Probabilistic Graphical Models: Parametric and Nonparametric Perspectives

Michael I. Jordan

Department of Statistics and Computer Science Division
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

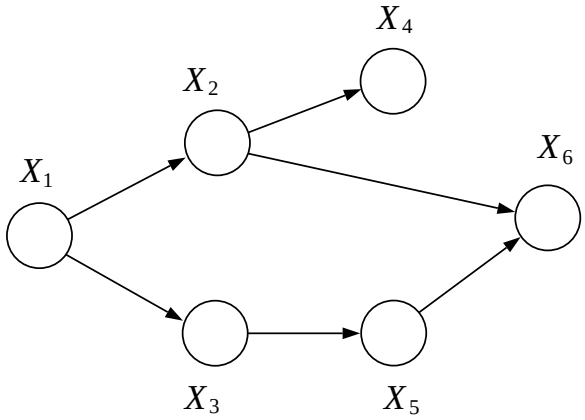
Acknowledgments: David Blei, Yee Whye Teh, Martin Wainwright

Graphical Models

- marriage of graph theory and probability theory
- special cases of the basic algorithms discovered in many (dis)guises:
  - state-space filters
  - error-control coding (decoding for low-density parity check codes)
  - statistical physics
  - hidden Markov models
  - genetics (peeling, pruning)
  - database theory (triangulation)
  - statistics
  - artificial intelligence
- fullest expression of symbolic inference: *junction tree algorithm*
  - important special case: *sum-product algorithm*
- numerous applications (bioinformatics, speech, vision, robotics, diagnostics, error-control coding, compression, channel modeling, control, optimization)

Directed Graphical Models

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:

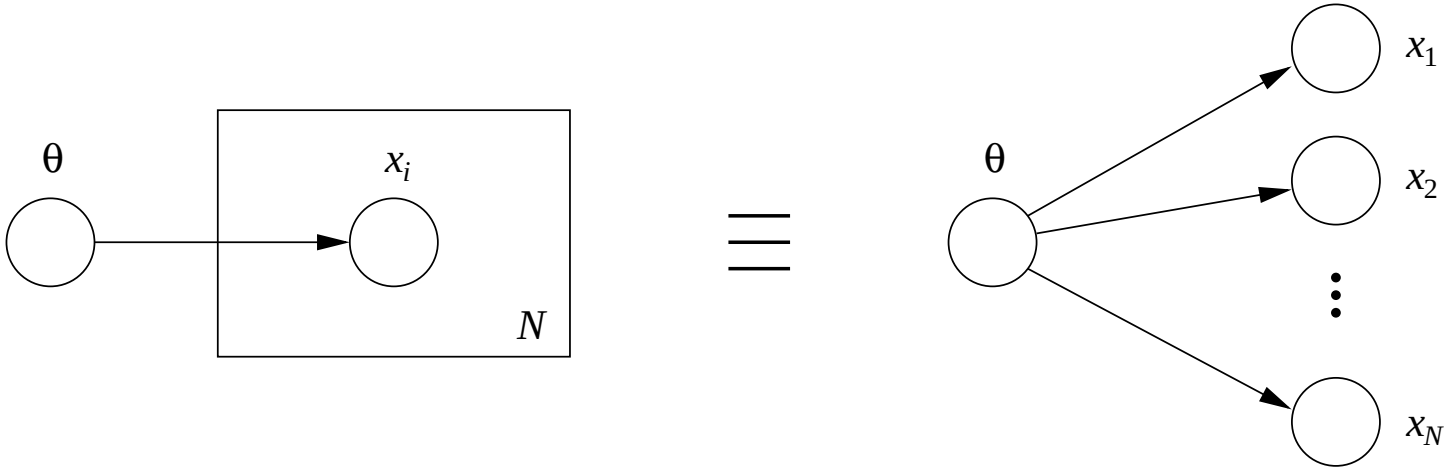


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

$$p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = p(x_1; \theta_1)\ p(x_2 \,|\, x_1; \theta_2)$$
$$p(x_3 \,|\, x_1; \theta_3)\ p(x_4 \,|\, x_2; \theta_4)\ p(x_5 \,|\, x_3; \theta_5)\ p(x_6 \,|\, x_2, x_5; \theta_6)$$

Plates

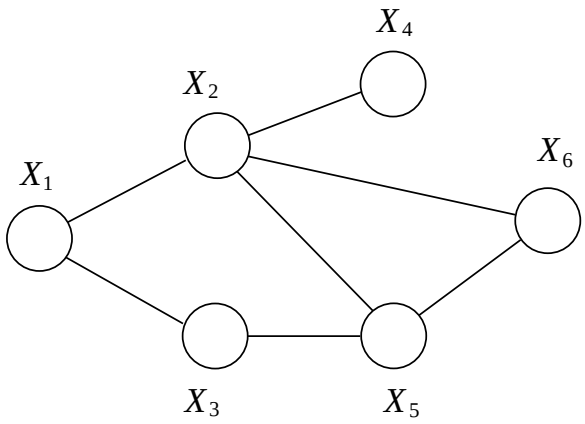
- A *plate* is a "macro" that allows subgraphs to be replicated:



- Graphical representation of an exchangeability assumption on $(X_1, X_2, \ldots)$

Undirected Graphical Models

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:

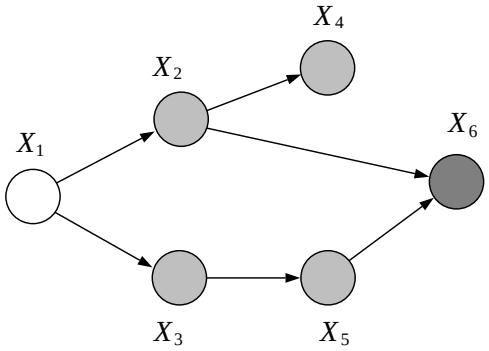


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the set of cliques defined by the edges $\mathcal{E}$:

$$p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = \frac{1}{Z}\, \psi(x_1, x_2; \theta_{12})\; \psi(x_1, x_3; \theta_{13})$$
$$\psi(x_2, x_4; \theta_{24})\; \psi(x_3, x_5; \theta_{35})\; \psi(x_2, x_5, x_6; \theta_{256})$$

Inference—Computing Conditional Probabilities

- Conditioning



- Marginalization:

$$p(x_1, x_6) = \int_{x_2} \int_{x_3} \int_{x_4} \int_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

- Conditional probabilities:

$$p(x_1 \,|\, x_6) = \frac{p(x_1, x_6)}{p(x_6)}$$

Inference Algorithms

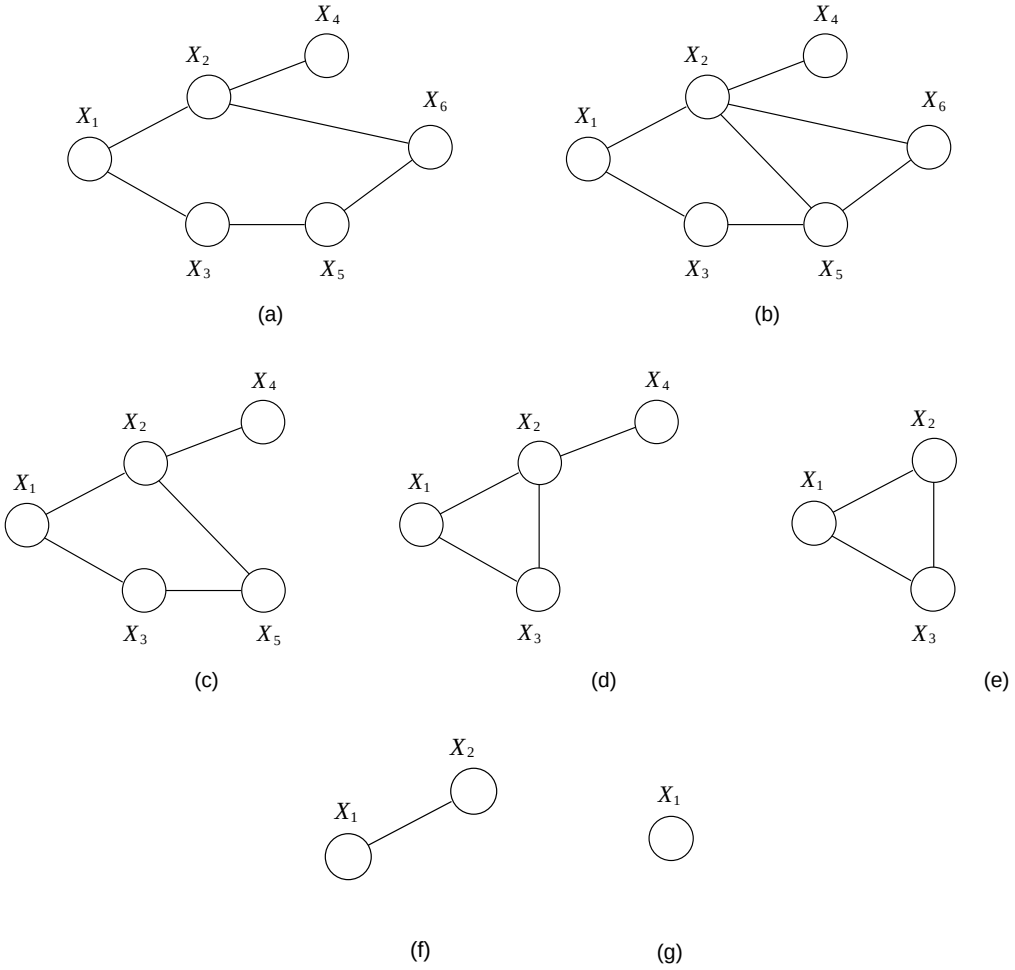
- *Exact algorithms*
  - elimination algorithm
  - sum-product algorithm
  - junction tree algorithm
- *Sampling algorithms*
  - importance sampling
  - Markov chain Monte Carlo (MCMC)
- *Variational algorithms*
  - mean field methods (e.g., Jordan et al., 1999)
  - sum-product algorithm and variations
    (e.g., Yedidia et al., 2001; Minka, 2001; McEliece & Yildirim, 2002)
  - semidefinite relaxations (Wainwright & Jordan, 2003)

Elimination Algorithm

$$
\begin{aligned}
p(x_1, x_6) &= \int_{x_2}\int_{x_3}\int_{x_4}\int_{x_5} p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5) \\
&= p(x_1)\int_{x_2} p(x_2 \mid x_1)\int_{x_3} p(x_3 \mid x_1)\int_{x_4} p(x_4 \mid x_2)\int_{x_5} p(x_5 \mid x_3)p(x_6 \mid x_2, x_5) \\
&= p(x_1)\int_{x_2} p(x_2 \mid x_1)\int_{x_3} p(x_3 \mid x_1)\int_{x_4} p(x_4 \mid x_2)\phi_{X_5}(x_2, x_3, x_6) \\
&= p(x_1)\int_{x_2} p(x_2 \mid x_1)\int_{x_3} p(x_3 \mid x_1)\phi_{X_5}(x_2, x_3, x_6)\int_{x_4} p(x_4 \mid x_2) \\
&= p(x_1)\int_{x_2} p(x_2 \mid x_1)\phi_{X_4}(x_2)\int_{x_3} p(x_3 \mid x_1)\phi_{X_5}(x_2, x_3, x_6) \\
&= p(x_1)\int_{x_2} p(x_2 \mid x_1)\phi_{X_4}(x_2)\phi_{X_3}(x_1, x_2, x_6) \\
&= p(x_1)\phi_{X_2}(x_1, x_6)
\end{aligned}
$$

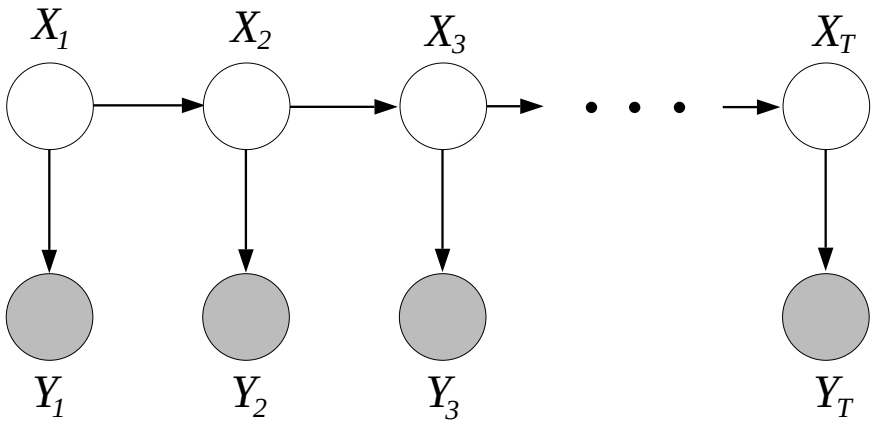
- The last term is proportional to the answer $p(x_1 \mid x_6)$
- The symbolic complexity can be characterized graph-theoretically

Graph Elimination



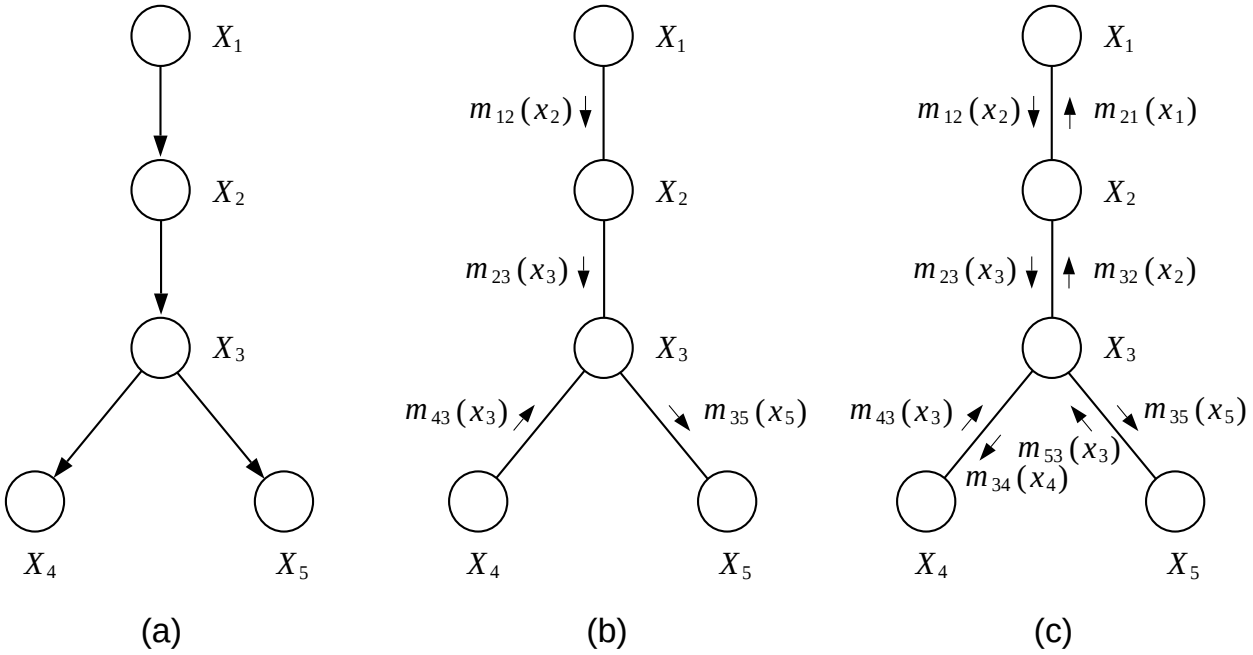
- Amounts to *triangulating* the graph
- Largest clique determines the computational complexity

Hidden Markov Models



- Generally wish to compute $p(x_i \mid y_1, y_2, \ldots, y_T)$, for all $i \in (1, \ldots, T)$
- Naive application of the Elimination Algorithm would require T different runs
- Want to cache and reuse the intermediate terms

Sum-product Algorithm



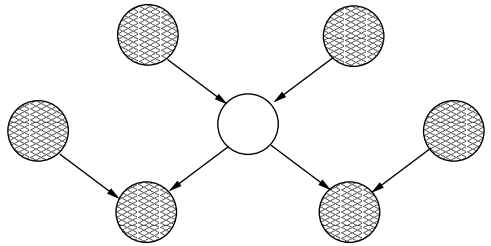
- Essentially the elimination algorithm along all possible paths
  - marginalization over a variable creates an intermediate term ("message")
  - messages are cached and reused
- The junction tree algorithm generalizes this to clique trees

Gibbs Sampling

- A widely-used Markov chain Monte Carlo (MCMC) method
- Given a set of variables X_V, we set up a Markov chain as follows:
  - initialize the X_i to arbitrary values
  - choose i randomly
  - sample from $p(x_i|x_{V\setminus i})$
  - iterate
- It is easy to prove that this scheme has $p(x_V)$ as its equilibrium distribution
- Gibbs sampling is often readily implemented in graphical models

Markov Blankets

- The *Markov blanket* of node X_i is the minimal set of nodes that renders X_i conditionally independent of all other nodes
- For undirected graphs, the Markov blanket is just the set of neighbors
- For directed graphs, the Markov blanket is the set of parents, children and co-parents:



- And the Gibbs conditional is a product of factors associated with these nodes:

$$p(x_i|x_{V\setminus i}) \propto p(x_i|x_{\pi_i}) \prod_{j \in \text{child}(i)} p(x_j|x_{\pi_j})$$

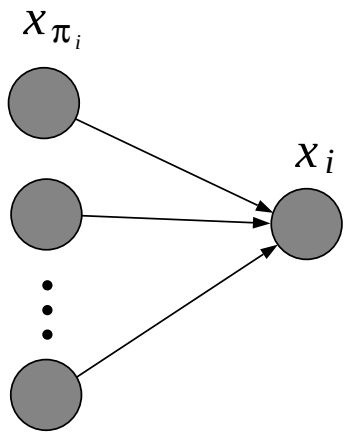
Variational Algorithms

- Three steps:
  - convert the inference problem into an optimization problem
  - relax the optimization problem into a simplified optimization problem
  - solve the relaxation

- Many variations
  - *mean field algorithms* (pretend the law of large numbers holds)
  - *sum-product algorithm* (pretend the graph is a tree)

Parameterization of Directed Models

- Local conditional probabilities, $p(x_i \mid x_{\pi_i}, \theta_i)$:



- E.g., generalized linear models (GLIMs)

- Joint probability distribution:

$$p(x \mid \theta) = \prod_i p(x_i \mid x_{\pi_i}; \theta_i)$$

Parameter Estimation

- *Maximum likelihood*

$$\theta^* = \text{argmax}_\theta \log\ p(x \mid \theta)$$

- Consider a data set in which each node is observed
  - The joint probability is a product, thus the log likelihood is a sum:

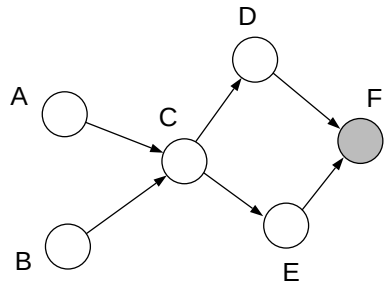
$$\begin{aligned} \log\ p(x|\theta) &= \log \prod_i p(x_i|x_{\pi_i}, \theta) \\ &= \sum_i \log\ p(x_i|x_{\pi_i}, \theta_i) \end{aligned}$$

  - I.e., the maximization problem decouples into a set of separate maximization problems

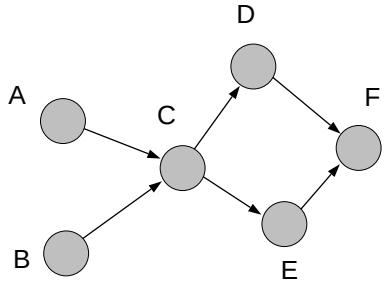
Parameter Estimation (cont.)

