g. Mixture of Gaussians

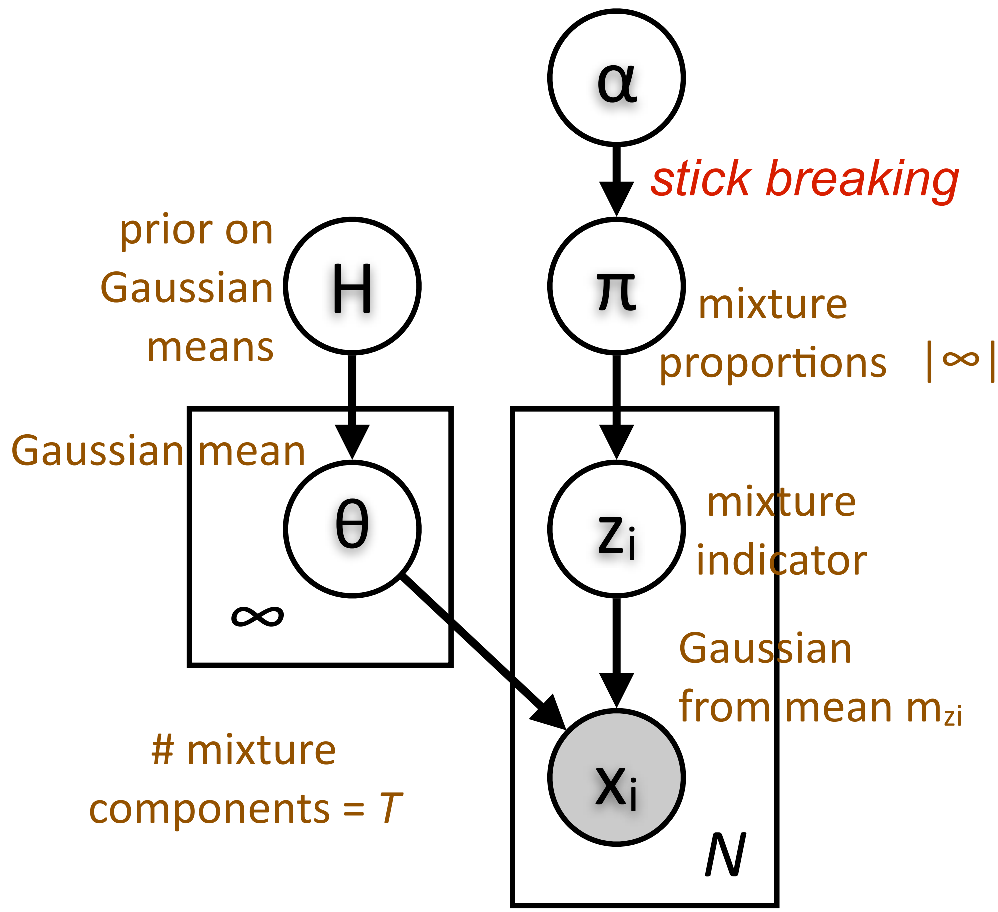
DP (Infinite) Mixture Model



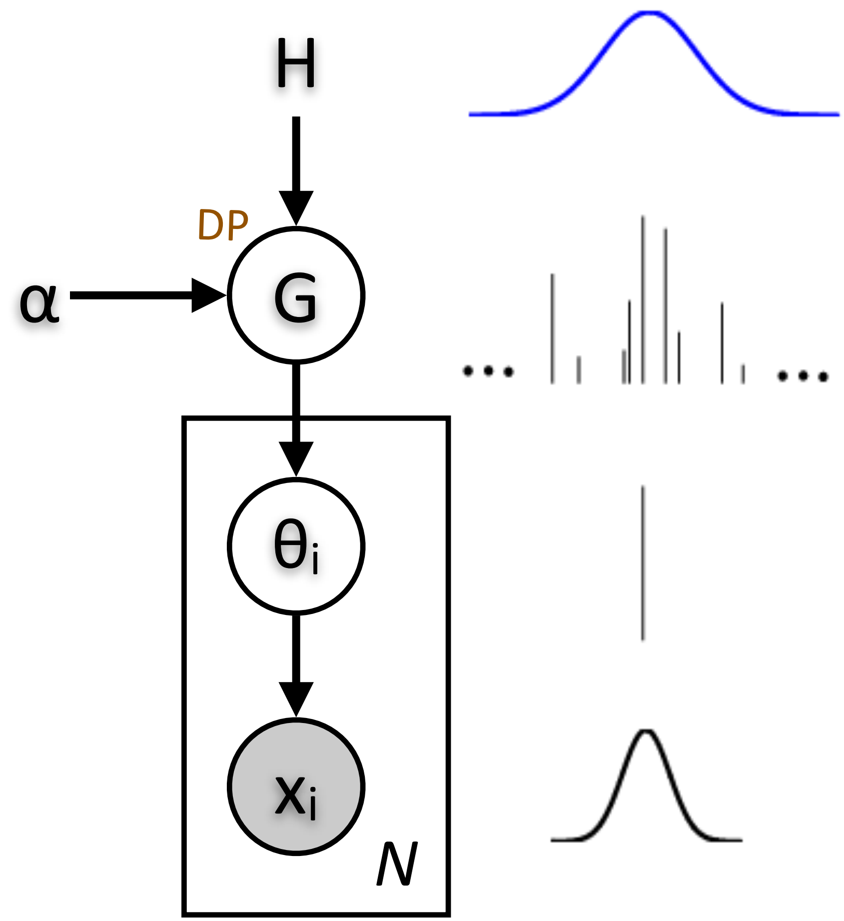
Alternative Diagram

DP (Infinite) Mixture Model