- When there are latent variables can use the Expectation-Maximization (EM) algorithm

- The E step of the EM algorithm involves calculating expectations of sufficient statistics associated with the latent variables
  - this is what the inference algorithms aim to provide

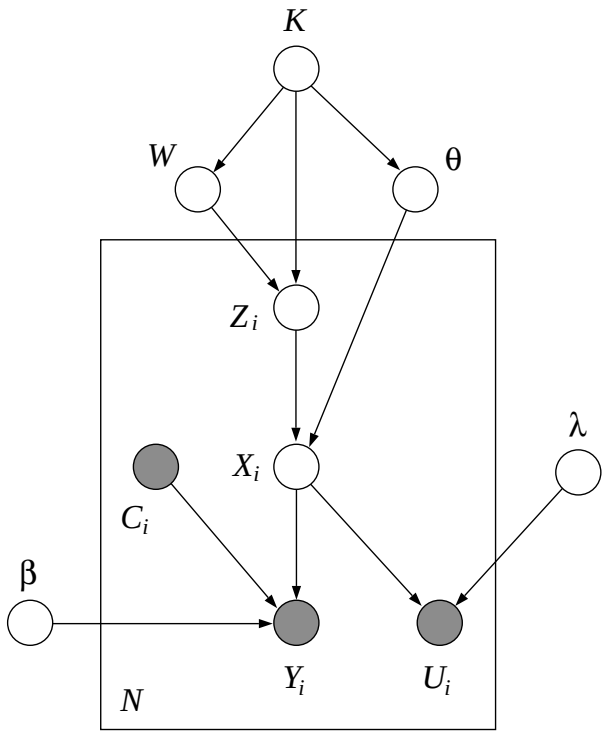


- The M step involves parameter estimation for a fully observed graph



Bayesian Methods

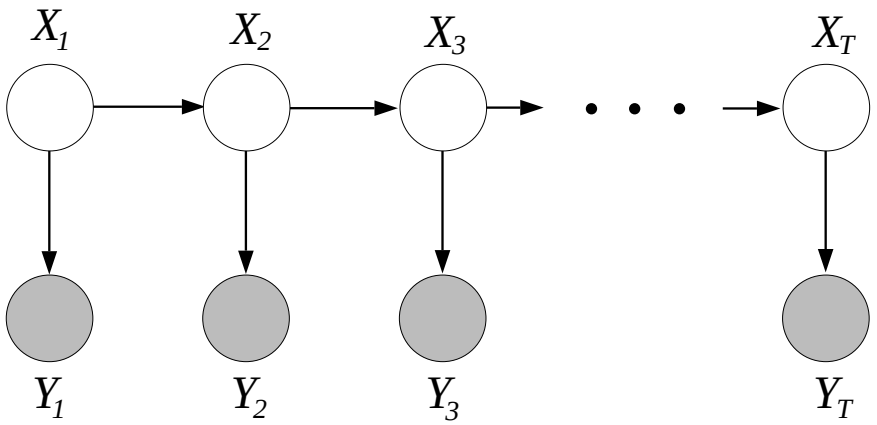
- Bayesian methods are based on posterior probabilities, marginal likelihoods, predictive probabilities, etc
  - this is what the inference algorithms aim to provide
- E.g., an error-in-covariates model:



Examples

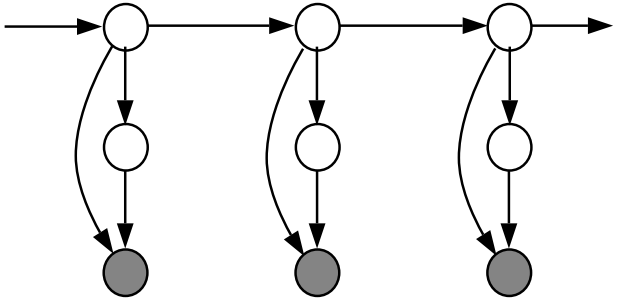
- Hidden Markov models
- Phylogenies
- Hidden Markov phylogenies
- Low-density parity check codes
- Medical diagnosis
- Latent Dirichlet allocation models

Hidden Markov Models

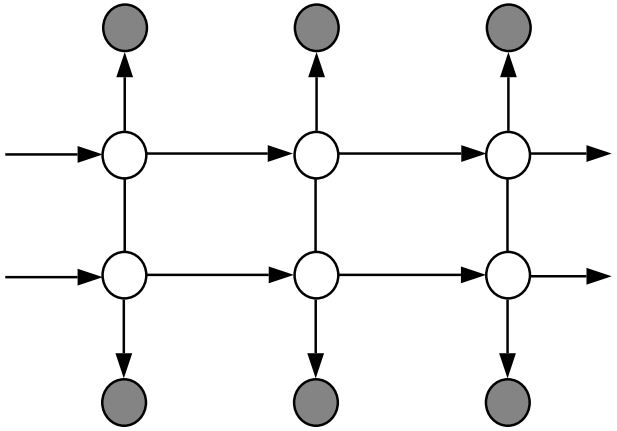


- Generally wish to compute $p(x_i \mid y_1, y_2, \ldots, y_T)$
- For continuous X_i, this is the Kalman filter/smoother
- Widely used in bioinformatics to represent segments of DNA or proteins
  - profile models of protein domains
  - gene finders and motif finders
  - models of secondary structure
  - models of transmembrane proteins

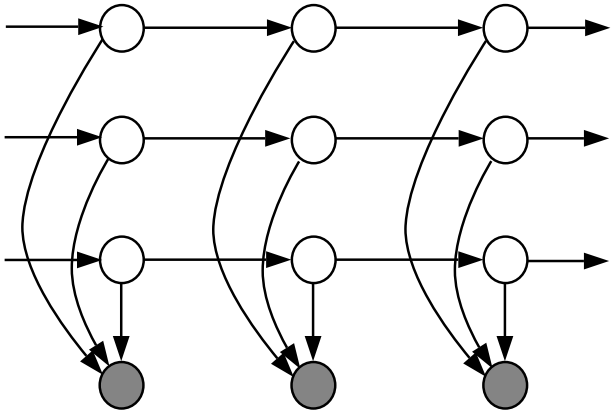
Hidden Markov Model Variations



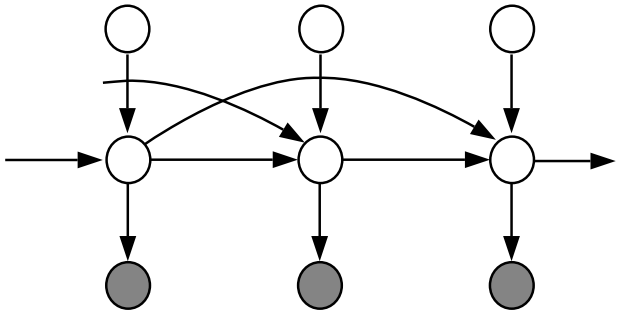
(a)



(b)

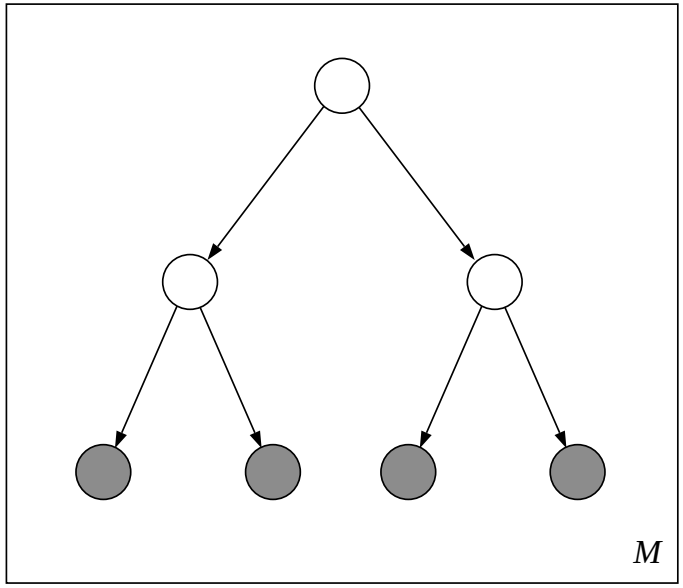


(c)



(d)

Phylogenies



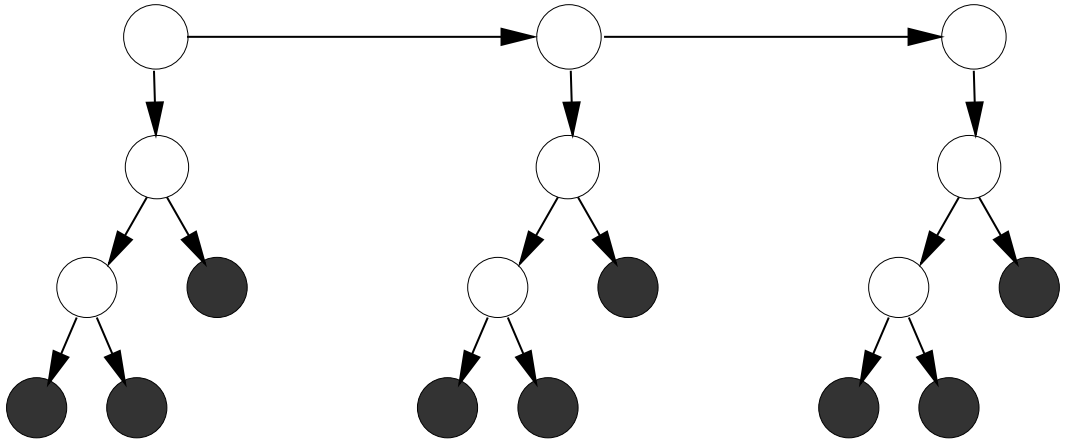
- The shaded nodes represent the observed nucleotides at a given site for a set of organisms
- Site independence model (note the plate)
- The unshaded nodes represent putative ancestral nucleotides
- Computing the likelihood involves summing over the unshaded nodes

Finding Genes in Genome Sequence

- Where do genes start and end? Where are the exon/intron boundaries within genes?
- Current gene finders are based on hidden Markov models
  - they have accuracies in the 30%-50% range
- Multiple species data is becoming available
  - how can we fuse data from multiple species to improve gene finding?

Hidden Markov Phylogeny

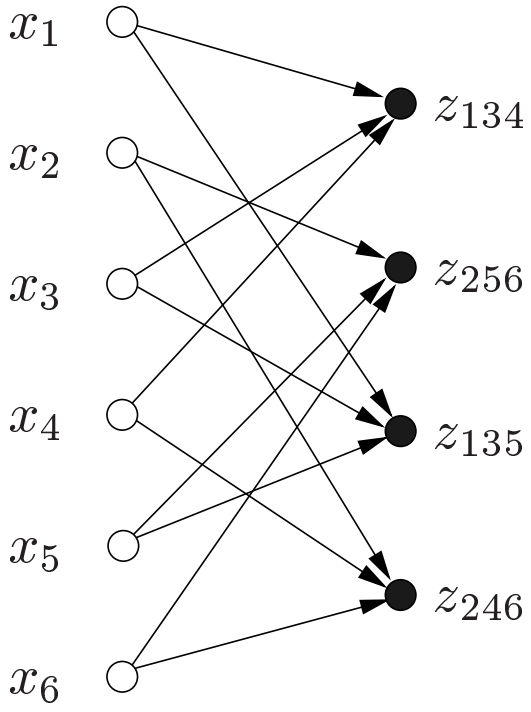
(McAuliffe, Pachter, & Jordan, 2003)



- This yields a gene finder that exploits evolutionary constraints
  - evolutionary rate is state-dependent
  - (edges from state to nodes in phylogeny are omitted for simplicity)

- Based on sequence data from 12-15 primate species, we obtain a nucleotide sensitivity of 100%, with a specificity of 89%
  - GENSCAN yields a sensitivity of 45%, with a specificity of 34%

Low-Density Parity Check Codes

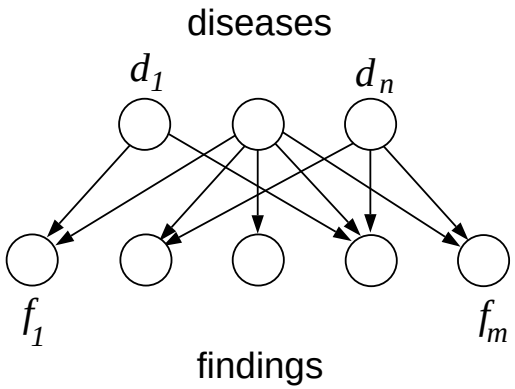


- The x_i denote the unknown message; the z_{ijk} denote the parity checks
- Compute the maximum a posteriori message
  - exact algorithms and MCMC algorithms are not viable
  - a variational algorithm ("max-product algorithm") is used instead, yielding impressive results

Quick Medical Reference (QMR) model

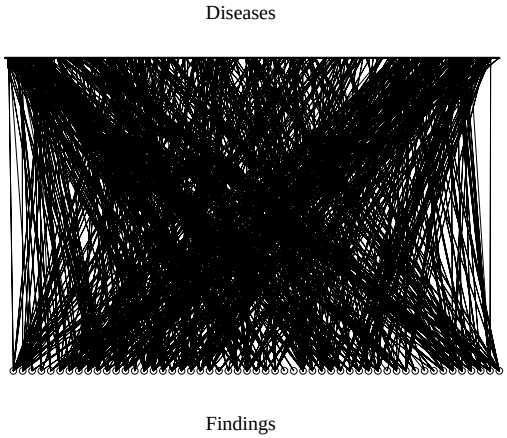
(Jaakkola & Jordan, 1999)

- A probabilistic graphical model for diagnosis with 600 *disease* nodes, 4000 *finding* nodes



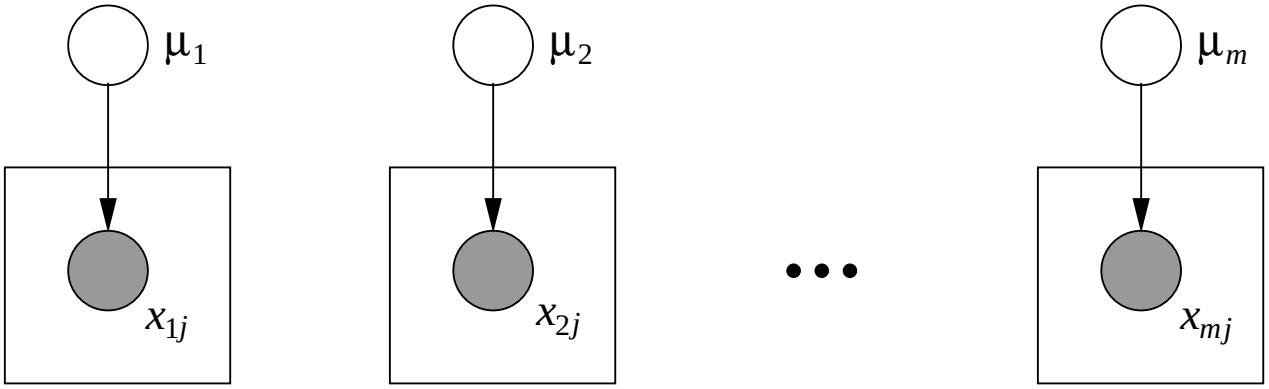
- Node probabilities $p(f_i|d)$ were assessed from an expert (Shwe, et al., 1991)
- Want to compute posteriors: $p(d_j|f)$
- Is this tractable?

Quick Medical Reference (cont.)



- Exact algorithms would take years to run
- MCMC algorithms take hours to run, and convergence is difficult to assess
- A mean field variational method due to Jaakkola and Jordan (1999) computes approximate posteriors in less than a second

Multiple Inference Problems



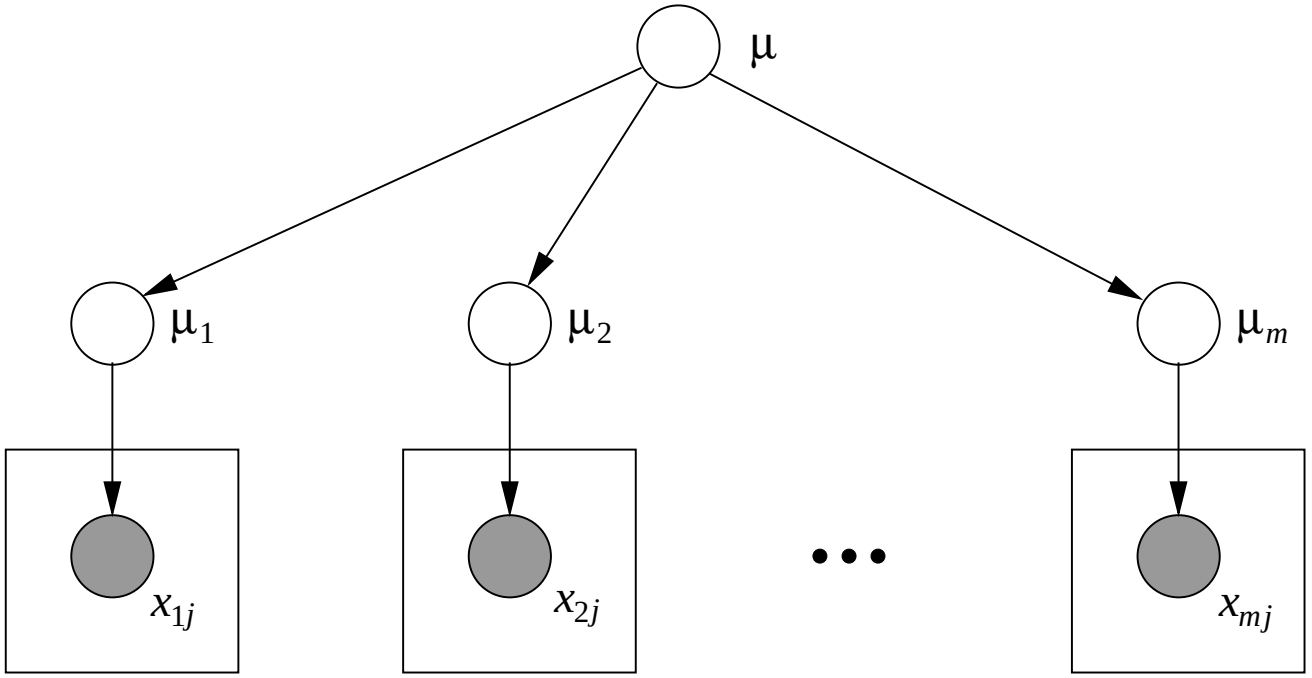
- Multiple Gaussian means (e.g., mean heights in various cities)

$$x_{ij} \sim N(\mu_i, \sigma_i^2)$$

- Maximum likelihood: $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

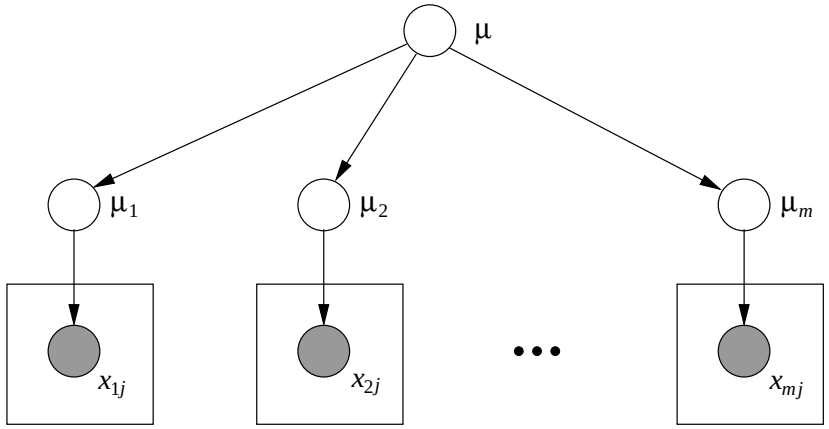
- Maximum likelihood often doesn't work very well
  - want to "share statistical strength" (i.e., "smooth")

Hierarchical Bayesian Modeling

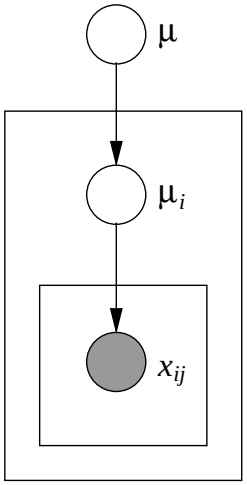


- Posterior mean is a shrinkage estimator

Hierarchical Modeling



- Recall the plate notation:



Probabilistic Modeling of Documents

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- *Goal*: a *joint probability distribution* over a corpus of such entities that can support activities of search, indexing, summarization, classification, text analysis, information extraction, etc

Classical *tf-idf* Approach

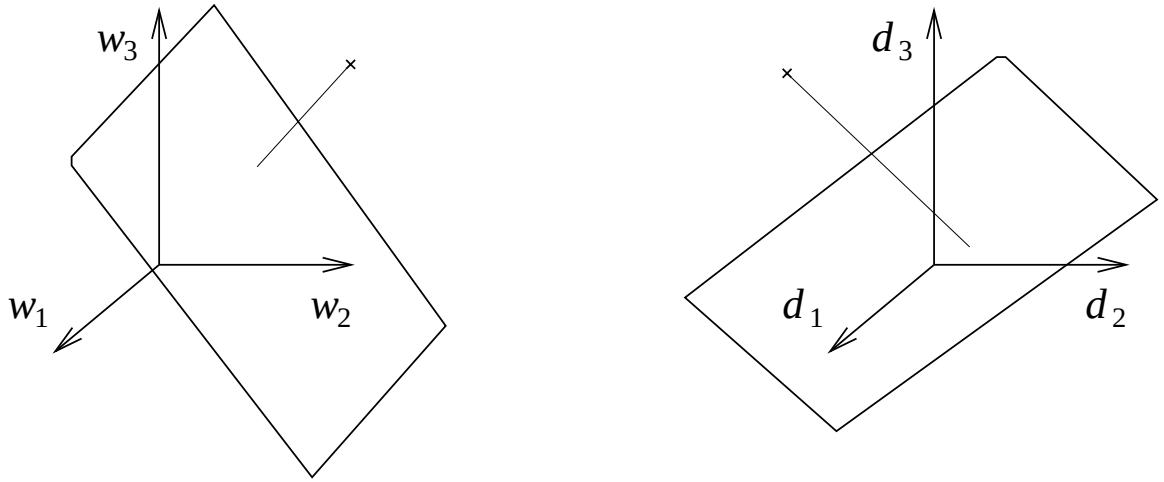
- Choose a vocabulary V of "words"
- Represent a "document" as a list of ratios of counts of word occurences:

$$t_i = \frac{f(\#\text{ of times word } i \text{ occurs in the document})}{g(\#\text{ of times word } i \text{ occurs across all documents})}$$

- Thoroughly adhoc, but somewhat of an empirical success story
- Still, limited data reduction and no structure revealed

Latent Semantic Indexing

- Arrange the *tf-idf* values into a word-by-document matrix X; do singular value decomposition of X; retain a small number of singular vectors



- Not clear what the "semantics" is here, nor why this geometry is appropriate

- Cf. probabilistic "duality"—*Bayes' rule*

  – a model in which we can compute $p(w \mid d)$ and $p(d \mid w)$

"Bag-of-words Assumption"

- Current methods in information retrieval are based on the "bag-of-words" assumption—word order within a document is ignored

- As a probabilistic assumption, this is **not** equivalent to "independent and identically distributed"

- Rather, it is equivalent to "exchangeable," which gives us a hint as to how to proceed:

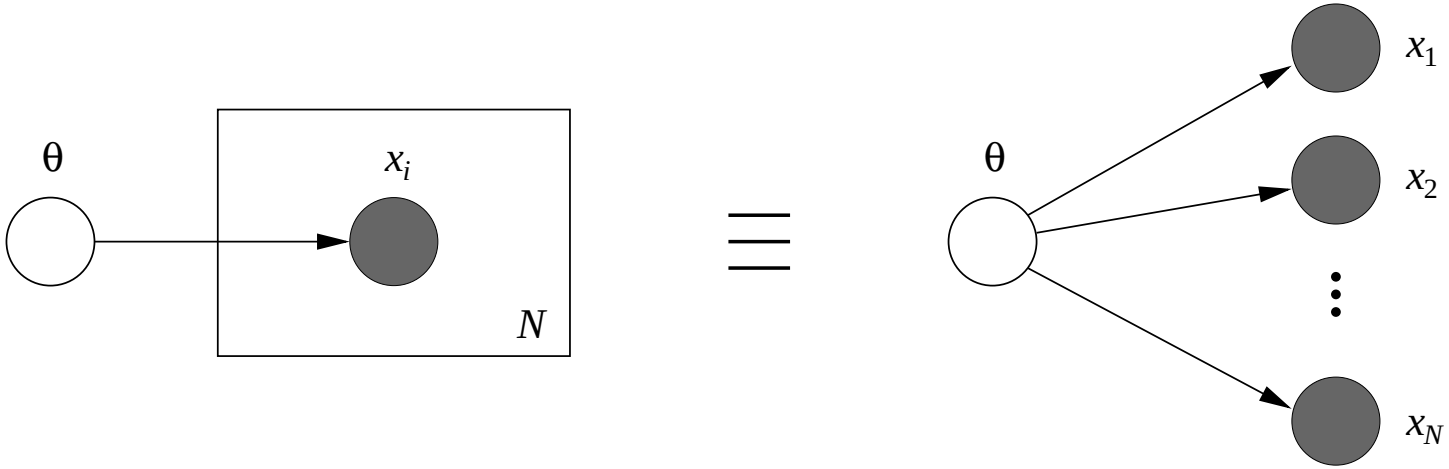
  Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture*:

$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) d\theta$$

  for some random variable θ.

"Bag-of-words Assumption (cont.)"

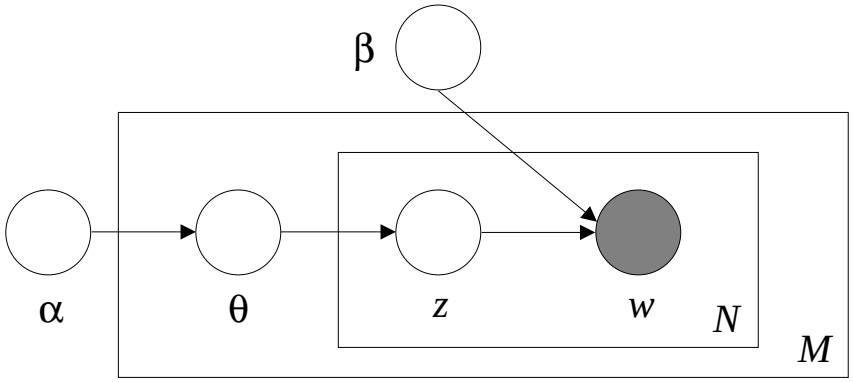
- A graphical representation of the De Finetti theorem:



$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left(\prod_{i=1}^{N} p(x_i \,|\, \theta) \right) d\theta$$

Latent Dirichlet Allocation (LDA) Model

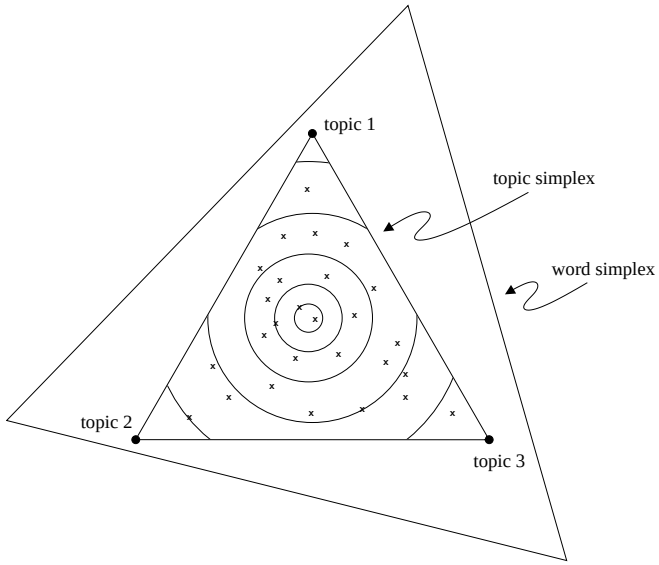
(Blei, Ng, & Jordan, 2003)



- *Random variables*:
  - A **word** is represented as a *multinomial* random variable w
  - A **topic** is represented as a *multinomial* random variable z
  - A **document** is represented as a *Dirichlet* random variable θ
- *Plates*:
  - *repeated* sampling of Dirichlet document variable within corpus
  - *repeated* sampling of multinomial topic variable within documents

The Topic Simplex

- Each corner of the simplex corresponds to a *topic*—a component of the vector z:



The topic simplex for $k = 3$.

- A document is modeled as a point in the simplex—a multinomial distribution over topics

- A corpus is modeled as a Dirichlet distribution on the simplex

Example: Nematode Abstracts

- A database of abstracts from articles on nematode biology
- Four of the resulting topics:

"Signaling"	"Genetics"	"Reproduction"	"Proteomics"
RECEPTOR	CHROMOSOME	MALE	ELEGANS
RESPONSE	RECOMBINATION	SEX	ACTIVITY
ELEGANS	MEIOTIC	SPERM	BINDING
ACETYLCHOLINE	ELEGANS	HERMAPHRODITES	NEMATODE
HABITUATION	DEFICIENCIES	TRA	PROTEIN
RESPONSES	CAENORHABDITIS	FEM	ELT
SIGNALING	DUPLICATIONS	ELEGANS	PURIFIED
RELEASE	LEFT	ANIMALS	KDA
LAG	LINKAGE	GENES	AFFINITY
GLUTAMATE	MAP	DETERMINATION	ENZYME

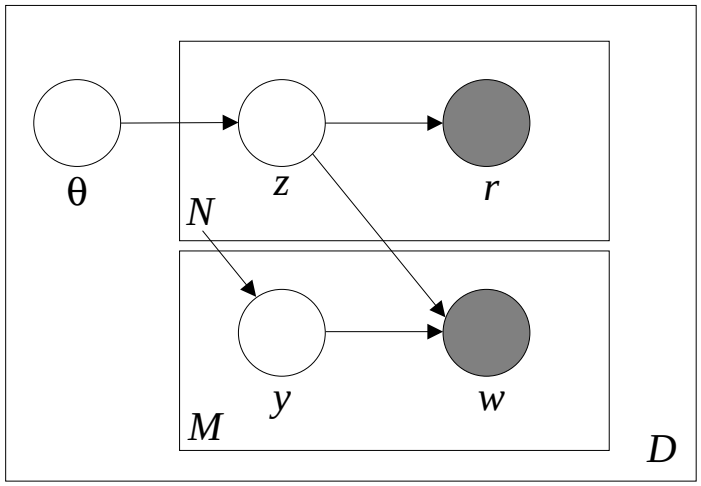
Probabilistic Modeling of Documents/Images

Sculpture, Statue, Stone

- *Goal*: a *joint probability distribution* that can support activities of search, indexing, text/image analysis, information extraction, etc
- Data are 11,000 images and their captions
- Images are segmented into regions, and each region is represented as a 47-dimensional Gaussian vector

Correspondence LDA Model

(Blei & Jordan, 2003)



- Image-topics and word-topics
  - a word is represented as a *multinomial* random variable w
  - an image region is represented as a *Gaussian* random variable r
  - a word-topic is represented as a *multinomial* random variable z
  - an image-topic is represented as a *multinomial* random variable y

- A mean field variational algorithm is used for inference

Automatic Annotation

True caption
market people

Corr−LDA
people market pattern textile display

GM−LDA
people tree light sky water

GM−Mixture
people market street costume temple

True caption
scotland water

Corr−LDA
scotland water flowers hills tree

GM−LDA
tree water people mountain sky

GM−Mixture
water sky clouds sunset scotland

(Use the top five words from $p(w|\mathbf{r})$ to annotate an image.)

Automatic Annotation

True caption
birds tree

Corr−LDA
birds nest leaves branch tree

GM−LDA
water birds nest tree sky

GM−Mixture
tree ocean fungus mushrooms coral

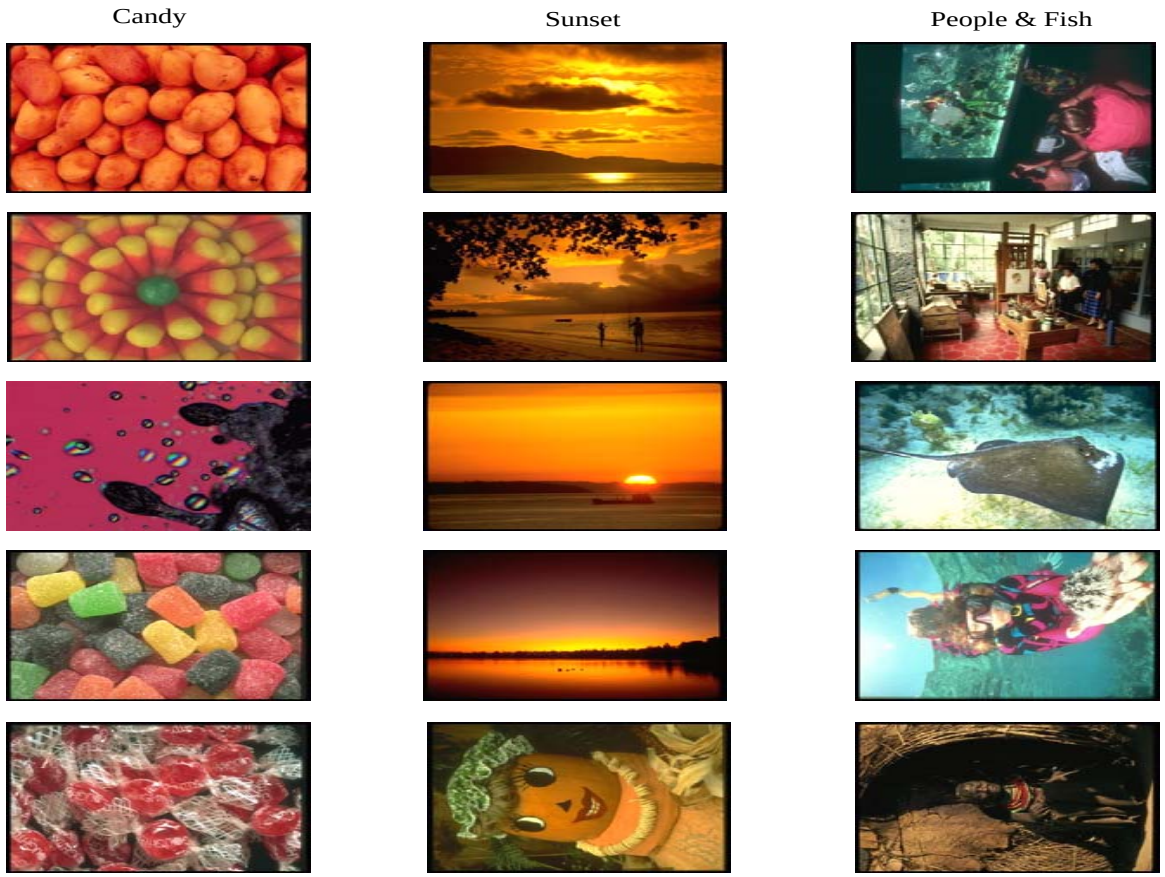
True caption
fish reefs water

Corr−LDA
fish water ocean tree coral

GM−LDA
water sky vegetables tree people

GM−Mixture
fungus mushrooms tree flowers leaves

Text-based Image Retrieval

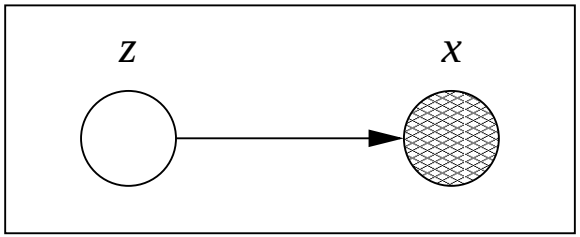


1. Compute $p(\mathbf{w}|\mathbf{r}_g)$ for each image in the test set.

2. Rank the images in order of conditional likelihood.

Mixture Models

- A probabilistic model for clustering



- The latent multinomial variable Z represents the clusters:

$$\begin{aligned} p(x|\theta) &= \sum_k p(Z=k)p(x|Z=k,\theta) \\ &= \sum_k \pi_k f_k(x|\theta_k), \end{aligned}$$

where π_k are the *mixing proportions* and $f_i(x|\theta_i)$ are the *mixture components* (e.g., Gaussians, where $\theta_i = (\mu_i, \Sigma_i)$)

Model Selection for Mixture Models

- Probabilistic model for clustering:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k),$$

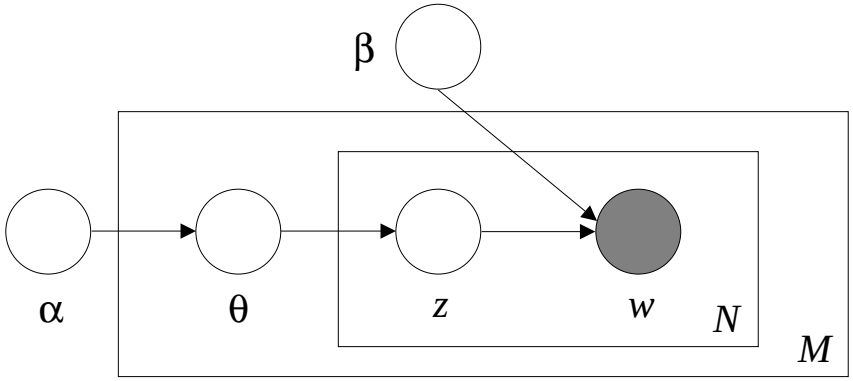
- How to choose K, the number of mixture components?

- Similarly, for the HMM, how to choose the number of states?

Clustering

- Many adhoc approaches (e.g., hierarchical clustering)
  - little theoretical understanding
  - hard to use as a building block
- Kernel methods
  - spectral clustering
  - how to choose K?
- Probabilistic graphical models
  - generalized mixture models
  - various methods for choosing K: cross-validation, bootstrap, AIC, BIC, Laplace, reversible jump, etc

Model Selection Problem for LDA



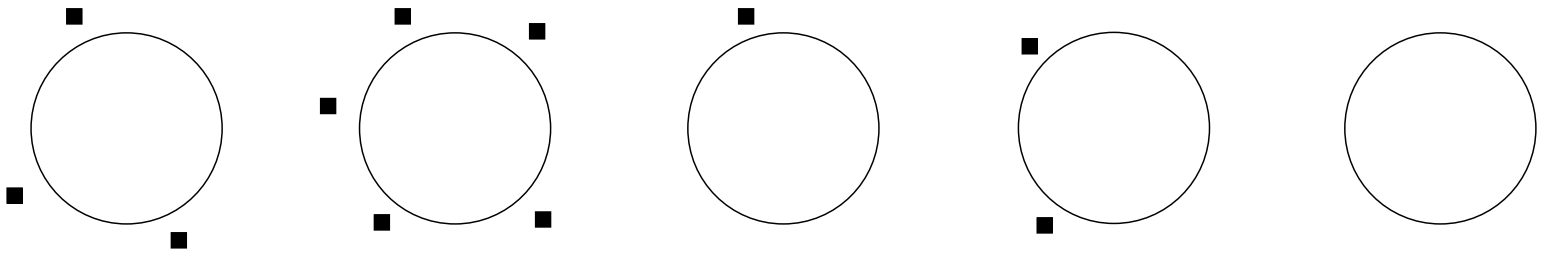
- How to choose the number of topics (the cardinality of z)?

Chinese Restaurant Process (CRP)

- A process in which n customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - mth subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} p(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto m_i \\ p(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \tag{1}$$

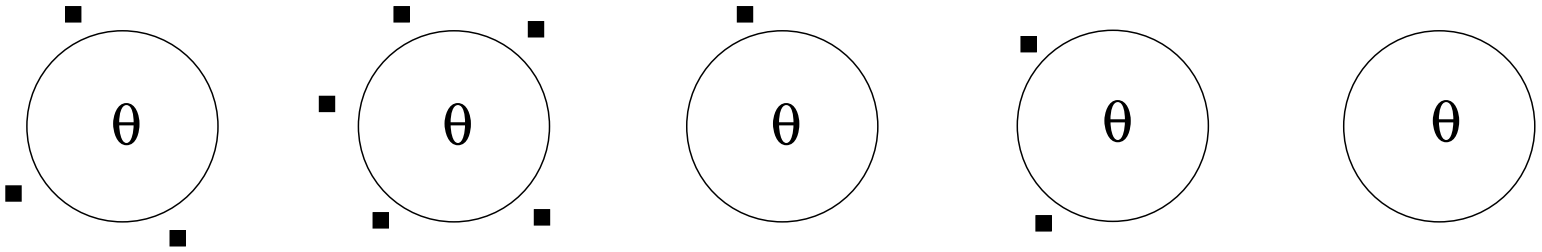
  where m_i is the number of customers currently at table i



- Defines an exchangeable distribution on partitions of customers

The CRP and Mixture Models

- Associate a mixture component with each table
  - the first customer to sit at a table draws a parameter from the prior



- This defines a prior on number of clusters and on the parameters associated with each cluster
- The likelihood is the usual mixture likelihood, but with an infinite number of mixture components
- Posterior inference can be performed via (e.g.) a Gibbs sampler

Gibbs Sampling

- For each data point:
  - pretend that it is the last point (by exchangeability)
  - choose a table using the Chinese restaurant dynamics
- For each table:
  - resample the parameter vector at that table, conditioning on all of the data points sitting at the table
- This will converge to a posterior distribution on partitions and parameters

Example

NIPS Data

- 1718 abstracts from NIPS 0 – 12

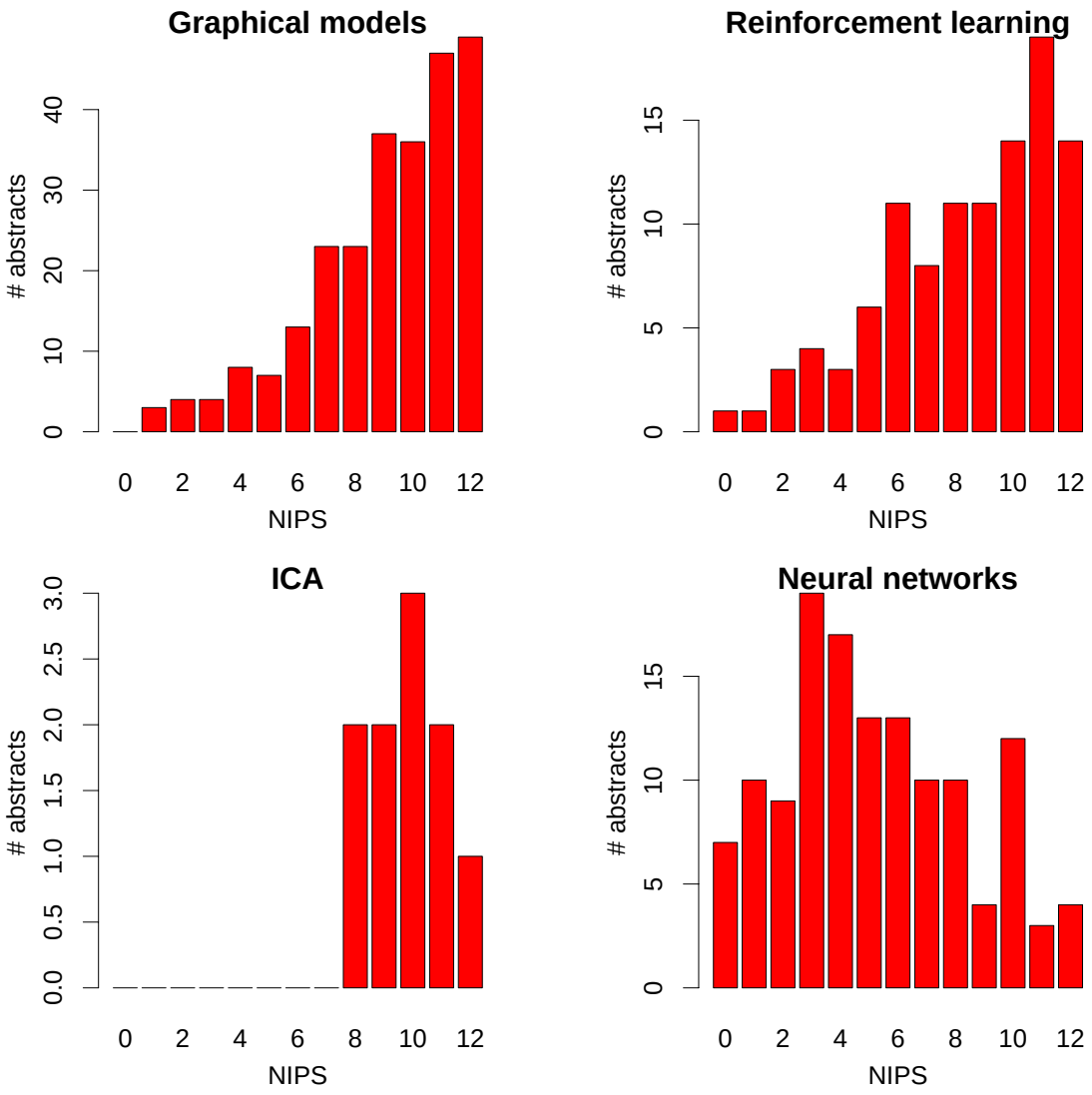
> Increasing attention has recently been paid to algorithms based on dynamic programming (DP) due to the suitability of DP for learning problems involving control. In stochastic environments where the system being controlled is only incompletely known, however, a unifying theoretical account of these methods has been missing. In this paper we relate DP-based learning algorithms to the powerful techniques of stochastic approximation via a new convergence theorem, enabling us to establish a class of convergent algorithms to which both TD(λ) and Q-learning belong.