E.g. Mixture of Gaussians



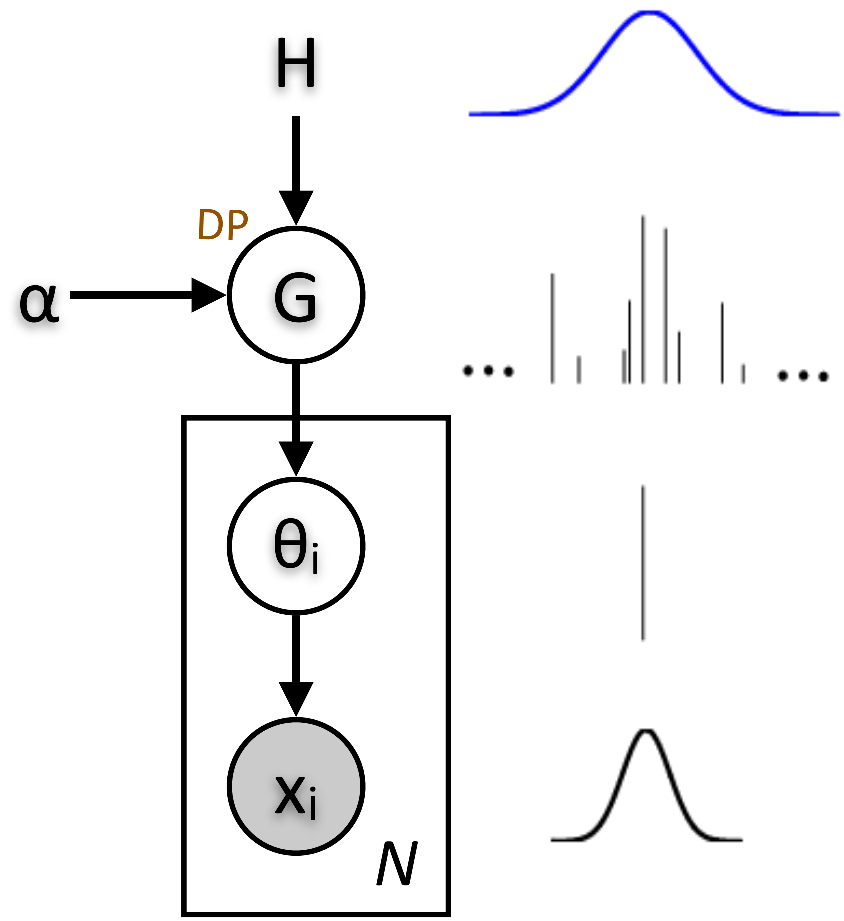
DP (Infinite) Mixture Model



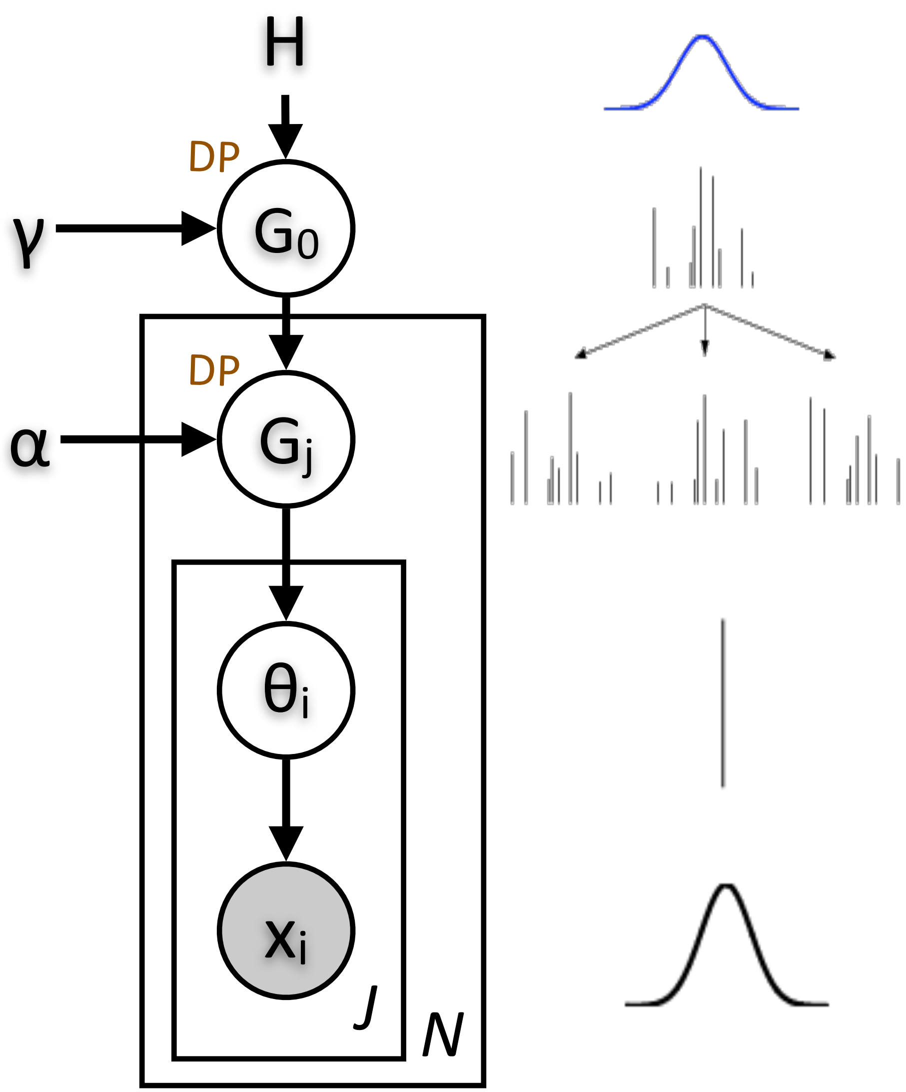
Getting back to LDA...