- Each cluster is a 4100-dimensional multinomial distribution

- The posterior mode was 26 clusters

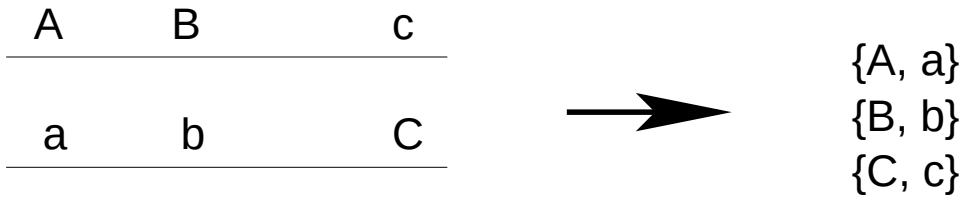
"GM"	"RL"	"NN"	"ICA"
data	learning	network	eeg
model	reinforcement	networks	ica
algorithm	control	learning	channel
learning	algorithm	neural	data
models	function	paper	signals
problem	policy	time	source
networks	problem	training	artifacts
show	optimal	recurrent	independent
method	paper	input	changes
approach	state	method	results
based	problems	architecture	problem
paper	value	structure	components
new	algorithms	rules	time
results	methods	units	analysis
bayesian	model	problem	cell

Visualizing Trends



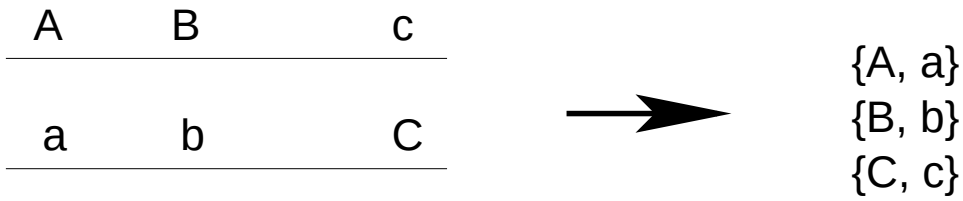
Haplotype Modeling

- Consider M binary markers in a genomic region
- There are 2^M possible *haplotypes*—i.e., states of a single chromosome
  - but in fact, far fewer are seen in human populations
- Given a sample of *genotypes* (unordered sets of pairs of markers)



  - estimate the underlying haplotypes
- This is a clustering problem

Haplotype Modeling (cont.)



- The genotype is a mixture over the population haplotypes:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1)p(h_2)p(g \,|\, h_1, h_2),$$

  (assuming Hardy-Weinberg equilibrium)

- So this is naturally treated as a mixture modeling problem

- What is the cardinality of $\mathcal{H}$?

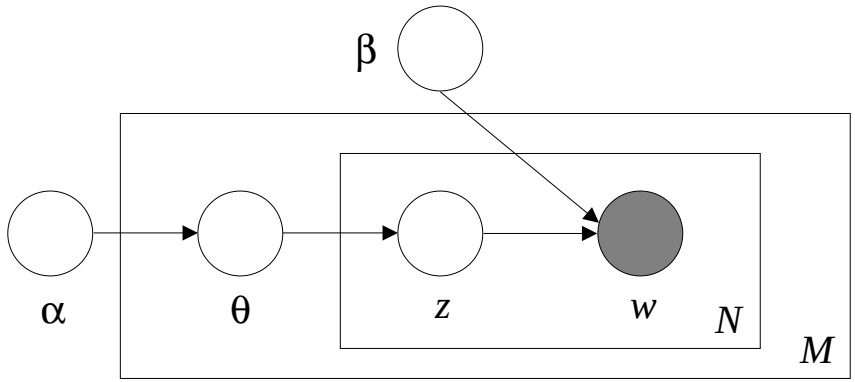
CRP-based Haplotype Model

(Xing, Sharan, & Jordan, 2004)

- Comparative performance of model on the data of Gabriel, et al (2002):

		DP			PHASE		
region	length	err_s	err_i	d_s	err_s	err_i	d_s
16a	13	0.185	0.480	0.141	0.174	0.440	0.130
1b	16	0.100	0.250	0.160	0.200	0.450	0.180
25a	14	0.135	0.353	0.115	0.212	0.588	0.212
7b	13	0.105	0.278	0.066	0.145	0.444	0.092

Model Selection Problem for LDA (cont.)



- Have we solved the problem of choosing the number of topics for LDA?
  - unfortunately, no
- We have *multiple* clustering problems—one per document
- We need to find a way to link multiple clustering problems

Haplotype Modeling (cont.)

- Suppose that we stratify the populations by ethnic group (e.g., African, Asian, European)

- Interesting to try to discover what haplotypes they have in common
  - thus we have multiple, linked clustering problems
  - need to choose the number of clusters in each group, and want to share clusters among groups?

- How to do this?

Dirichlet Distribution

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ be a point in the $(m-1)$simplex

  – i.e., $0 < \theta_i < 1$ and $\sum_{i=1}^{m} \theta_i = 1$

- Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a set of parameters, where $\alpha_i > 0$

- The Dirichlet distribution:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_m^{\alpha_m - 1}$$

  is an exponential family distribution on the simplex, with $\mathbb{E}(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i}$

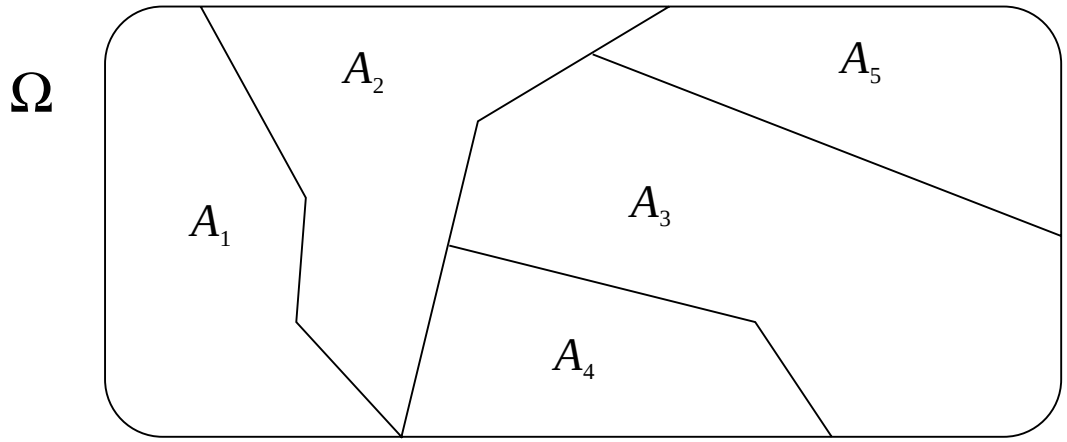
- Often used in a Bayesian setting, where θ is a parameter for a multinomial

Dirichlet Process

Definition 1. *Let $(\Omega, \mathcal{B})$ by a measurable space, with G_0 a probability measure on the space, and let α_0 be a positive real number. A* Dirichlet process *is the distribution of a random probability measure G over $(\Omega, \mathcal{B})$ such that, for any finite partition $(A_1, \ldots, A_r)$ of Ω, the random vector $(G(A_1), \ldots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution:*

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}\big(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)\big) \tag{2}$$

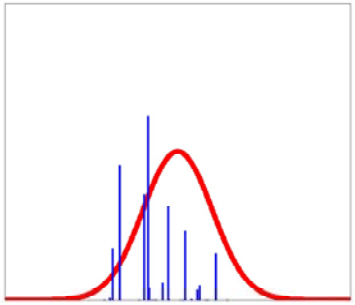
We write $G \sim \mathrm{DP}(\alpha_0, G_0)$ if G is a random probability measure distributed according to the Dirichlet process. Call G_0 the base measure of G and call α_0 the concentration parameter.



Stick-Breaking Representation

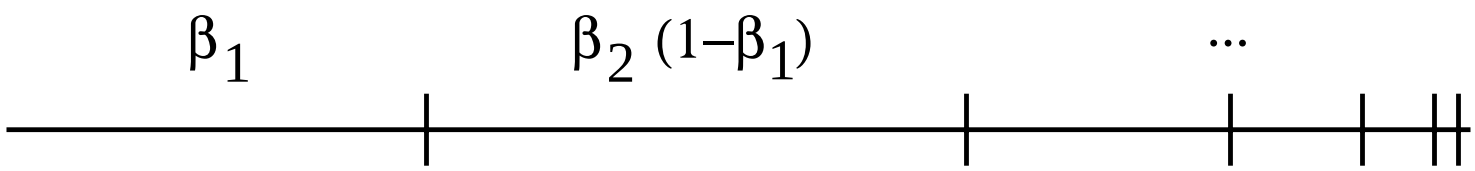
- Sethuraman (1994) gave an explicit representation for a draw from a Dirichlet process:

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

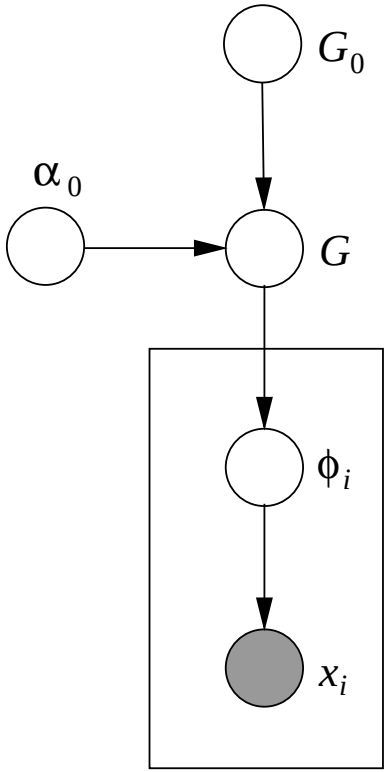


where

$$\theta_k \sim G_0 \qquad \beta_k \sim \text{Beta}(1, \alpha_0) \qquad \pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_k)$$

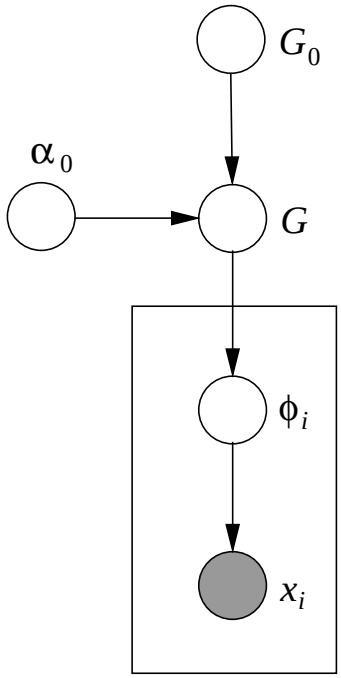


Dirichlet Process Mixture Models



$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha_0, G_0) \\
\phi_i \,|\, G &\sim G & i \in 1, \ldots, n \\
x_i \,|\, \phi_i &\sim F(x_i \,|\, \phi_i) & i \in 1, \ldots, n
\end{aligned}
$$

Integrating Out G



- Integrating out G yields a joint distribution on $\phi_1, \phi_2, \ldots, \phi_N$
- This distribution is precisely that given by the Chinese restaurant process!

$$\phi_i \,|\, \phi_1, \ldots, \phi_{i-1}, \alpha_0, G_0 \;\sim\; \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0}\delta_{\phi_l} + \frac{\alpha_0}{i-1+\alpha_0}G_0$$

Inference for Dirichlet Process Mixtures

- Gibbs sampling
  - based on the Chinese restaurant process
  - based on the stick-breaking representation
- Variational inference
  - based on the stick-breaking representation

Variational inference

(Blei, & Jordan, 2005)

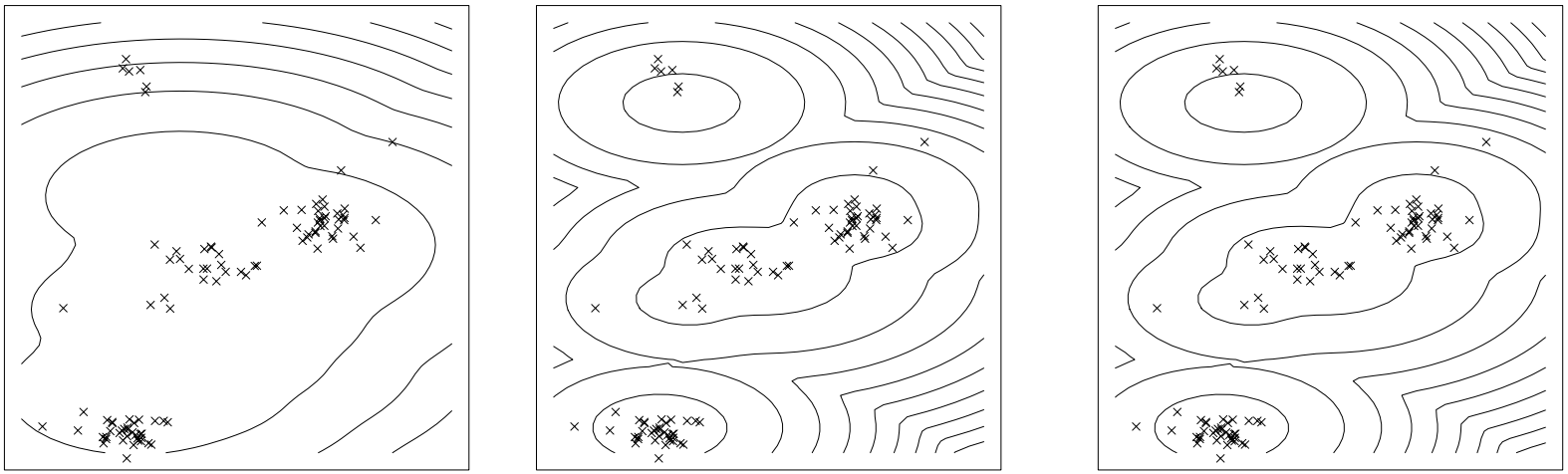
- Variational inference equations for a conjugate DP mixture in the exponential family

$$\begin{aligned}
\gamma_{i,1} &= 1 + \textstyle\sum_n \phi_{n,i} \\
\gamma_{i,2} &= \alpha + \textstyle\sum_n \sum_{j=i+1}^{K} \phi_{n,j} \\
\tau_{i,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,i} x_n \\
\tau_{i,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,i} \\
\phi_{n,i} &\propto \exp(S),
\end{aligned}$$

where

$$S = E[\log V_i \mid \gamma_i] + E[\eta_i \mid \tau_i]^T X_n - E[a(\eta_i) \mid \tau_i] - \textstyle\sum_{j=i+1}^{K} E[\log(1 - V_j) \mid \gamma_j].$$

Example: DP-Gaussian Mixture



Initial state 1st iteration 5th (and last) iteration

Figure 1. The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

Example: DP-Gaussian Mixture

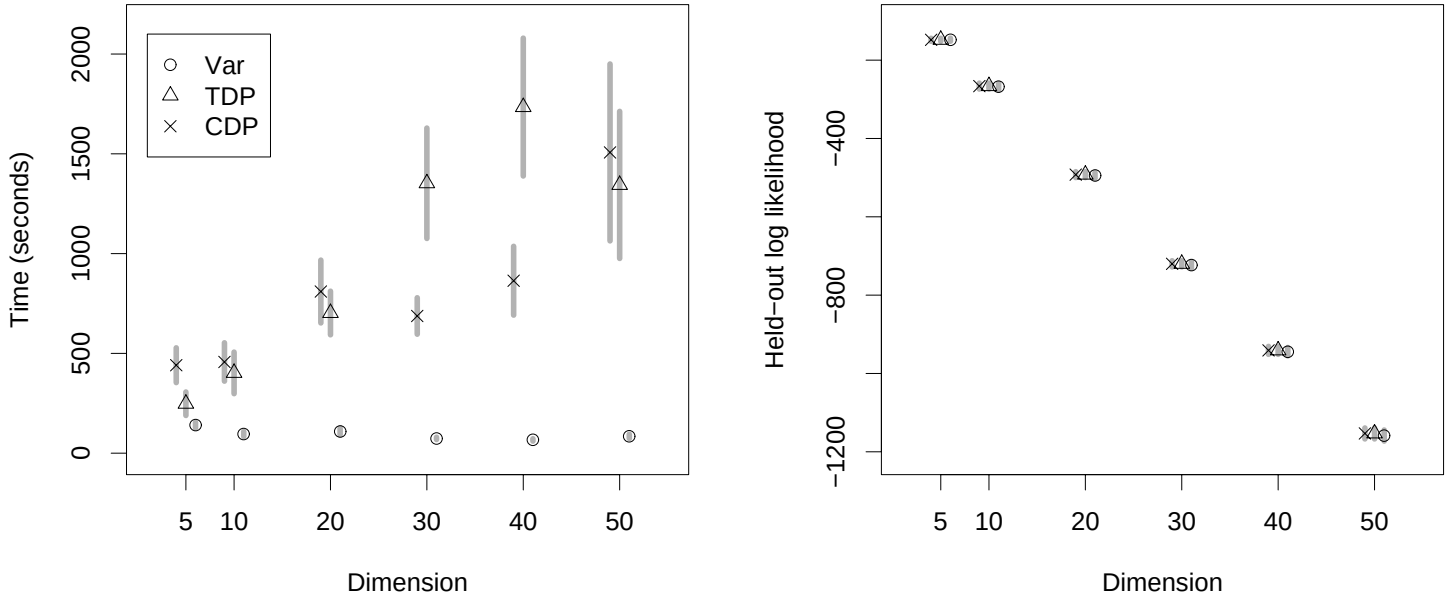
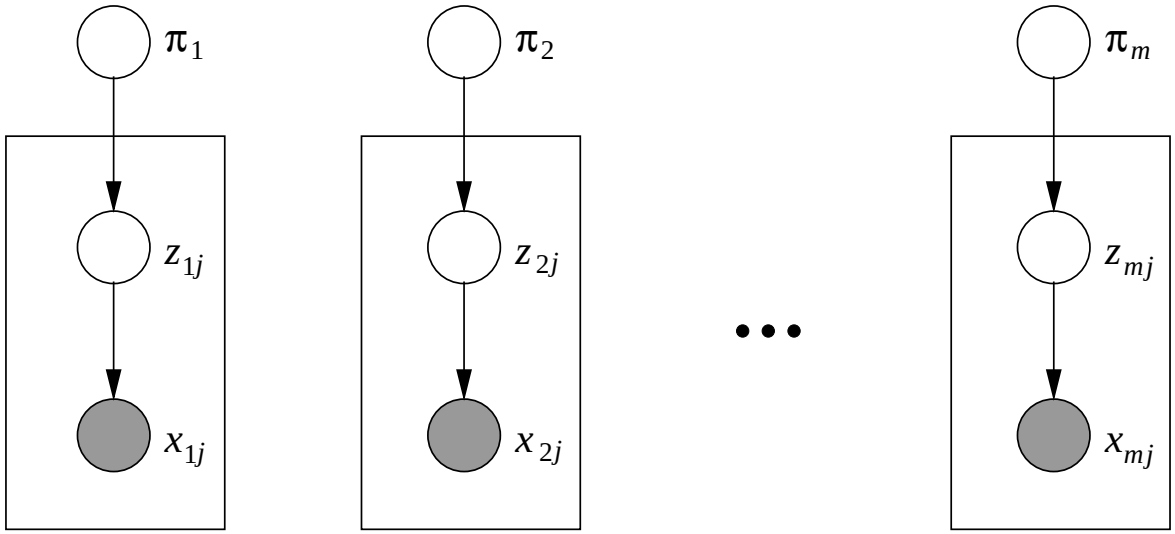


Figure 2. (Left) Convergence time per dimension across ten datasets for variational inference (Var), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

Multiple Clustering Problems

- Suppose that we have not one clustering problem, but a set of clustering problems
- Suppose that we view these problems as linked; in particular, we want to share clusters among groups
  - recall the haplotype problem with sub-populations
  - recall the LDA problem

Multiple Clustering Problems



- Mixture models with explicit allocations z_{ij}
- For each i: $p(x_{ij}|\pi_i, \theta_i, k_i) = \sum_{l=1}^{k_i} p(z_{ij} = l|\pi_i) p(x_{ij}|z_{ij} = l, \theta_i)$
- What to share: π_i?, θ_i?, k_i?
- How to share?

Multiple Clustering Problems (cont.)