- We want to generate a corpus of documents from a set of shared "topics"
- The DP Mixture Model does not explicitly enforce any sharing. (Alternatively: the DP Mixture confounds the mixture values and mixture proportions.)
- We need something more...

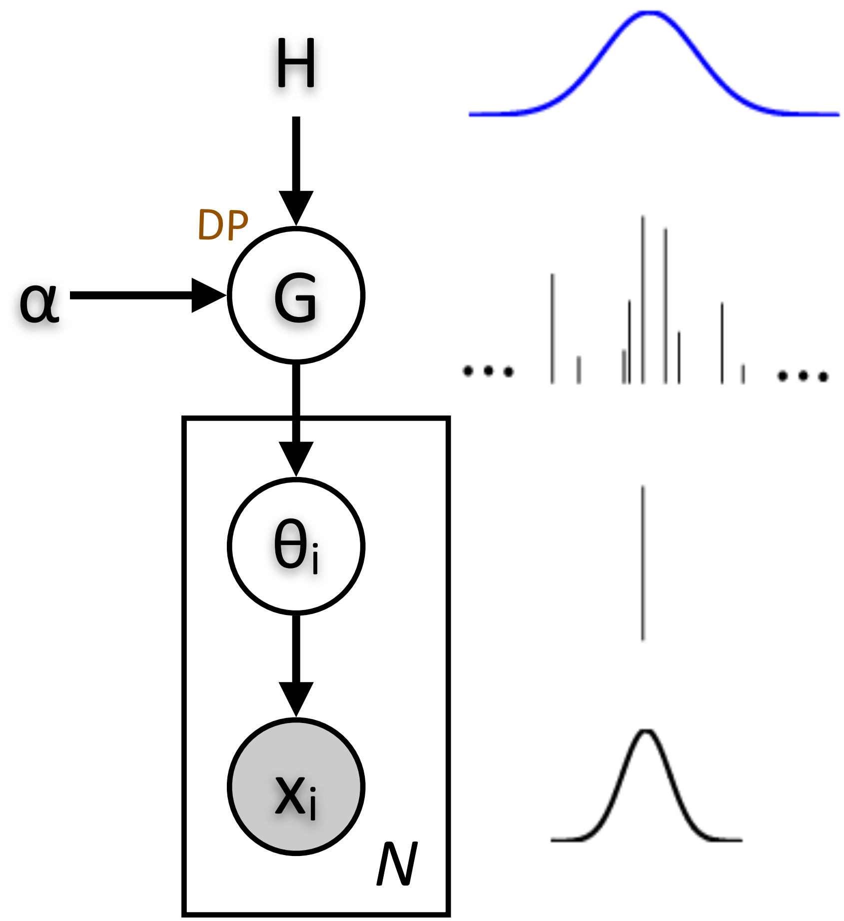
DP (Infinite) Mixture Model