- Idea: associate a Dirichlet process with each group
- How to link them?
  - let them share the same underlying G_0?
- Problem: the atoms generated by these processes will be distinct with probability one
  - i.e., there will be sharing of statistical strength, but no sharing of clusters!
- Need to have the base measure G_0 be discrete
  - but also need it to be flexible and random

Hierarchical Dirichlet Process

(Teh, Jordan, Beal & Blei, 2004)

- Let G_0 itself be distributed according to a DP:

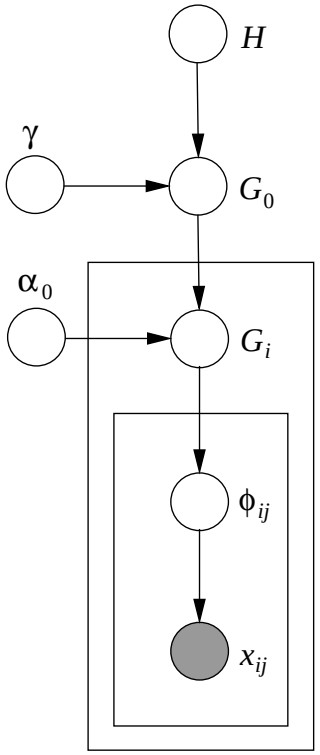
$$G_0 \,|\, \gamma, H \;\sim\; \mathrm{DP}(\gamma, H)$$

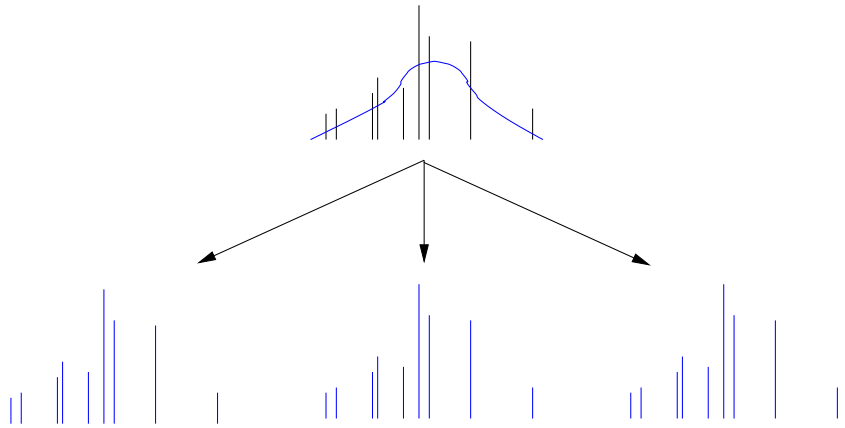
- Then

$$G \,|\, \alpha, G_0 \;\sim\; \mathrm{DP}(\alpha_0, G_0)$$

  has as its base measure an atomic distribution—samples of G will resample from these atoms

Hierarchical Dirichlet Process Mixture





$$
\begin{aligned}
G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\
G_i \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) \\
\phi_{ij} \,|\, G_i &\sim G_i \\
x_{ij} \,|\, \phi_{ij} &\sim F(x_{ij}, \phi_{ij})
\end{aligned}
$$

Stick-Breaking Representation

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}$$

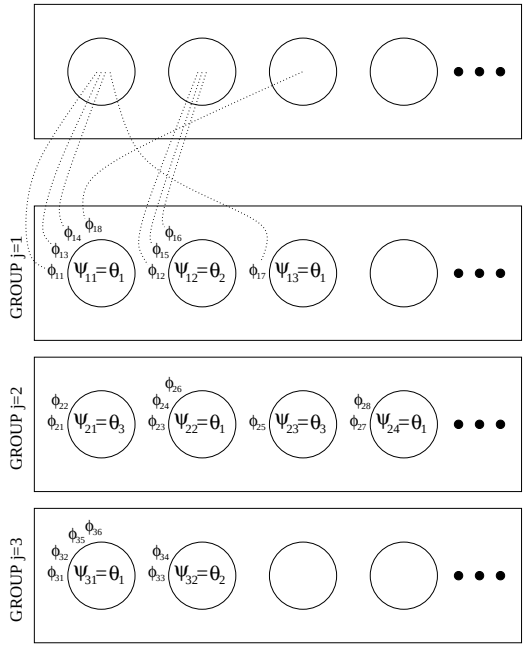
$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}) \qquad \pi'_{jk} \sim \text{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l\right)\right)$$

Inference for HDP mixtures

- Gibbs sampling
  - based on the Chinese restaurant process
  - based on the stick-breaking representation
- Variational inference
  - based on the stick-breaking representation

Chinese restaurant franchise (CRF)

- First integrate out the G_i, then integrate out G_0



- Set of *restaurants* with an unbounded number of *tables* in each restaurant
- One *menu* with an unbounded number of *dishes* on the menu
- Reinforcement effects—customers prefer to sit at tables with many other customers, and prefer to choose dishes that are chosen by many other customers

Gibbs sampler based on the CRF

$$p(t_{ji} = t \mid \boldsymbol{t} \backslash t_{ji}, \boldsymbol{k}, \boldsymbol{\theta}, \mathbf{x}) \quad \propto \quad \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt}}) & \text{if } t = t^{\text{new}} \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & \text{if } t \text{ currently used} \end{cases}$$

$$p(k_{jt} = k \mid \boldsymbol{t}, \boldsymbol{k} \backslash k_{jt}, \boldsymbol{\theta}, \mathbf{x}) \quad \propto \quad \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k = k^{\text{new}} \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k \text{ currently used} \end{cases}$$

$$p(\theta_k \mid \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\theta} \backslash \theta_k, \mathbf{x}) \quad \propto \quad h(\theta_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\theta_k)$$

Results: Nematode database

- data set of 5838 abstracts on nematode biology
- removed standard stop words and words appearing less than 10 times: yields a vocabulary size of 5699, with 476441 words in total
- the basic model is a multinomial mixture, one per abstract

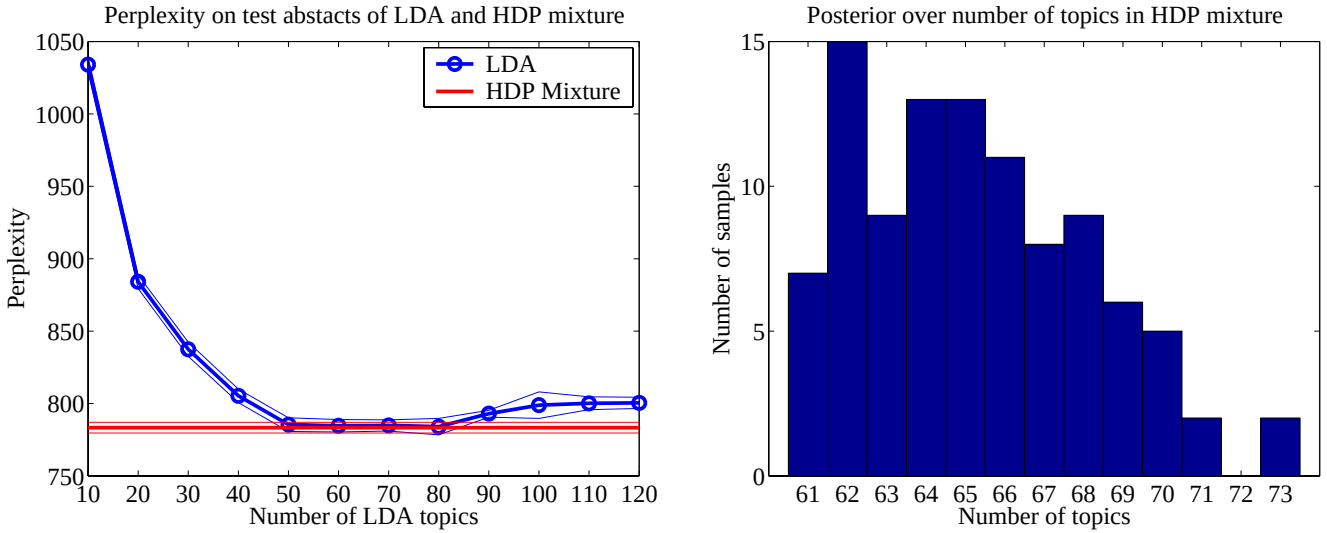


Figure 3. Left: perplexity on test abstracts of LDA and HDP mixture. Right: histogram of the number of topics over 100 posterior samples.

NIPS Conference articles (1988-2001)

- articles from the conference are divided into sections:

AA	algorithms and architectures
AP	applications
CS	cognitive science
CN	control and navigation
IM	implementations
NS	neuroscience
SP	signal processing
LT	learning theory
VS	vision

- each article is represented as a mixture model (over words in the vocabulary)
- an HDP is used to discover and share clusters ("topics") among articles within each section
- want to examine relationships among the sections

Models

- Each article is a DP mixture model
- Each section is a collection of mixture models—thus a section is modeled via an HDP mixture
- We have multiple sections
  - thus we require another level of the hierarchy to link the section HDPs—easily done
- Models:

"none"	a separate HDP for each section
"flat"	a single HDP for all sections
"hier"	a linked set of HDPs

- In presenting the results, we focus on one section (VS) and consider one other section at a time

Results

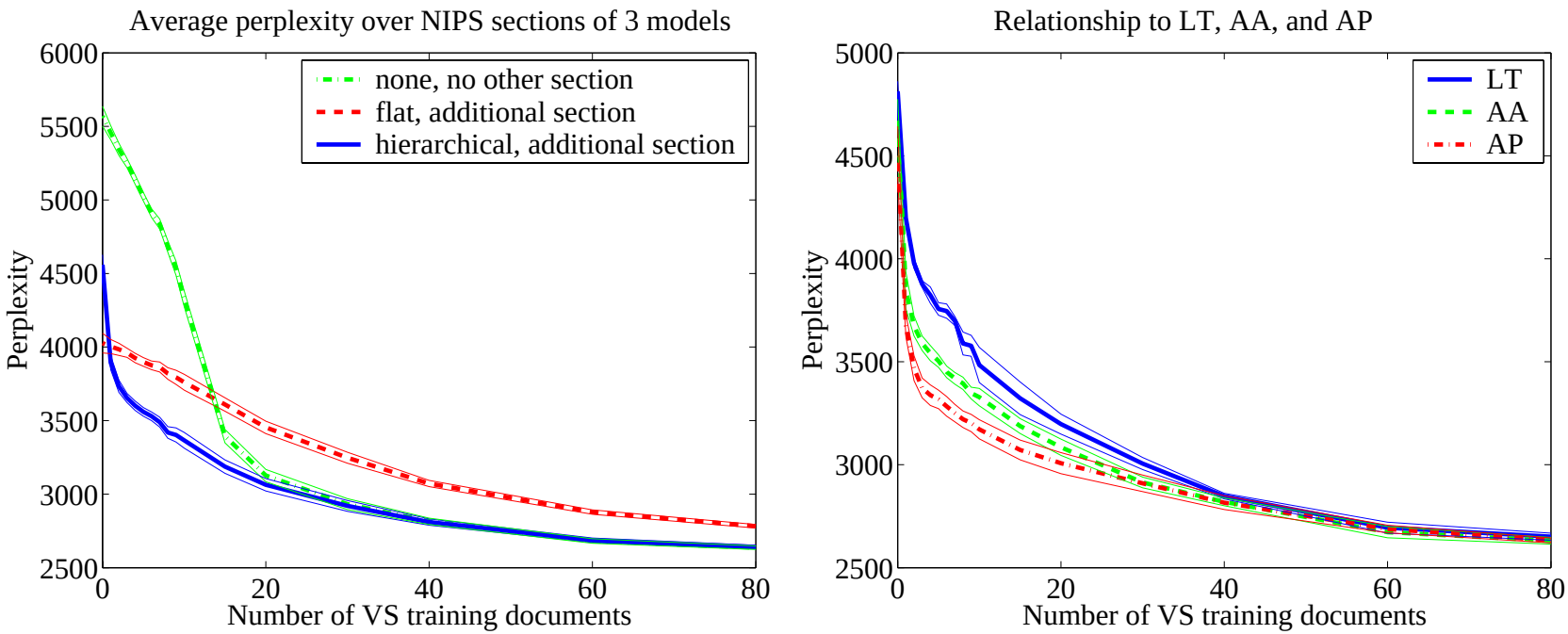


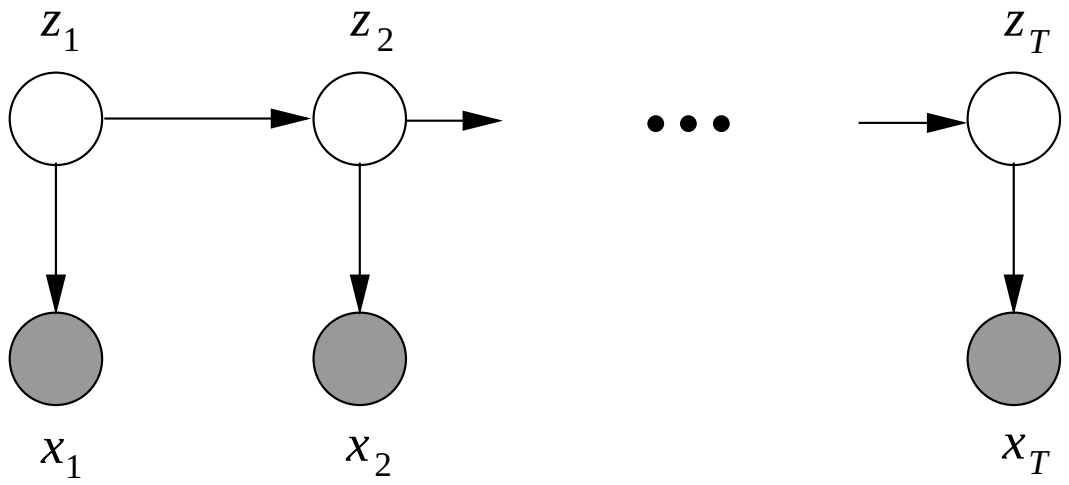
Figure 4. Left: Average perplexity of test VS documents given training documents from VS and another section for 3 different models. curves shown are averaged over the other sections and other 5 runs. Right: Average perplexity of test VS documents given LT, AA and AP documents respectively using M3, averaged over 5 runs.

Shared topics

- Topics shared between VS and the other sections
  - the two highest probability topics are displayed

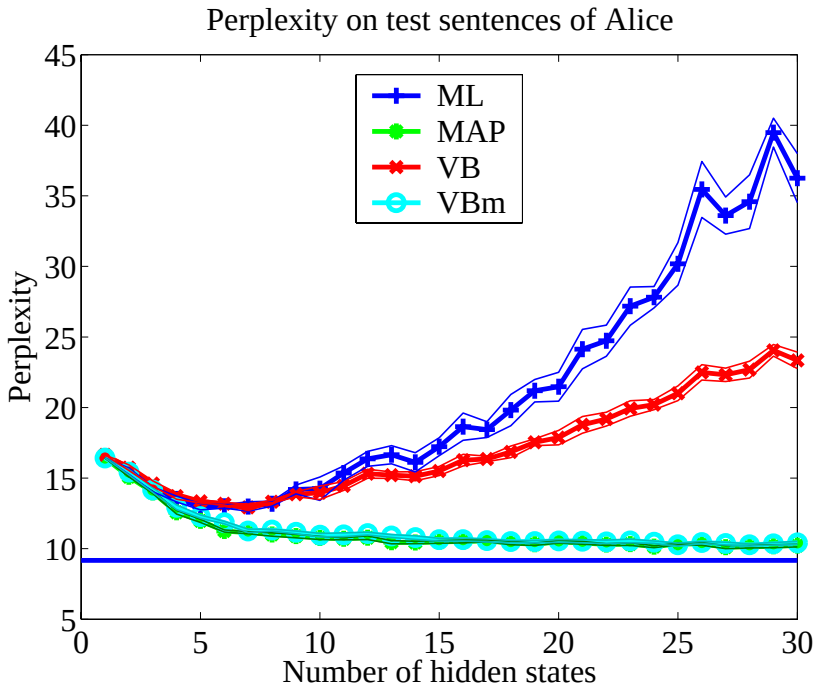
CS	NS	LT	AA	IM	SP	AP	CN
task representation pattern processing trained representations three process unit patterns	cells cell activity response neuron visual patterns pattern single	signal layer gaussian cells figure nonlinear rate equation cell	algorithms test approach methods based point problems large paper	processing pattern approach architecture single shows simple based large	visual images video language image pixel acoustic delta lowpass	approach based trained test layer features table classification rate paper	tree pomdp observable strategy class stochastic history strategies density
examples concept similarity bayesian hypotheses generalization numbers positive classes hypothesis	visual cells cortical orientation receptive contrast spatial cortex stimulus tuning	large examples form point see parameter consider random small optimal	distance tangent image images transformation transformations pattern vectors convolution simard	motion visual velocity flow target chip eye smooth direction optical	signals separation signal sources source matrix blind mixing gradient eq	image images face similarity pixel visual database matching facial examples	policy optimal reinforcement control action states actions step problems goal

Hidden Markov models



- The "infinite hidden Markov model"—an HMM with an unbounded number of states
- Straightforward to use the HDP framework
  - multiple mixture models—one for each value of the "current state"
  - the DP creates new states, and the HDP approach links the transition distributions

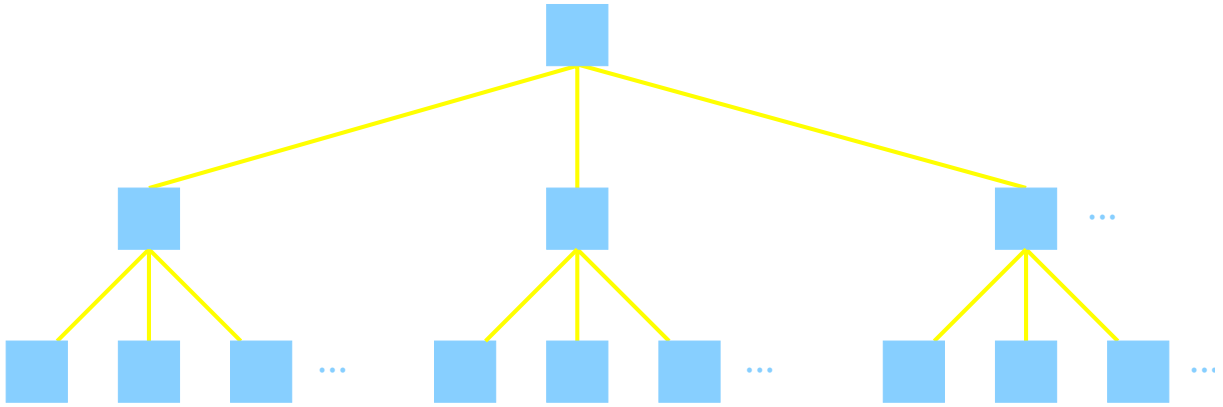
Alice in Wonderland



- Perplexity of test sentences taken from Lewis Carroll's *Alice in Wonderland*

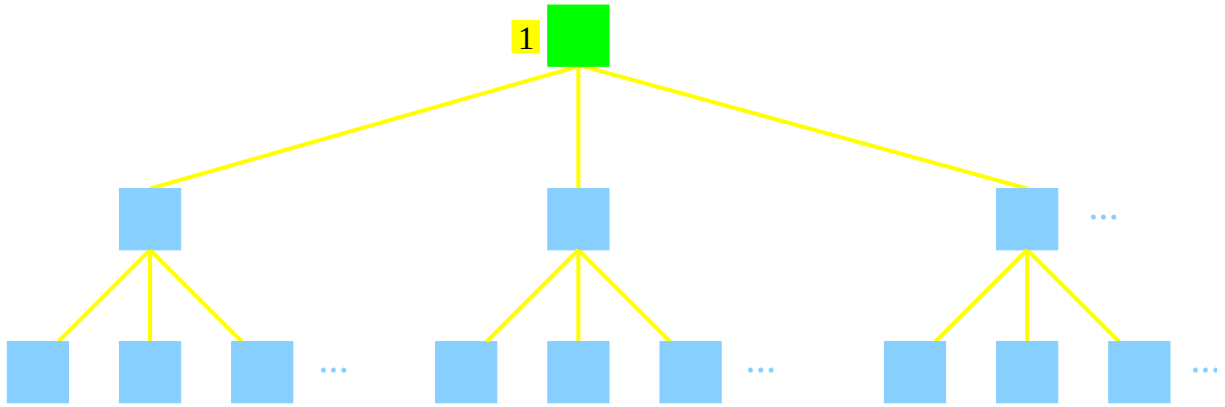
Hierarchical topic models

(Blei, Griffiths, Jordan, & Tenenbaum, 2004)



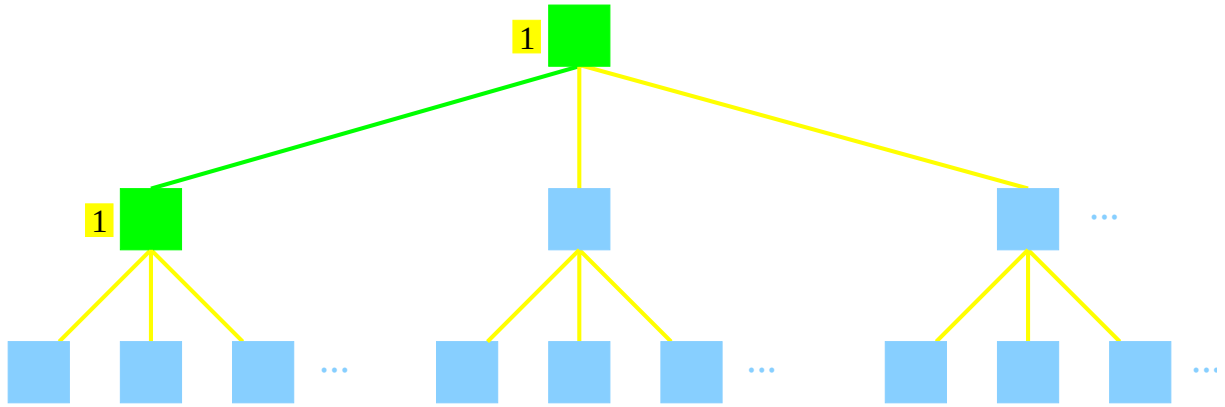
- Infinite number of restaurants in a city:
  - One restaurant is the root restaurant and on each of its infinite tables is a card with the name of another restaurant
  - On each of the tables in those restaurants are cards that refer to other restaurants, and this structure repeats
- Restaurants are organized into an infinitely branching tree

Nested Chinese restaurant process



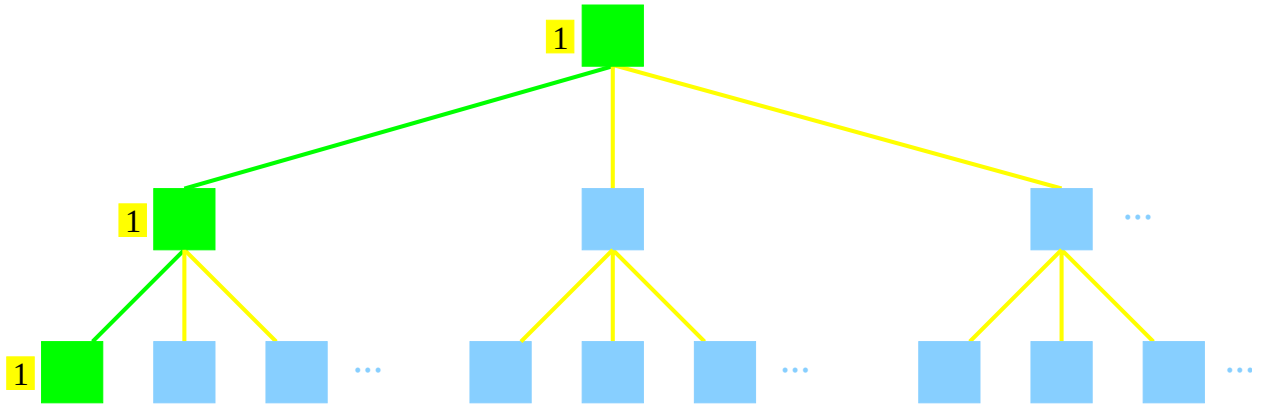
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