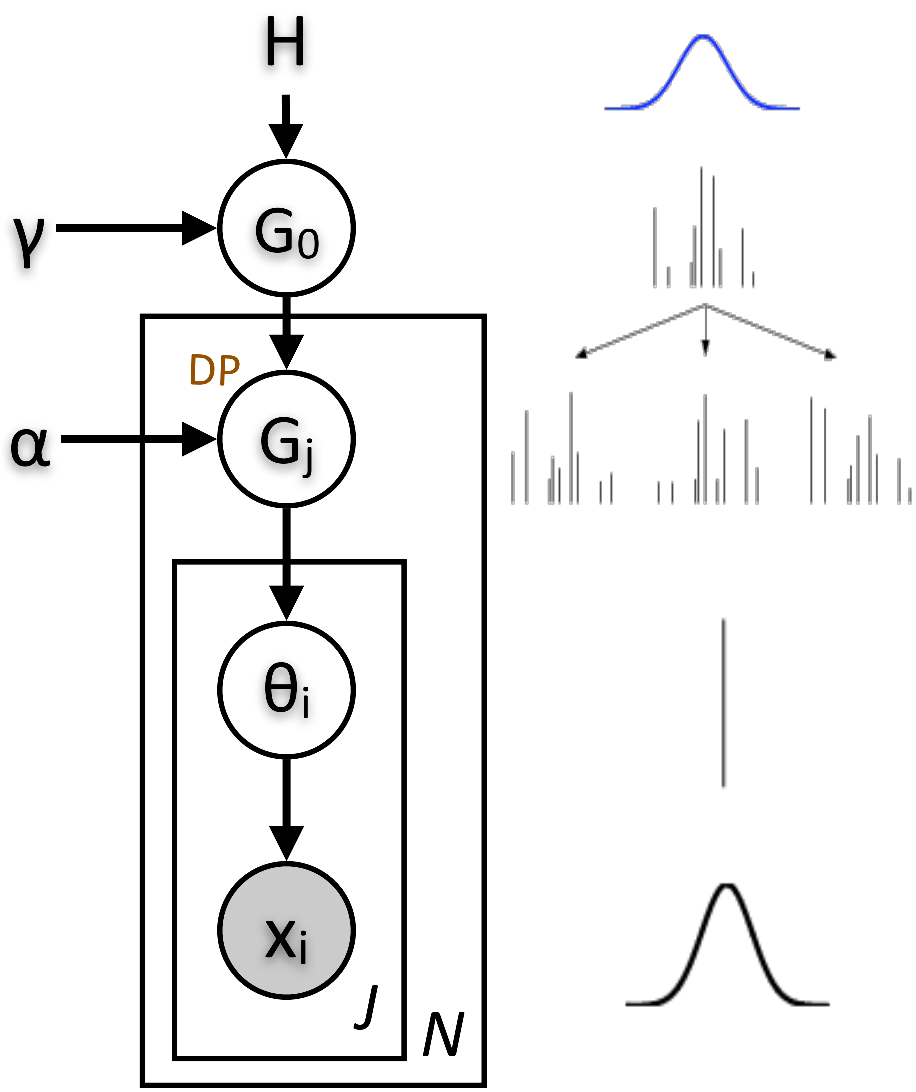
Hierarchical DP Mixture Model



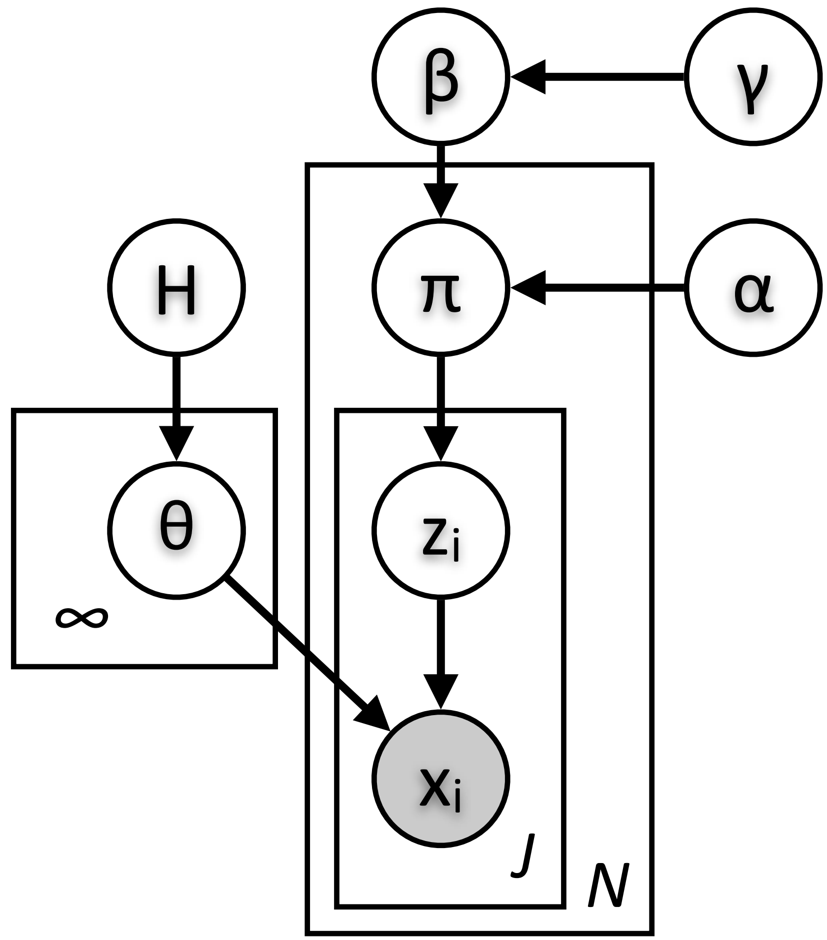
DP (Infinite) Mixture Model



Hierarchical DP Mixture Model



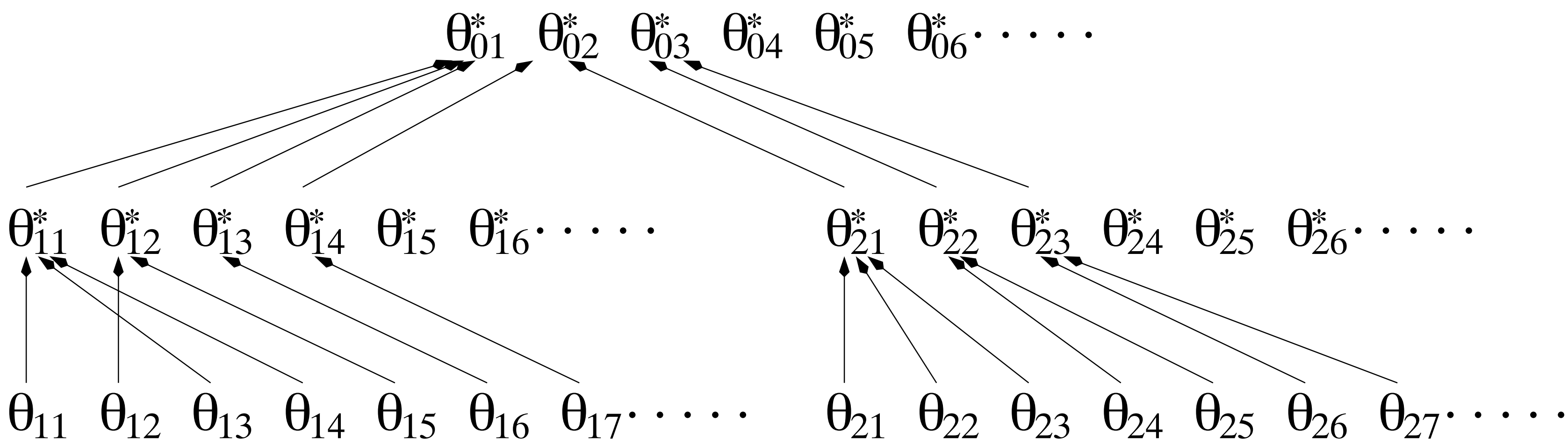
Alternative Diagram



Another Picture of the HDP

Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.