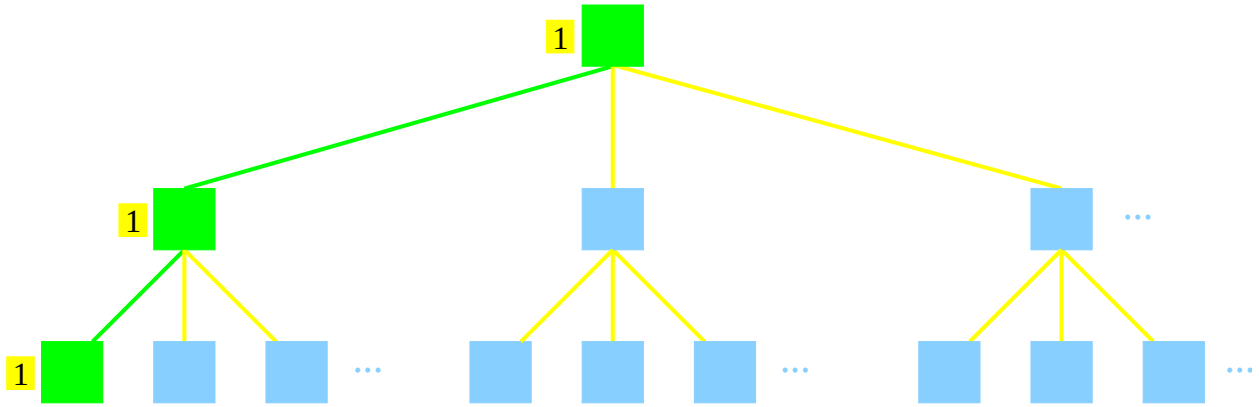
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



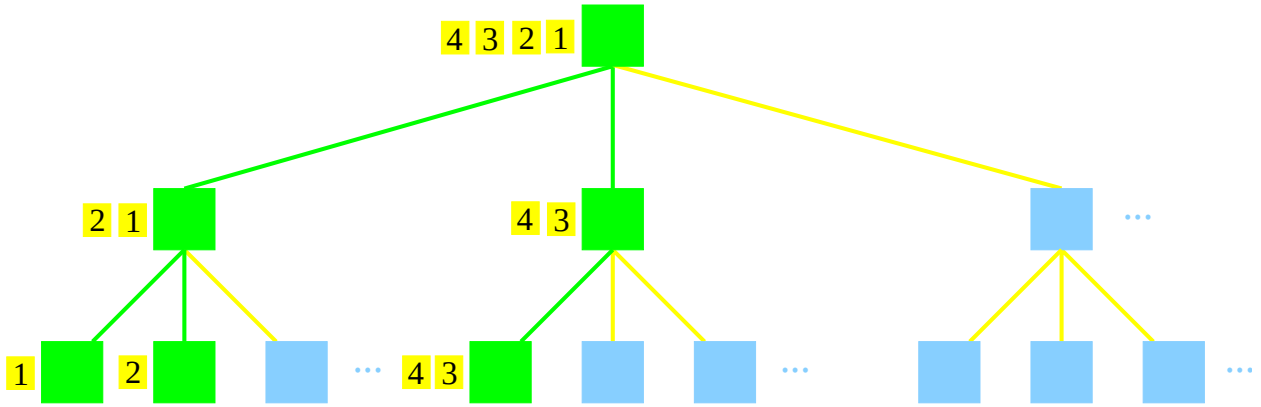
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree

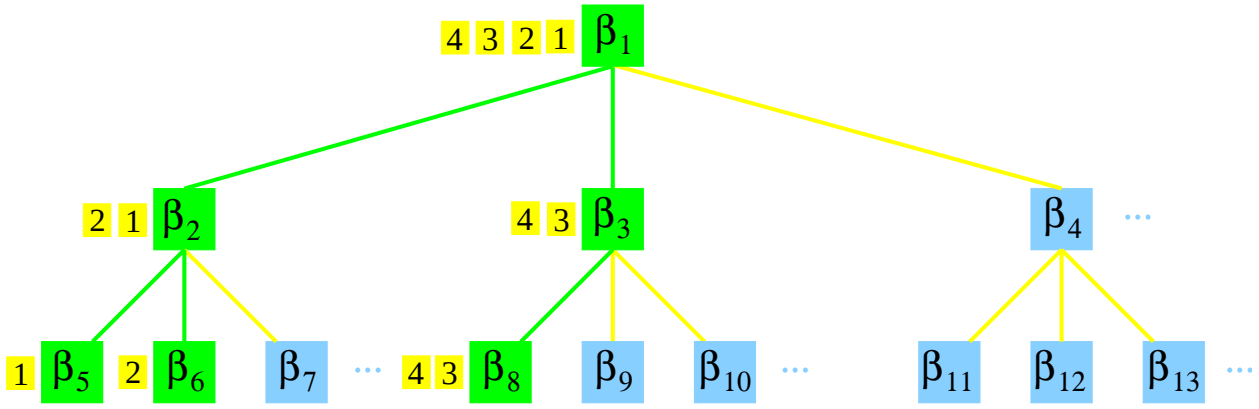
Nested Chinese restaurant process



- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree

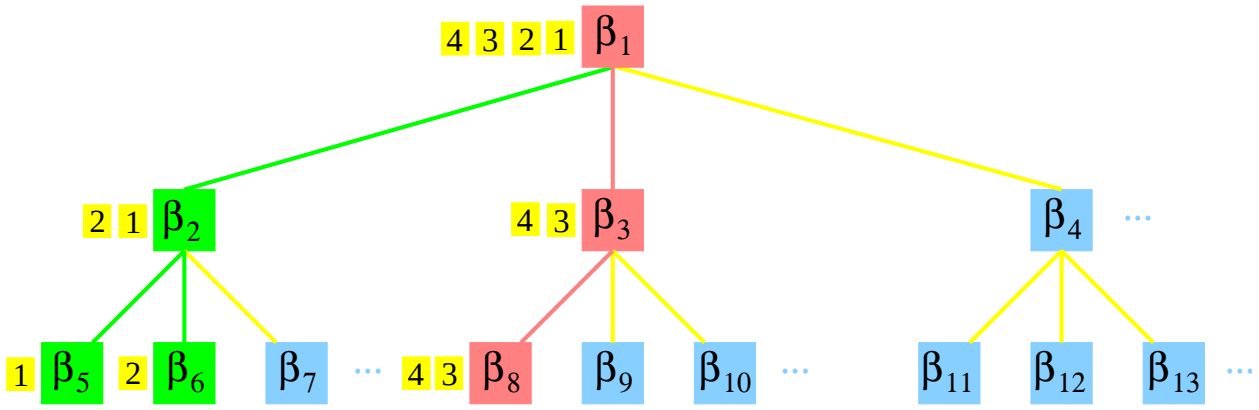
- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree

Nested Chinese restaurant process



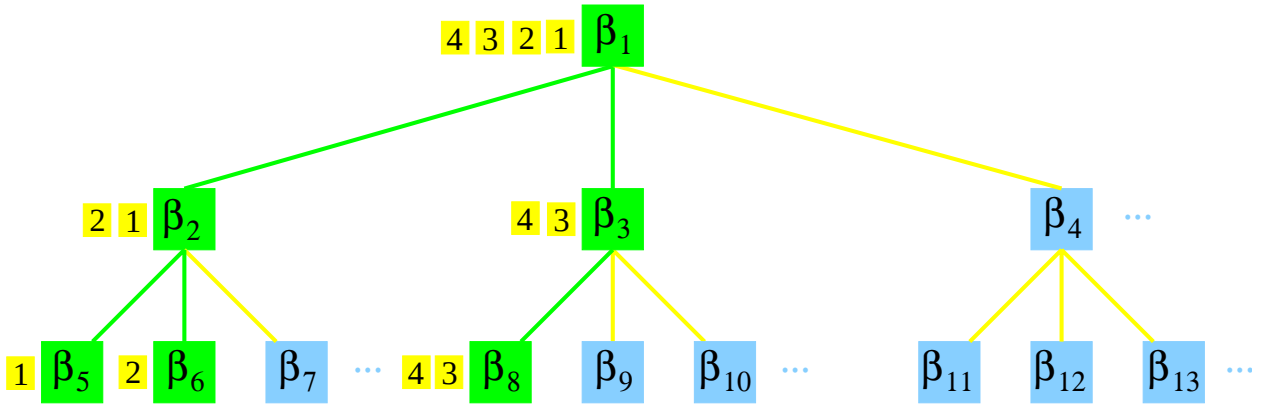
- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree
- Assigning each restaurant to a parameter, we can use each tour as a topic path for a document

Nested Chinese restaurant process



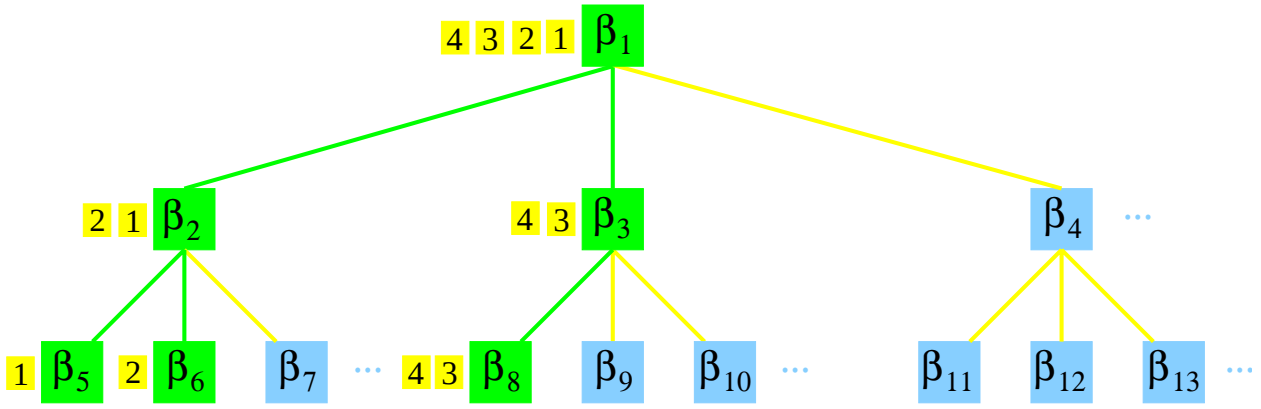
- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree
- Assigning each restaurant to a parameter, we can use each tour as a topic path for a document

Hierarchical LDA



- Choose a path through the infinite tree of restaurants
- Choose a distribution θ over levels
- For each word
  - Choose a level from $\mathrm{Mult}(\theta)$
  - Draw the word from the topic in the restaurant at that level

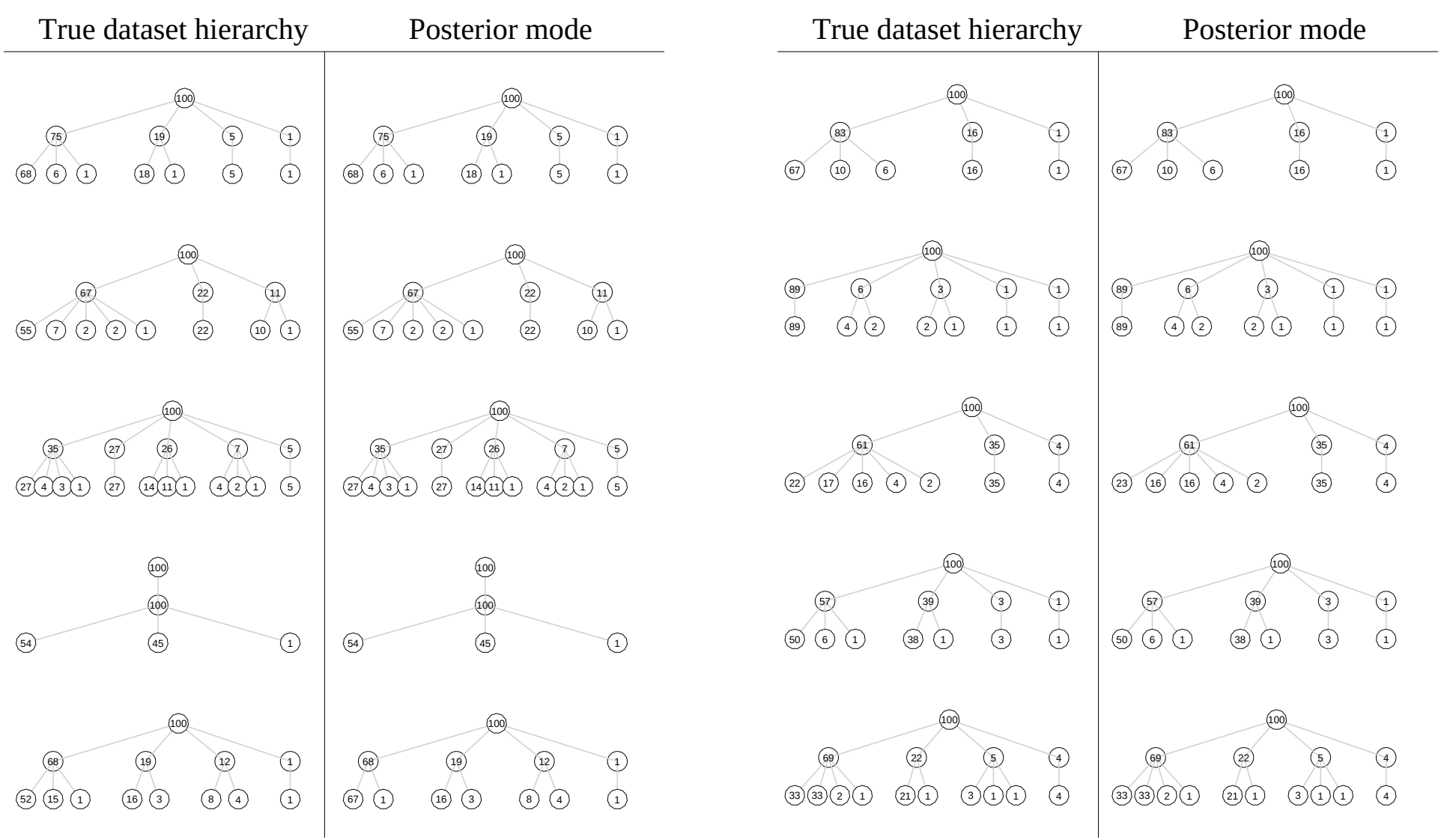
Hierarchical LDA



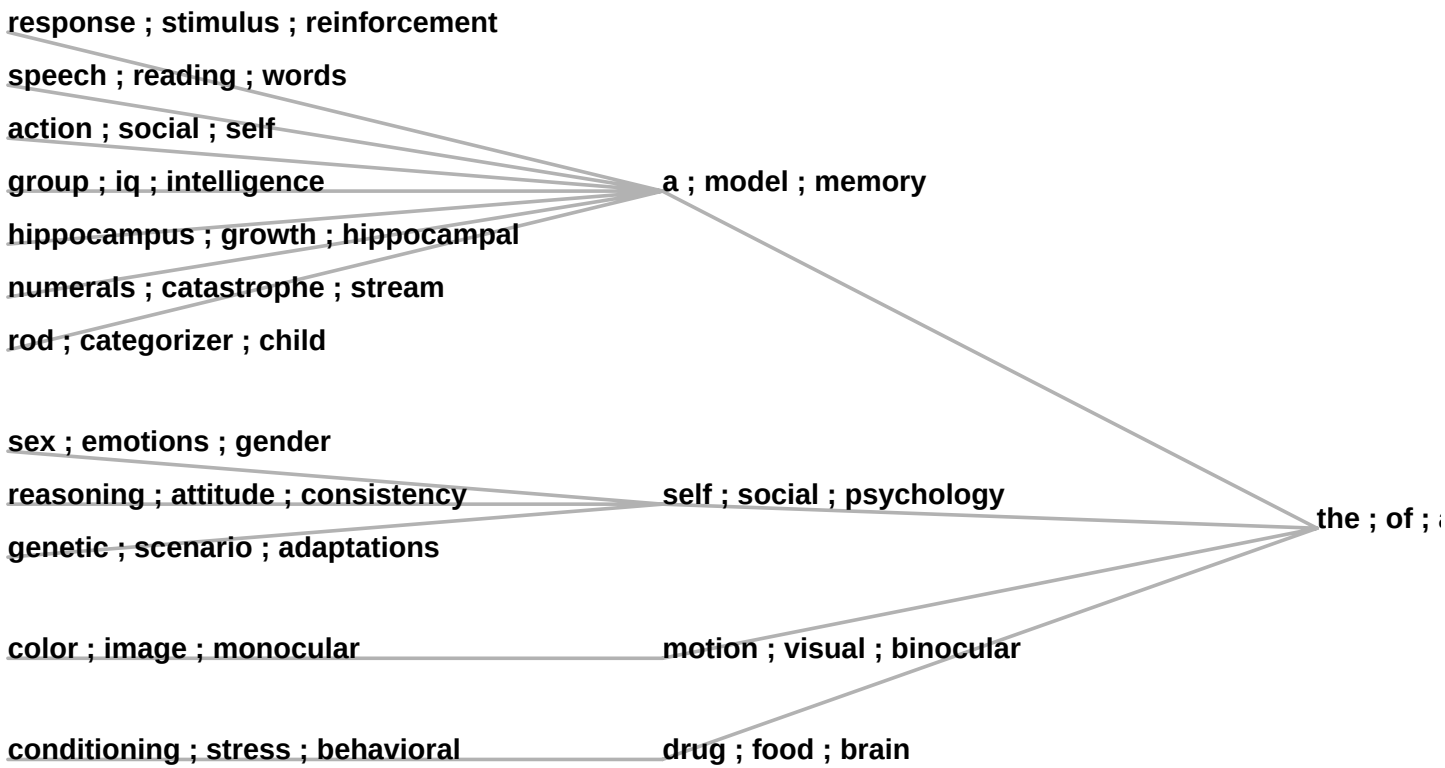
- Given a document collection, posterior is a distribution of
  - The structure of the hierarchy
  - Assignment of documents to paths, words to levels
  - Topics which populate the hierarchy

- Allows new documents to fill unoccupied parts of the tree

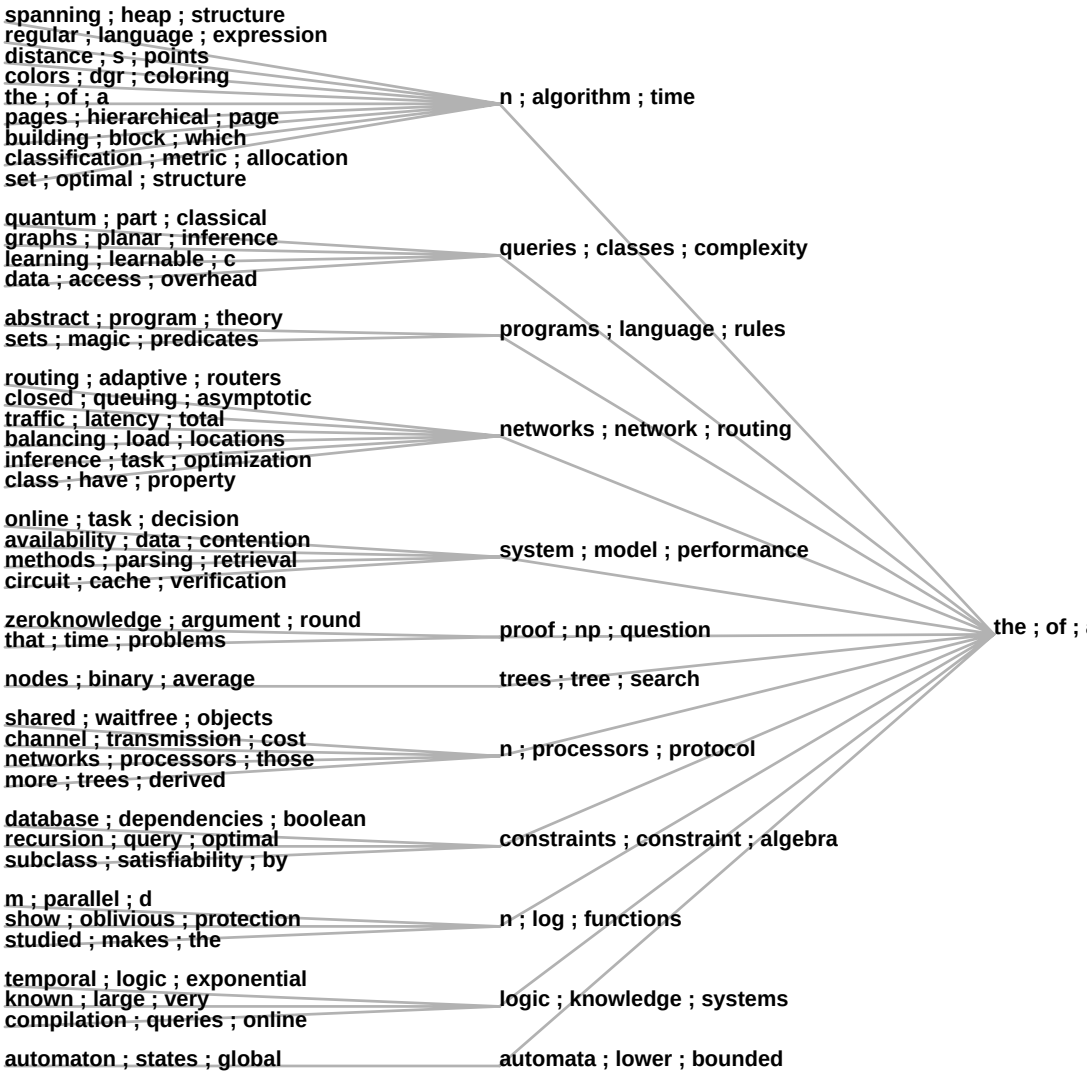
Estimating the correct hierarchy



Topic hierarchy from *Psychology*



Topic hierarchy from $JACM$



Variational Algorithms

- Three steps:
  - convert the inference problem into an optimization problem
  - relax the optimization problem into a simplified optimization problem
  - solve the relaxation

- Many variations
  - *mean field algorithms* (pretend the law of large numbers holds)
  - *sum-product algorithm* (pretend the graph is a tree)

Conjugate Duality Refresher

- For a convex function $f(x)$, we have:

$$\begin{aligned} f(x) &= \sup_{\mu} \left\{ \mu x - f^*(\mu) \right\} \\ f^*(\mu) &= \sup_{x} \left\{ \mu x - f(x) \right\}, \end{aligned}$$

  where $f^*(\mu)$ is the *conjugate function*.

- E.g., conjugate duality for e^x:

$$e^x = \sup_{\mu} \left\{ \mu x - \mu \log \mu + \mu \right\}$$

- Implies a family of bounds, indexed by the "variational parameter" μ:

$$e^x \geq \mu x - \mu \log \mu + \mu$$

- Setting μ equal to one yields a simple "convexity bound":

$$e^x \geq x + 1$$

Mean Field Intuition

- Recall the family of bounds:

$$e^x \geq \mu x - \mu \log \mu + \mu$$

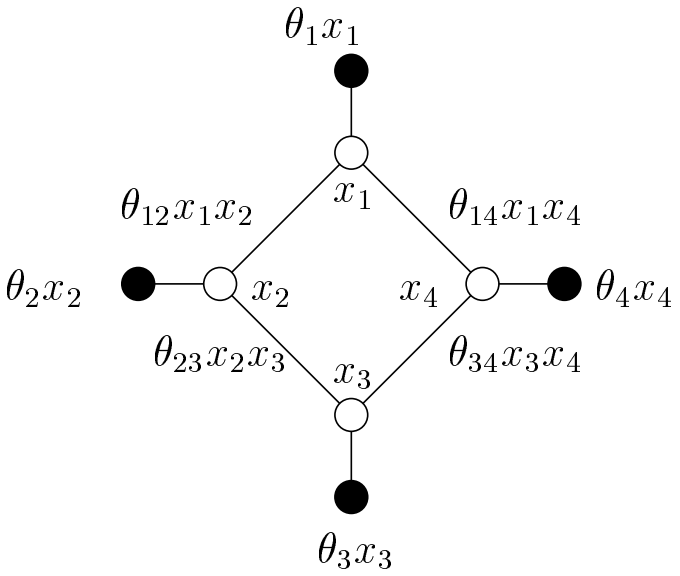
- Useful in a probabilistic setting if there is a concentration in x.
  - the bound is tight for $x \approx \log \mu$
  - turns a nonlinearity into a linearity

- Need to find a value of μ that allows us to exploit the (posited) concentration in x.

- When there are many coupled variables, need to solve a system of equations involving multiple variational parameters $\{\mu_i\}$, one for each variable.

Example—The Ising Model

- Binary variables on a graph with pairwise cliques



$$\begin{aligned} \boldsymbol{\phi} &= \{\, x_s \mid s \in V \,\} \cup \{\, x_s x_t \mid (s,t) \in E \,\} \\ \mathcal{I} &= V \cup E \\ \mathcal{X}^n &= \{0,1\}^n \end{aligned}$$

$$p(\mathbf{x};\theta) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta) \right\}$$

Inference for Ising Model

- *Gibbs sampler*

$$x_s \leftarrow \begin{cases} 1 & \text{if } u \leq \{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} x_t)]\}^{-1} \\ 0 & \text{otherwise} \end{cases},$$

  where $u \sim \mathcal{U}(0, 1)$.

- *Naive mean field algorithm*

$$\mu_s \leftarrow \left\{ 1 + \exp\left[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t) \right] \right\}^{-1},$$

  where $\mu_s \in [0, 1]$ are *variational parameters*.

Inference for Ising model (cont.)

- *Sum-product algorithm*

$$\mu_{ts}(x_s) \leftarrow \sum_{x'_t} \left\{ \theta_{st} x_s x'_t \prod_{u \in \mathcal{N}(t)/s} \mu_{ut}(x'_t) \right\}$$

$$\mu_s(x_s) \propto \theta_s x_s \prod_{t \in \mathcal{N}(s)} \mu_{ts}(x_s),$$

where $\mu_s \in [0,1]$ and $\mu_{st} \in [0,1]$ are *variational parameters*.

Exponential Representations

- Parameterized family of distributions:

$$p(\mathbf{x};\theta) \quad = \quad \exp\Big\{ \sum_{\alpha} \theta_\alpha \phi_\alpha(\mathbf{x}) \; - \; \Phi(\theta) \Big\}$$

- Cumulant generating function (aka, log partition function):

$$\Phi(\theta) \quad = \quad \log \Big(\sum_{\mathbf{x} \in \mathcal{X}^n} \exp\{ \sum_{\alpha} \theta_\alpha \phi_\alpha(\mathbf{x}) \} \Big)$$

$$\begin{array}{lcl} \boldsymbol{\phi} = \{ \, \phi_\alpha \mid \alpha \in \mathcal{I} \, \} & \equiv & \text{sufficient statistics (aka, potential functions)} \\ \theta = \{ \, \theta_\alpha \mid \alpha \in \mathcal{I} \, \} & \equiv & \text{canonical parameters} \end{array}$$

Variational Approach

- **Basic idea:** Represent a quantity of interest $\widehat{z}$ as the solution of an optimization problem:
  - study $\widehat{z}$ via the optimization problem.
  - approximate $\widehat{z}$ by approximating the optimization problem.

Variational Approach

- **Basic idea:** Represent a quantity of interest $\widehat{z}$ as the solution of an optimization problem:
  - study $\widehat{z}$ via the optimization problem.
  - approximate $\widehat{z}$ by approximating the optimization problem.

- **Goal:** Obtain a variational representation for:
  - the log partition function.
  - the inference problem of computing $\mu_\alpha := \mathbb{E}[\phi_\alpha(\mathbf{x})]$.

The Marginal Polytope

- **Dual perspective:** Define the optimization problem in terms of *only* the mean parameters:

$$\mu_\alpha \quad := \quad \sum_{\mathbf{x}} p(\mathbf{x})\phi_\alpha(\mathbf{x})$$

- **Question:** What set do these mean parameters range over?

The Marginal Polytope

- **Dual perspective:** Define the optimization problem in terms of *only* the mean parameters:

$$\mu_\alpha \quad := \quad \sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x})$$

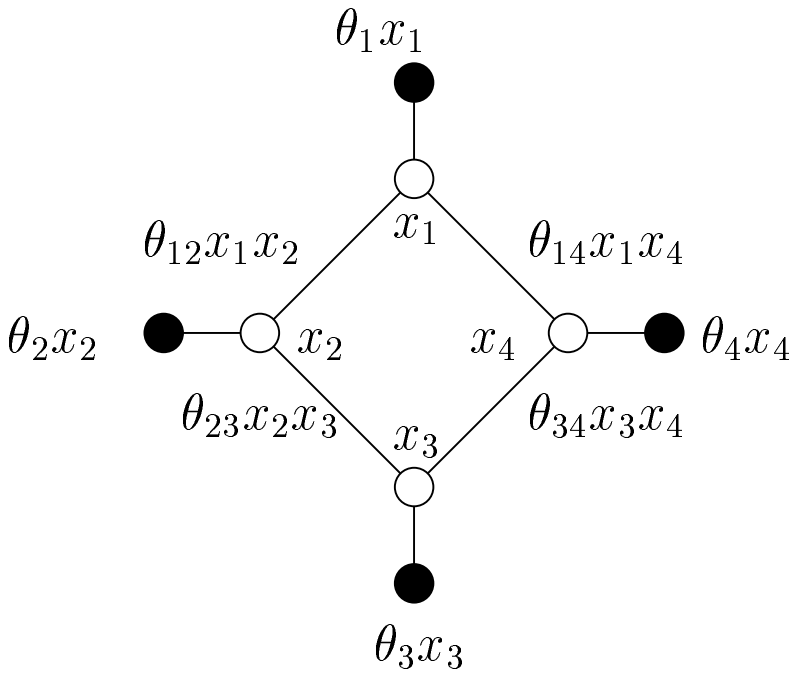
- **Question:** What set do these mean parameters range over?

- Define $\mathcal{M}(G; \boldsymbol{\phi})$ as the set of realizable or globally consistent marginals:

$$\mathcal{M}(G; \boldsymbol{\phi}) \quad = \quad \left\{ \mu \in \mathbb{R}^d \;\middle|\; \mu = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}) \qquad \text{for } \textit{some } p(\cdot) \right\}$$

- For discrete families, we refer to this set as the marginal polytope, and denote it as $\mathrm{MARG}(G; \boldsymbol{\phi})$

Ising Model Example

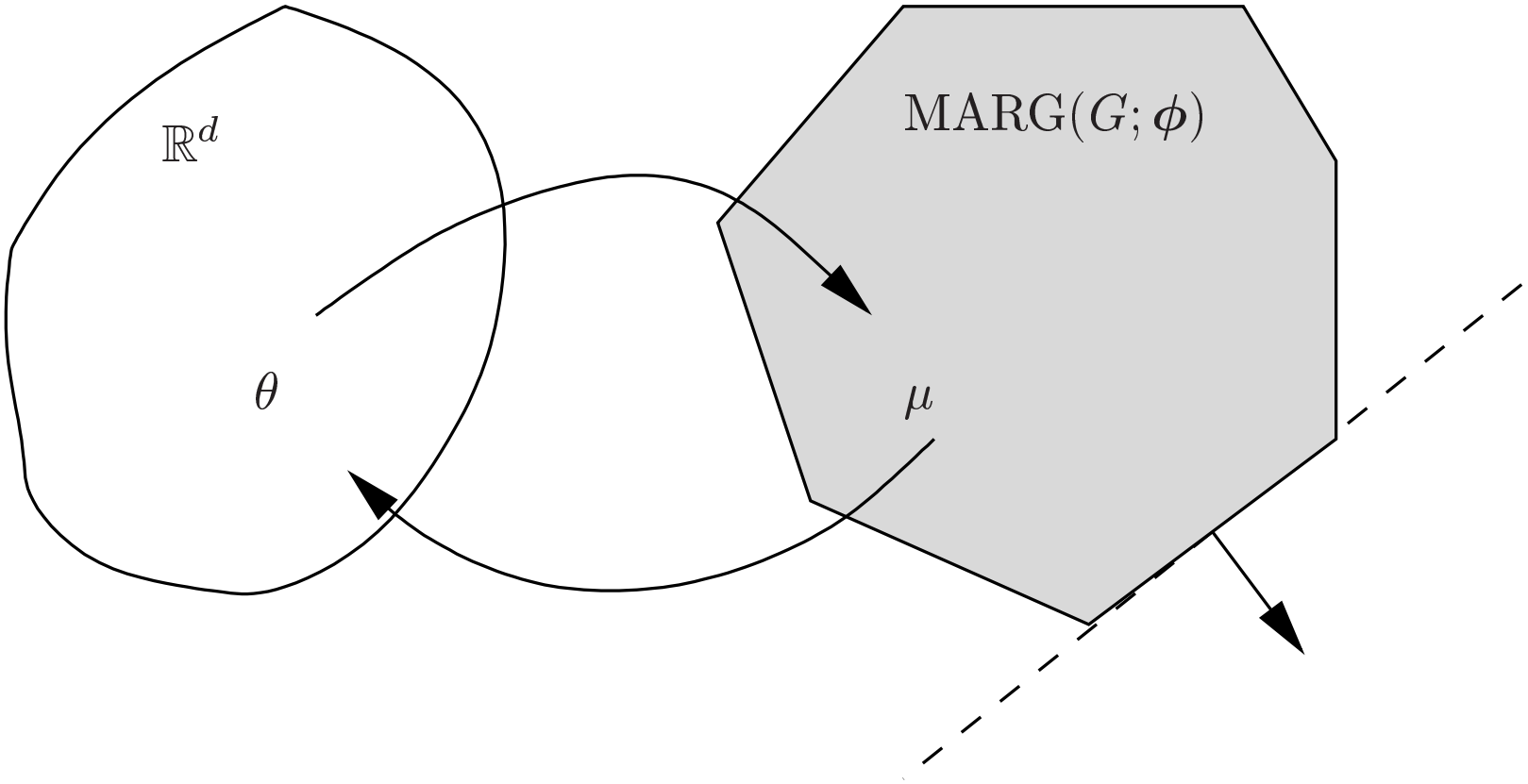


Potentials $\quad \phi = \{x_s \mid s \in V\} \cup \{x_s x_t \mid (s,t) \in E\}$

Relevant marginals $\quad \mu_s = \mathbb{E}_\theta[x_s] \qquad \mu_{st} = \mathbb{E}_\theta[x_s x_t]$

- Associated constraint set is known as the *correlation polytope* or the *binary quadric polytope*. (e.g., Deza & Laurent, 1997)

Geometry and Moment Mapping



The Conjugate Dual of the Log Partition Function

- Given $\mu \in \mathrm{MARG}(G; \boldsymbol{\phi})$, let $\theta(\mu)$ denote the corresponding canonical parameter.

- Compute the conjugate dual:

$$\begin{aligned} \Phi^*(\mu) &= \max_{\theta} \quad \{\langle \mu, \theta \rangle \quad - \quad \Phi(\theta)\} \\ &= \qquad \{\langle \mu, \theta(\mu) \rangle \quad - \quad \Phi(\theta(\mu))\}. \end{aligned}$$

- The entropy of a distribution in the exponential family:

$$\begin{aligned} H(p(\mathbf{x}; \theta(\mu))) &= -\sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta(\mu)) \log p(\mathbf{x}; \theta(\mu)) \\ &= - \quad \{\langle \mu, \theta(\mu) \rangle \quad - \quad \Phi(\theta(\mu))\}. \end{aligned}$$

- I.e., for $\mu \in \mathrm{MARG}(G; \boldsymbol{\phi})$, the conjugate dual function is just the negative entropy.

Variational Principle in Terms of Marginals

- It turns out that outside of $\mathrm{MARG}(G; \boldsymbol{\phi})$, the conjugate dual function is infinite. Thus:

$$\Phi^*(\mu) = \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \in \mathrm{MARG}(G; \boldsymbol{\phi}) \\ +\infty & \text{otherwise.} \end{cases}$$

Variational Principle in Terms of Marginals

- It turns out that outside of $\mathrm{MARG}(G; \boldsymbol{\phi})$, the conjugate dual function is infinite. Thus:

$$\Phi^*(\mu) = \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \in \mathrm{MARG}(G; \boldsymbol{\phi}) \\ +\infty & \text{otherwise.} \end{cases}$$

- Plugging in to the general conjugacy formula, this leads to a representation of Φ in terms of Φ^*:

$$\underbrace{\Phi(\theta)}_{\text{log partition function}} = \underbrace{\max_{\mu \in \mathrm{MARG}(G; \boldsymbol{\phi})} \{\langle \mu, \theta \rangle - \Phi^*(\mu)\}}_{\text{convex optimization problem over marginal polytope}}$$

Variational Principle in Terms of Marginals

- It turns out that outside of $\mathrm{MARG}(G; \boldsymbol{\phi})$, the conjugate dual function is infinite. Thus:

$$\Phi^*(\mu) \quad = \quad \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \in \mathrm{MARG}(G; \boldsymbol{\phi}) \\ +\infty & \text{otherwise.} \end{cases}$$

- Plugging in to the general conjugacy formula, this leads to a representation of Φ in terms of Φ^*:

$$\underbrace{\Phi(\theta)}_{\text{log partition function}} \quad = \quad \underbrace{\max_{\mu \in \mathrm{MARG}(G; \boldsymbol{\phi})} \{\langle \mu, \theta \rangle - \Phi^*(\mu)\}}_{\text{convex optimization problem over marginal polytope}}$$

- Moreover, maximum is attained uniquely at desired marginals:

$$\mu_\alpha \quad = \quad \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta) \phi_\alpha(\mathbf{x}) \quad = \quad \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})].$$

Mean Field Algorithms

- Let H represent a *tractable subgraph*—a subgraph of G over which it feasible to perform exact calculations (e.g., the completely disconnected graph).

- Set of exponential parameters corresponding to distributions structured according to H:

$$\mathcal{E}(H) \quad := \quad \{\theta \in \Theta \mid \theta_\alpha = 0 \quad \forall \ \alpha \in \mathcal{I} \backslash \mathcal{I}(H)\},$$

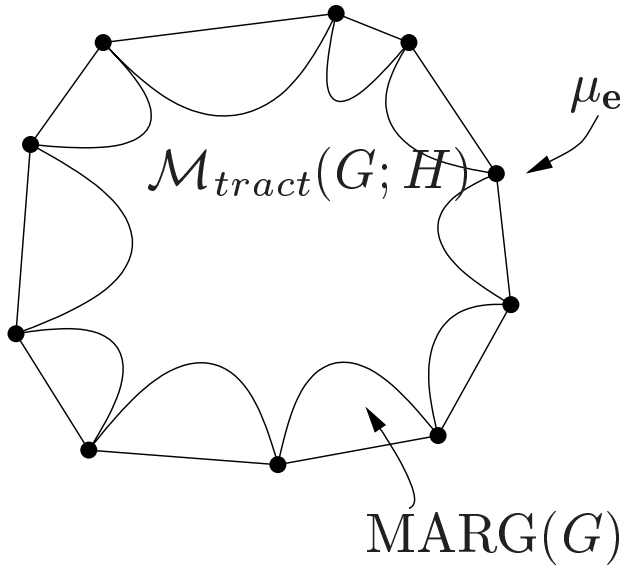
  where $\mathcal{I}(H)$ is the subset of indices associated with cliques in H.

- Consider the set of all possible mean parameters that are realizable by tractable distributions:

$$\mathcal{M}_{tract}(G; H) \quad := \quad \{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \text{ for some } \theta \in \mathcal{E}(H)\}.$$

Mean Field Algorithms (cont.)

- Since any μ that arises from a tractable distribution is certainly a valid mean parameter, the inclusion $\mathcal{M}_{tract}(G; H) \subseteq \mathrm{MARG}(G; \phi)$ always holds. I.e., $\mathcal{M}_{tract}$ is an *inner approximation*:



- Note that the set of tractable distributions is a *non-convex* set.

Mean Field Algorithms (cont.)

- Optimizing over $\mathcal{M}_{tract}$ instead of $\mathcal{M}$ yields an *approximation* to the variational principle:

$$\underbrace{\Phi(\theta)}_{\text{log partition function}} \quad \geq \quad \underbrace{\max_{\mu \in \mathcal{M}_{tract}(G;H)} \{\langle \mu, \theta \rangle - \Phi^*(\mu)\}}_{\text{optimization over set of tractable distributions}}$$

- The entropy $\Phi^*(\mu)$ can be computed exactly because (by assumption) we are restricted to tractable distributions

- We obtain a *lower bound* on $\Phi(\theta)$, because we optimize the same expression as before over a smaller set.

Naive Mean Field for the Ising Model

- Completely disconnected graph $H_0 = (V, \emptyset)$
- Permissible parameters belong to the subspace $\mathcal{E}(H_0) := \{\theta \in \Theta \mid \theta_{st} = 0, \ \forall \ (s,t) \in E\}$.
  - the associated distributions are of the product form $p(\mathbf{x}; \theta) = \prod_{s \in V} p(x_s; \theta_s)$.
- The approximate variational principle becomes:

$$\max_{\{\mu_s\} \in [0,1]^n} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} \left[\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s) \right] \right\}$$

- **Coordinate ascent:** with all $\{\mu_t, t \neq s\}$ fixed, problem is strictly concave in μ_s and optimum is attained at

$$\mu_s \quad \longleftarrow \quad \left\{ 1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t)] \right\}^{-1}$$

Alternative View: Minimizing a KL Divergence

- The *mixed form* of the KL divergence between $p(\mathbf{x};\theta)$ and $p(\mathbf{x};\widetilde{\theta})$:

$$D(\widetilde{\mu}\,||\,\theta) \quad = \quad \Phi(\theta) + \Phi^*(\widetilde{\mu}) - \langle\widetilde{\mu},\,\theta\rangle$$

- Try to find the "best" approximation to $p(\mathbf{x};\theta)$ in the sense of KL divergence

- This can be written as

$$\inf_{\widetilde{\mu}\in\mathcal{M}_{tract}} D(\widetilde{\mu}\,||\,\theta) \quad = \quad \Phi(\theta) + \inf_{\widetilde{\mu}\in\mathcal{M}_{tract}} \left\{\Phi^*(\widetilde{\mu}) - \langle\widetilde{\mu},\,\theta\rangle\right\}$$

- Hence, in the mean field setting ($\mathcal{M}_{tract} \subseteq \mathcal{M}$), finding the best approximation (in the KL sense) to $p(\mathbf{x};\theta)$ from distributions with $\widetilde{\mu} \in \mathcal{M}_{tract}$ is equivalent to solving our variational problem

Mean Field Algorithms for the Dirichlet Process

(Blei & Jordan, 2005)