The hierarchical Pòlya urn scheme to generate draws from G_1, G_2:



Inference in the Dirichlet Process Mixture Model

Collapsed Gibbs Recipe for DP Mixture

The big picture:

- For each data point
  - pretend it is the last point (by exchangeability)
  - the prior is just the Chinese restaurant dynamics
  - the likelihood is just the usual mixture likelihood

Collapsed Gibbs Recipe for DP Mixture

The slightly more detailed picture, still skipping the evidence (word) likelihood:

- For the *n*th word:
  - with probability $\propto \alpha$, draw $z_n \sim G_0$.
    - To draw from G_0: with probability $\propto \gamma$, draw $z_n \sim H$, with prob $\propto$ n-1, draw a topic from those already in G_0 proportionally according to their counts.
  - With probability $\propto n_j - 1$, draw a topic from those already in G_j proportionally according to counts.

From previously...

The finite collapsed Gibbs sampler

Sample each z_i conditioned on $\mathbf{z}_{-i}$

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha}$$

For the DP (infinite) case:

- Very similar, but include the possibility of picking a "new" topic, using Chinese restaurant dynamics.
- When you pick a "new" topic for a document, first go to G_0 and consider using a "old new" topic from another document, otherwise create a "new new" topic.

Generic Collapsed Gibbs Sampler for DP Mixture Model

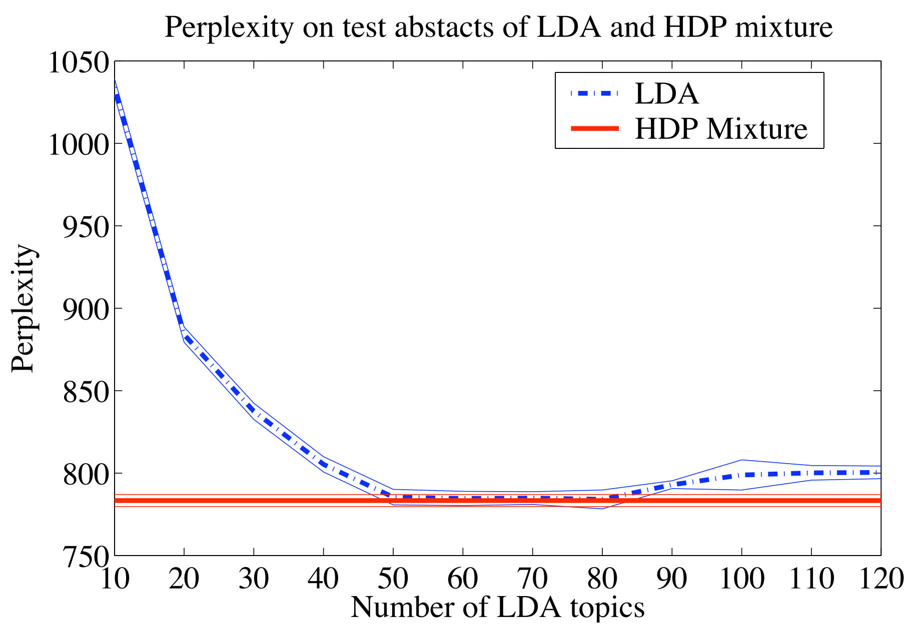
[Sudderth PhD]
[Neal 2000, Alg #2]

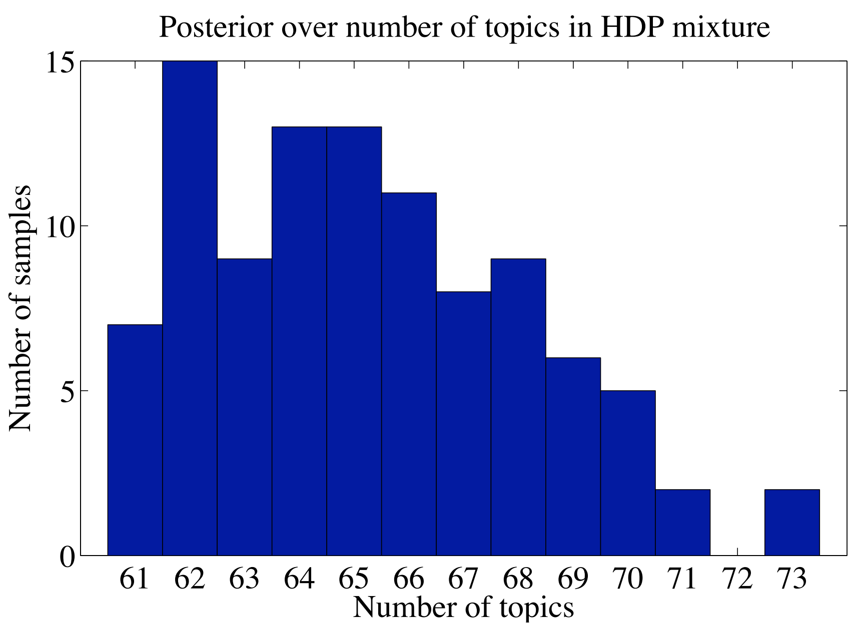
Given the previous concentration parameter $\alpha^{(t-1)}$, cluster assignments $z^{(t-1)}$, and cached statistics for the K current clusters, sequentially sample new assignments as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $\alpha = \alpha^{(t-1)}$ and $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, resample z_i as follows:
   (a) For each of the K existing clusters, determine the predictive likelihood
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   This likelihood can be computed from cached sufficient statistics via Prop. 2.1.4. Also determine the likelihood $f_{\bar{k}}(x_i)$ of a potential new cluster $\bar{k}$ via eq. (2.189).
   (b) Sample a new cluster assignment z_i from the following $(K+1)$–dim. multinomial:
   $$z_i \sim \frac{1}{Z_i}\left(\alpha f_{\bar{k}}(x_i)\delta(z_i, \bar{k}) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)\delta(z_i, k)\right) \qquad Z_i = \alpha f_{\bar{k}}(x_i) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)$$
   N_k^{-i} is the number of other observations currently assigned to cluster k.
   (c) Update cached sufficient statistics to reflect the assignment of x_i to cluster z_i. If $z_i = \bar{k}$, create a new cluster and increment K.
3. Set $z^{(t)} = z$. Optionally, mixture parameters for the K currently instantiated clusters may be sampled as in step 3 of Alg. 2.1.
4. If any current clusters are empty ($N_k = 0$), remove them and decrement K accordingly.
5. If $\alpha \sim \text{Gamma}(a, b)$, sample $\alpha^{(t)} \sim p(\alpha \mid K, N, a, b)$ via auxiliary variable methods [76].

HDP Mixture Experimental Results

- Compared against latent Dirichlet allocation, a parametric version of the HDP mixture for topic modelling.

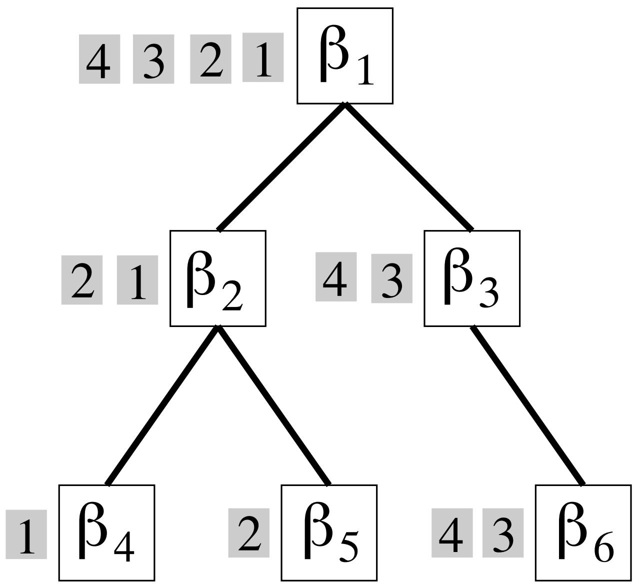


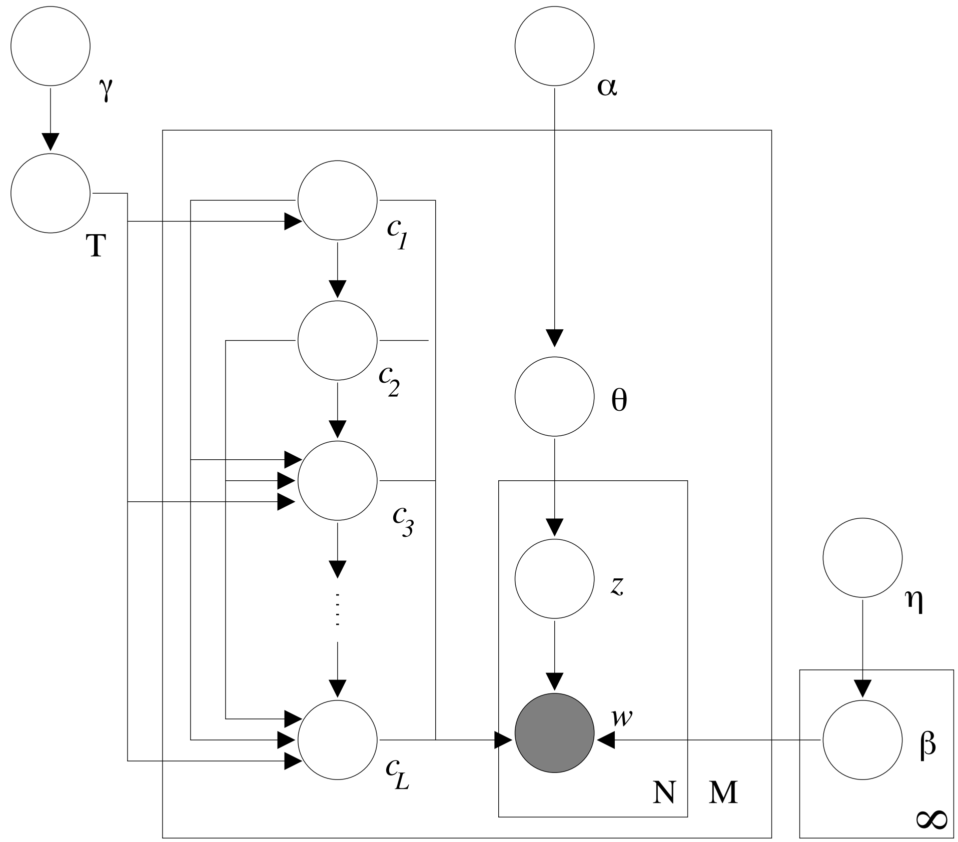


Further Variations

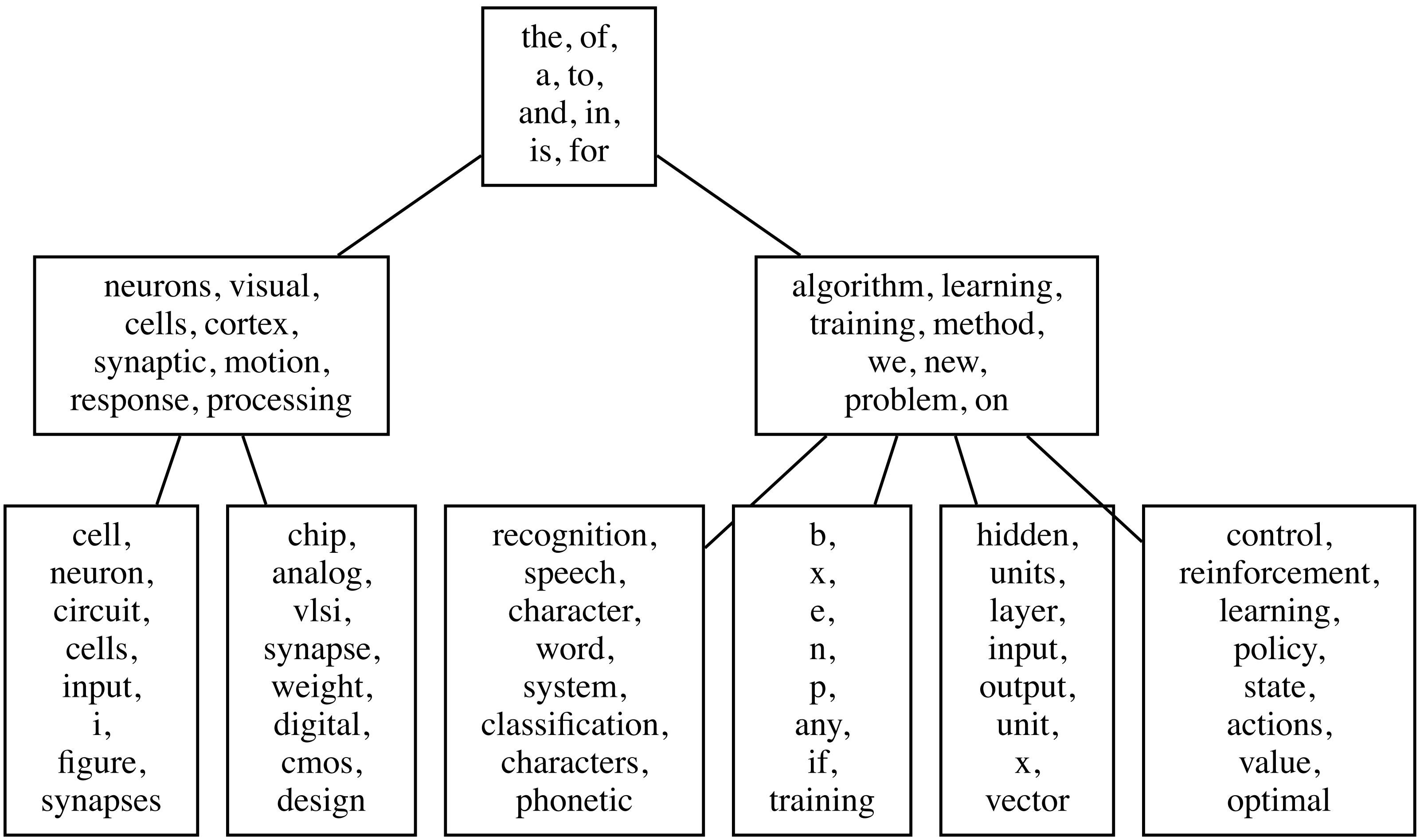
Nested Chinese Restaurant Process

[Blei et al 2003]





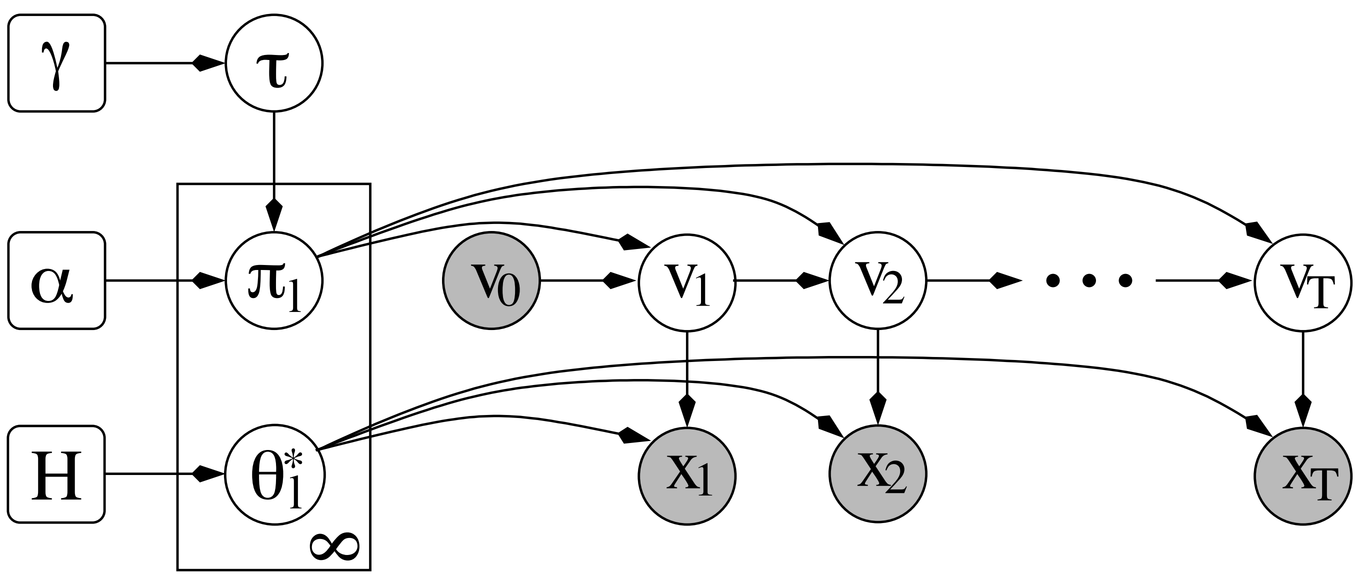
Nested Chinese Restaurant Process



Infinite Hidden Markov Model

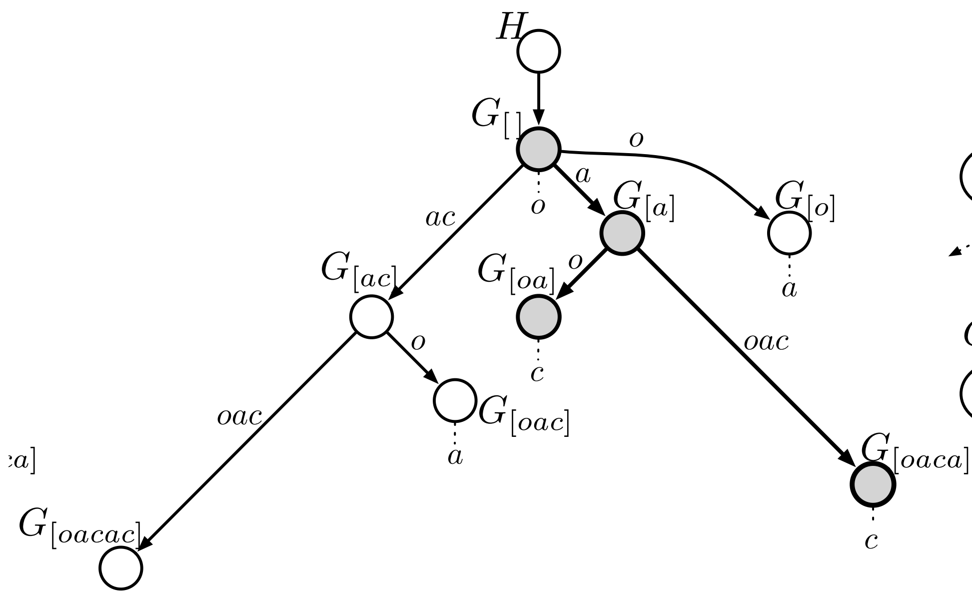
Implement sharing of next states using a HDP:

$$(\tau_1, \tau_2, \ldots) \sim \text{GEM}(\gamma)$$
$$(\pi_{1l}, \pi_{2l}, \ldots) | \tau \sim \text{DP}(\alpha, \tau)$$

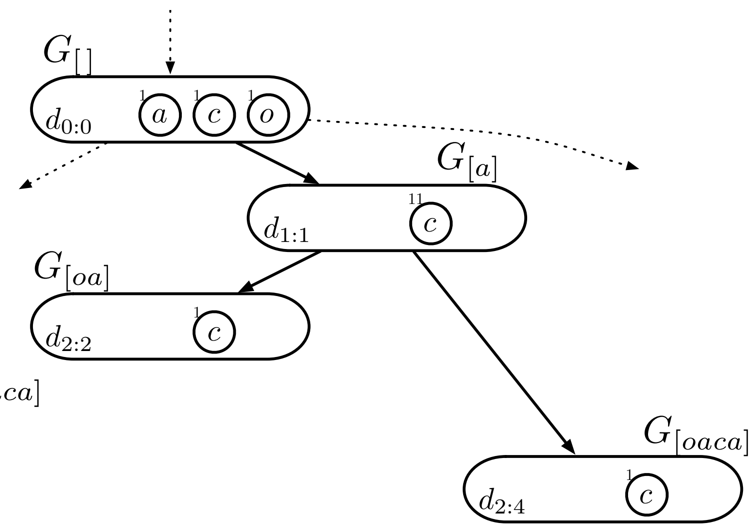


Infinite N-gram Model

"A Stochastic Memoizer for Sequence Data"
[Wood, Archambeau, Gasthaus, James, Teh, 2009]



(b) Prefix tree for *oacac*.



(c) Initialisation.

Readings for More Detail

- Erik Sudderth's PhD thesis
  http://www.cs.brown.edu/~sudderth/papers/sudderthPhD.pdf
- Yee Whye Teh's Dirichlet Process Tutorial
  http://www.gatsby.ucl.ac.uk/~ywteh/teaching/npbayes/mlss2007.pdf
- HDP introduction, LDA with infinite topics
  http://www.cse.buffalo.edu/faculty/mbeal/papers/hdp.pdf
- HDP implementation by Teh.
  http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/npbayes-r21.tgz