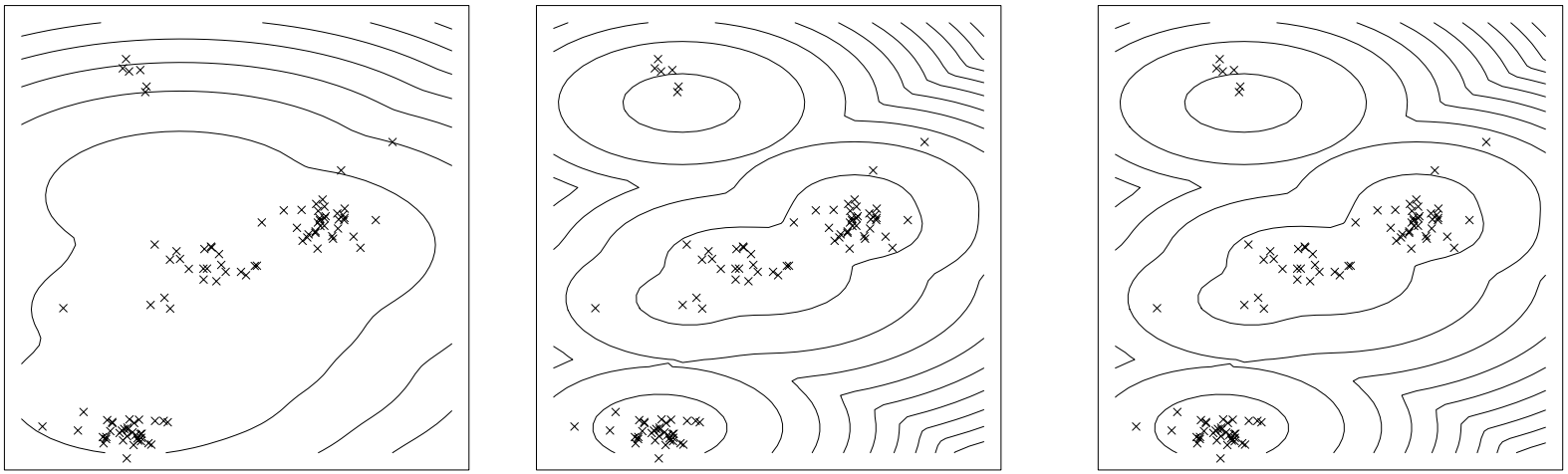
- Express the Dirichlet process using the stick-breaking representation; this is $\mathcal{M}$
- Let $\mathcal{M}_{tract}$ be the set of *truncated* stick-breaking processes
- Optimize the KL divergence; this yields (suppressing details):

$$\begin{aligned}
\gamma_{i,1} &= 1 + \textstyle\sum_n \phi_{n,i} \\
\gamma_{i,2} &= \alpha + \textstyle\sum_n \sum_{j=i+1}^{K} \phi_{n,j} \\
\tau_{i,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,i} x_n \\
\tau_{i,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,i} \\
\phi_{n,i} &\propto \exp(S),
\end{aligned}$$

where

$$S = E[\log V_i \mid \gamma_i] + E[\eta_i \mid \tau_i]^T X_n - E[a(\eta_i) \mid \tau_i] - \textstyle\sum_{j=i+1}^{K} E[\log(1 - V_j) \mid \gamma_j].$$

Example: DP-Gaussian Mixture



Initial state 1st iteration 5th (and last) iteration

Figure 5. The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

Example: DP-Gaussian Mixture

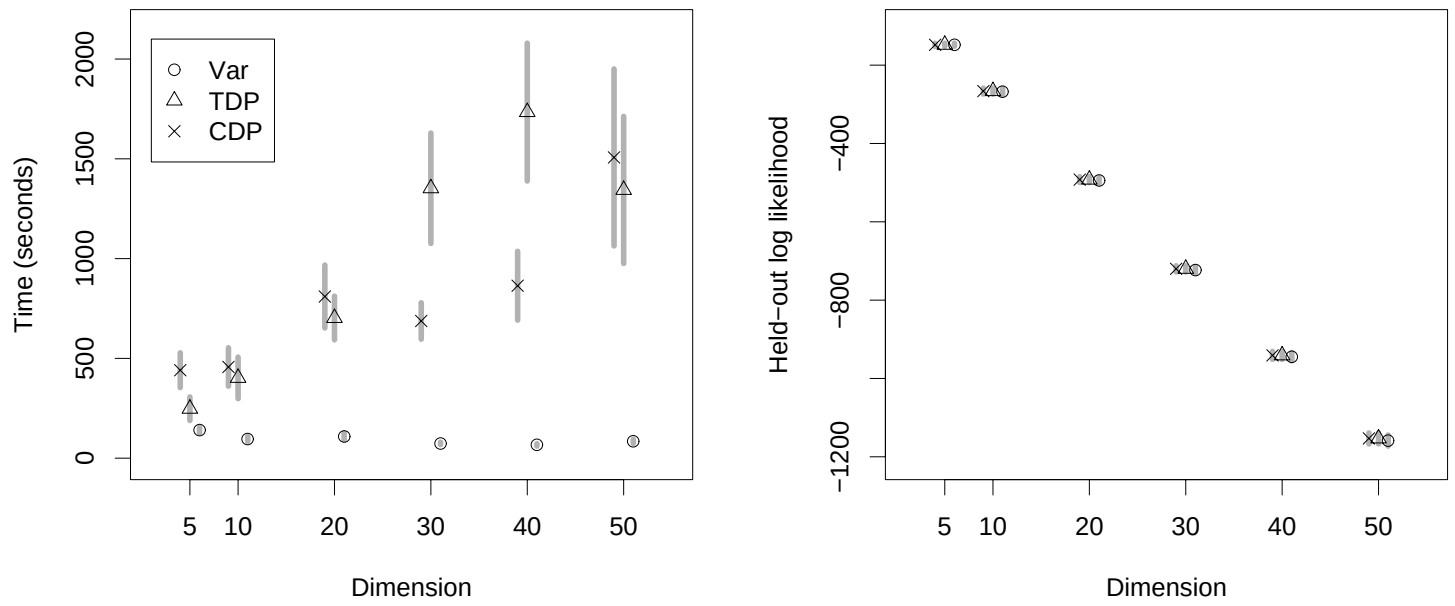


Figure 6. (Left) Convergence time per dimension across ten datasets for variational inference (Var), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

The Bethe Approximation

(Yedidia, Freeman & Weiss, 2001)

- Relax the constraint that the "marginals" that we obtain from the optimization are consistent with any joint probability distribution (e.g., they need not be globally consistent)
  - we'll refer to such quantities as *pseudomarginals* (often referred to as *beliefs*)

- Focus on a *pairwise* undirected graphical model:

$$p(\mathbf{x};\theta) \quad \propto \quad \exp\Big\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\Big\}$$

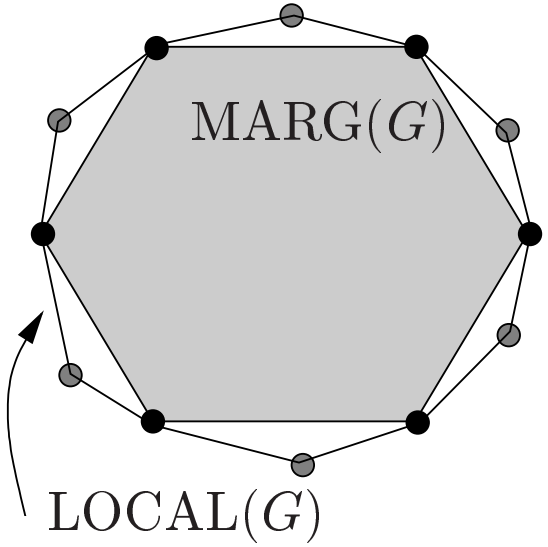
- Denote the corresponding marginals as $\mu_s(x_s)$ and $\mu_{st}(x_s,x_t)$.

The Bethe Approximation (cont.)

- Consider the relaxed constraint set:

$$\mathrm{LOCAL}(G) \quad = \quad \{\, \mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \;\; \sum_{x_s} \mu_{st}(x_s, x_t) = \mu_t(x_t) \,\}.$$

- These constraints are necessary conditions on marginals; thus we obtain an *outer approximation* to $\mathrm{MARG}(G; \boldsymbol{\phi})$:



The Bethe Approximation (cont.)

- We must approximate the entropy
  - we're no longer working with tractable distributions
  - indeed, we're no longer necessarily working with distributions at all

- The *Bethe entropy approximation*:

$$H_{Bethe}(\mu) \quad := \quad \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}),$$

  where $I_{st}(\mu_{st}) = H_s(\mu_s) + H_t(\mu_t) - H_{st}(\mu_{st})$ is the mutual information.

- This expression is exact on a tree; in general it is an approximation.

The Bethe Approximation (cont.)

- Combining the entropy approximation H_{Bethe} with the tree-based constraint set $\mathrm{LOCAL}(G)$ leads to the *Bethe variational problem*:

$$\max_{\mu \in \mathrm{LOCAL}(G)} \Big\{ \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) \Big\}.$$

- Although $\mathrm{LOCAL}(G)$ is a convex set, $\sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$ is not a convex function, so the problem overall is not convex.

- Taking derivatives with respect to the pseudomarginals yields the sum-product updates presented earlier.

Summary of Current Variational Algorithms

- Obtain algorithms by *relaxation* of original problem
  - can consider inner or outer approximations to $\mathrm{MARG}(G; \boldsymbol{\phi})$
  - can approximate $\Phi^*(\mu)$ in various ways

- The sum-product algorithm involves an *outer approximation* to $\mathrm{MARG}(G; \boldsymbol{\phi})$, and the *Bethe approximation* to the entropy $\Phi^*(\mu)$ ("tree-consistent" pseudomarginals)

- Mean field algorithms involve an *inner approximation* to $\mathrm{MARG}(G; \boldsymbol{\phi})$. No approximation is needed for the entropy $\Phi^*(\mu)$.
  - thus, mean field algorithms yield a lower bound on the log partition function (sum-product yields no bound).

- Neither the mean field approach nor the Bethe approach yield a convex relaxation.

Convex Relaxations

(Wainwright & Jordan, 2003)

- **Goal:** Obtain upper bounds by a *convex relaxation*. This will yield an algorithm with a single global optimum.

- **Requirements:**
  - convex outer approximation to marginal polytope $\mathrm{MARG}(G; \boldsymbol{\phi})$.
  - concave upper bound on entropy function $-\Phi^*(\mu)$.

Semidefinite Outer Bounds on Marginal Polytopes

- Focus on: (a) binary case with "spins" $\mathbf{x} \in \{-1, +1\}^n$.
  (b) complete graph K_n on n nodes.

- Refer to the associated marginal polytope as $\mathrm{MARG}(K_n)$.

- Relevant marginals:
$$\mu_s = \mathbb{E}_\theta[x_s] \quad \text{for all } s = 1, \ldots n$$
$$\mu_{st} = \mathbb{E}_\theta[x_s x_t] \quad \text{for all } (s,t)$$

- Sequence of semidefinite relaxations on the binary marginal polytope $\mathrm{MARG}(K_n)$ (e.g., Lasserre, 2001)

Covariance Matrix

- The covariance matrix of $\mathbf{x}$ must be positive semidefinite:

$$\mathrm{cov}(\mathbf{x}) \;=\; \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^T] \;\succeq\; 0$$

Covariance Matrix

- The covariance matrix of $\mathbf{x}$ must be positive semidefinite:

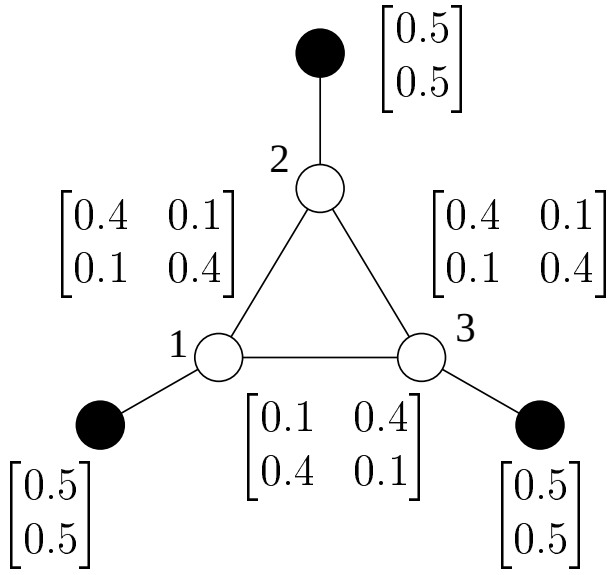
$$\mathrm{cov}(\mathbf{x}) \;=\; \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^T] \;\succeq\; 0$$

- By Schur complement, equivalent to enforce PSD constraint on

$$\mathbb{E}\left\{ \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \right\} \;=\; \left[\begin{array}{c|cccccc} 1 & \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_n \\ \hline \mu_1 & 1 & \mu_{12} & \mu_{13} & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & 1 & \mu_{23} & \cdots & \mu_{2n} \\ \mu_3 & \mu_{31} & \mu_{32} & 1 & \vdots & \mu_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_n & \mu_{n1} & \mu_{n2} & \mu_{n3} & \cdots & 1 \end{array}\right]$$

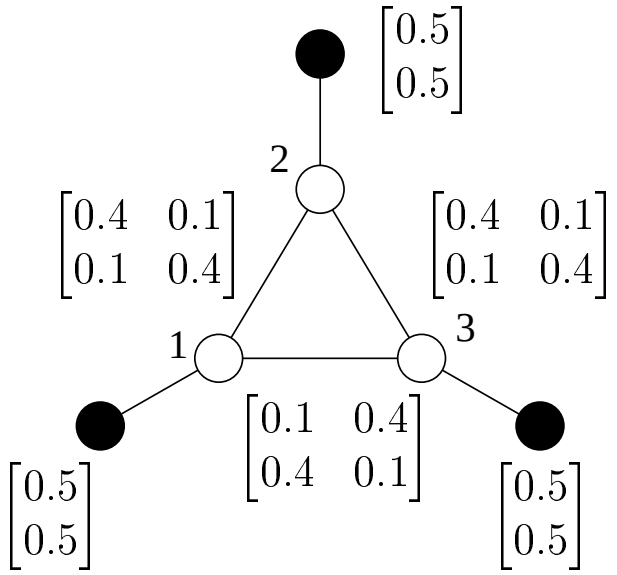
Illustrative Example

Tree-consistent
(pseudo)marginals



Illustrative Example

Tree-consistent
(pseudo)marginals



Second-order
moment matrix

$$\begin{bmatrix}\mu_1 & \mu_{12} & \mu_{13}\\ \mu_{21} & \mu_2 & \mu_{23}\\ \mu_{31} & \mu_{32} & \mu_3\end{bmatrix} = \begin{bmatrix}0.5 & 0.4 & 0.1\\ 0.4 & 0.5 & 0.4\\ 0.1 & 0.4 & 0.5\end{bmatrix}$$

Not positive-semidefinite!

Concave Upper Bound on Entropy

- **Challenge:** Recall that entropy function $-\Phi^*(\mu)$ in terms of *only* μ lacks an explicit form.

- For the Ising model, we have second-order information:

$$\mu_s := \mathbb{E}[x_s] \;\; \forall \; s \in V, \qquad \mu_{st} := \mathbb{E}[x_s x_t] \;\; \forall \; (s,t) \in E$$

Concave Upper Bound on Entropy

- **Challenge:** Recall that entropy function $-\Phi^*(\mu)$ in terms of *only* μ lacks an explicit form.

- For the Ising model, we have second-order information:

$$\mu_s := \mathbb{E}[x_s] \;\; \forall \; s \in V, \qquad \mu_{st} := \mathbb{E}[x_s x_t] \;\; \forall \; (s,t) \in E$$

- **Lemma:** The differential entropy of any $\widetilde{\mathbf{x}}$ is upper-bounded by the covariance-matched Gaussian as follows:

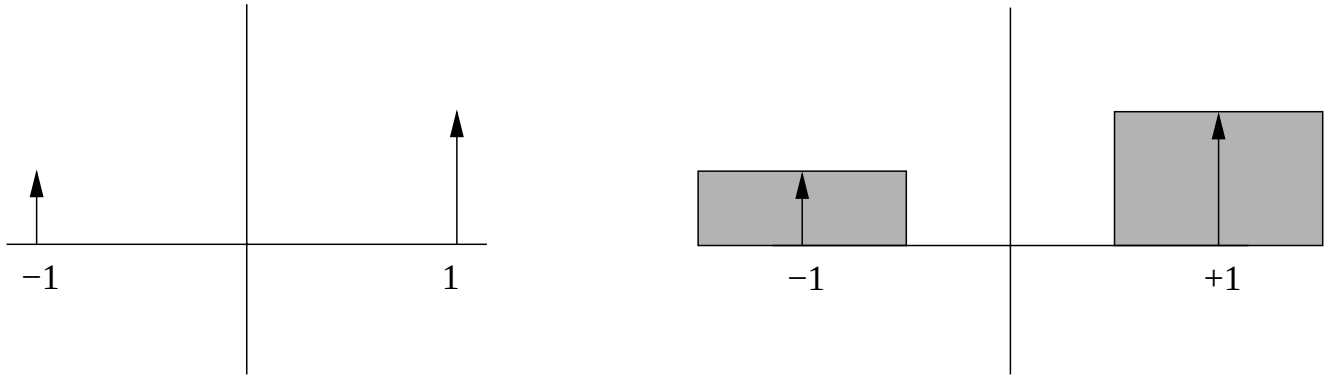
$$h(\widetilde{\mathbf{x}}) \quad \leq \quad \frac{1}{2} \log \det \operatorname{cov}(\widetilde{\mathbf{x}}) + \frac{n}{2} \log(2\pi e)$$

(Cover & Thomas, 1990)

Note: The *differential entropy* $h(\widetilde{\mathbf{x}}) := -\int p(\widetilde{\mathbf{x}}) \log p(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}}$.

From Discrete to Differential Entropy

- Need to relate the discrete entropy $H(\mathbf{x})$ to the differential entropy
- **Solution:** "Smoothing" by addition of independent randomness. Let $\mathbf{u} \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ be uniformly distributed.



Discrete $\mathbf{x}$ Cts. $\widetilde{\mathbf{x}} = \mathbf{x} + \mathbf{u}$

- **Lemma:** The differential entropy of the smoothed version $\widetilde{\mathbf{x}}$ is matched to the discrete entropy of $\mathbf{x}$. (I.e., $H(\mathbf{x}) = h(\widetilde{\mathbf{x}})$.)

Log-determinant Relaxation

- Consider an outer bound $\mathrm{OUT}(K_n)$ that satisfies:

$$\mathrm{MARG}(K_n) \quad \subseteq \quad \mathrm{OUT}(K_n) \; \subseteq \; \mathrm{SDEF}_1(K_n)$$

- Let $M_1(\mu) \in \mathrm{OUT}(K_n)$ be a covariance matrix. Note that constraints imply that $M_1[\mu] \succeq 0$.

Log-determinant Relaxation

- Consider an outer bound $\mathrm{OUT}(K_n)$ that satisfies:

$$\mathrm{MARG}(K_n) \quad \subseteq \quad \mathrm{OUT}(K_n) \; \subseteq \; \mathrm{SDEF}_1(K_n)$$

- Let $M_1(\mu) \in \mathrm{OUT}(K_n)$ be a covariance matrix. Note that constraints imply that $M_1[\mu] \succeq 0$.

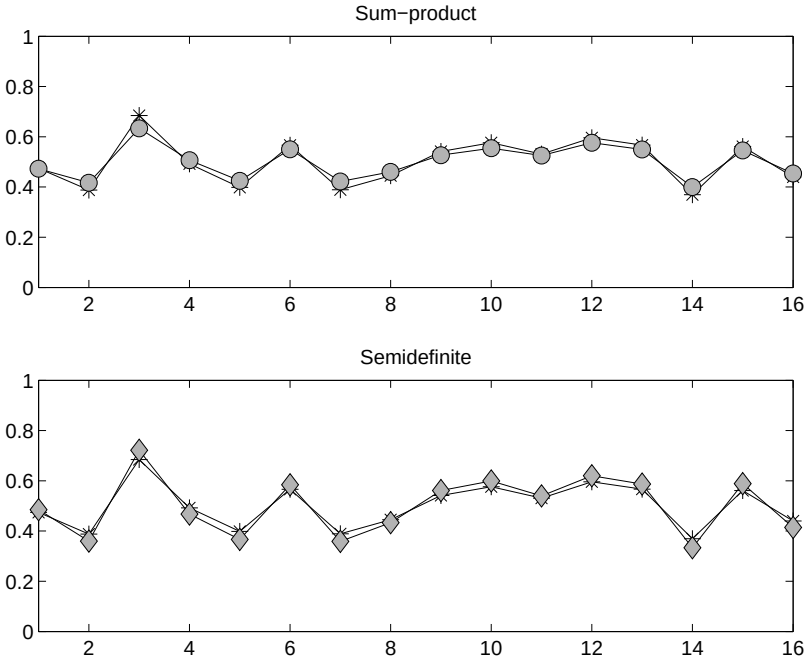
- **Log-det relaxation:** For any such $\mathrm{OUT}(K_n)$, $\Phi(\theta)$ is upper bounded by:

$$\max_{\mu \in \mathrm{OUT}(K_n)} \left\{ \langle \theta, \, \mu \rangle + \frac{1}{2} \log \det \left[M_1(\mu) + \frac{1}{3} \, \mathrm{blkdiag}[0, I_n] \right] \right\} + \frac{n}{2} \log(\frac{\pi e}{2})$$

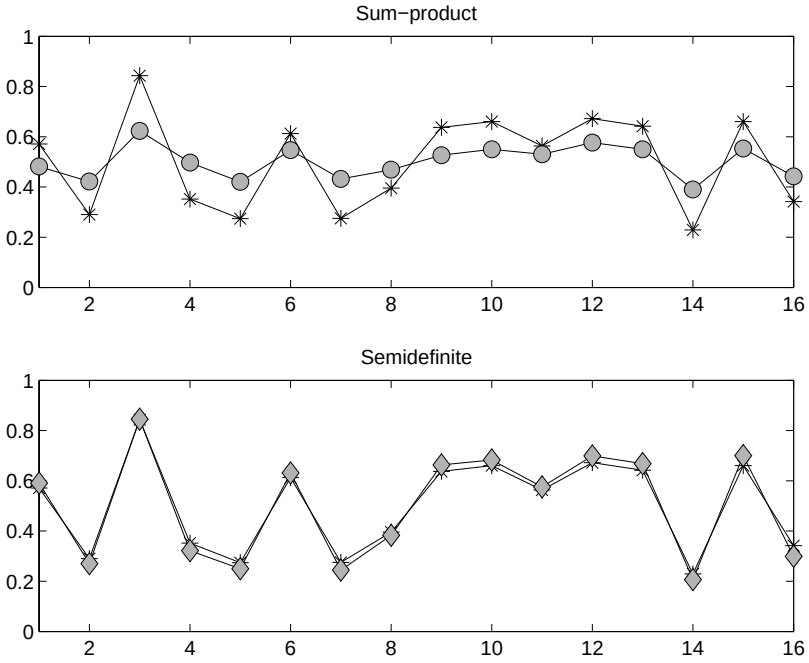
Note: Such a log-det problem with LMI constraints can be solved efficiently by an interior-point method. (Vandenberghe, Boyd, & Wu, 1998)

Simple Illustration

Binary vector $\mathbf{x}$ on complete graph K_{16}.

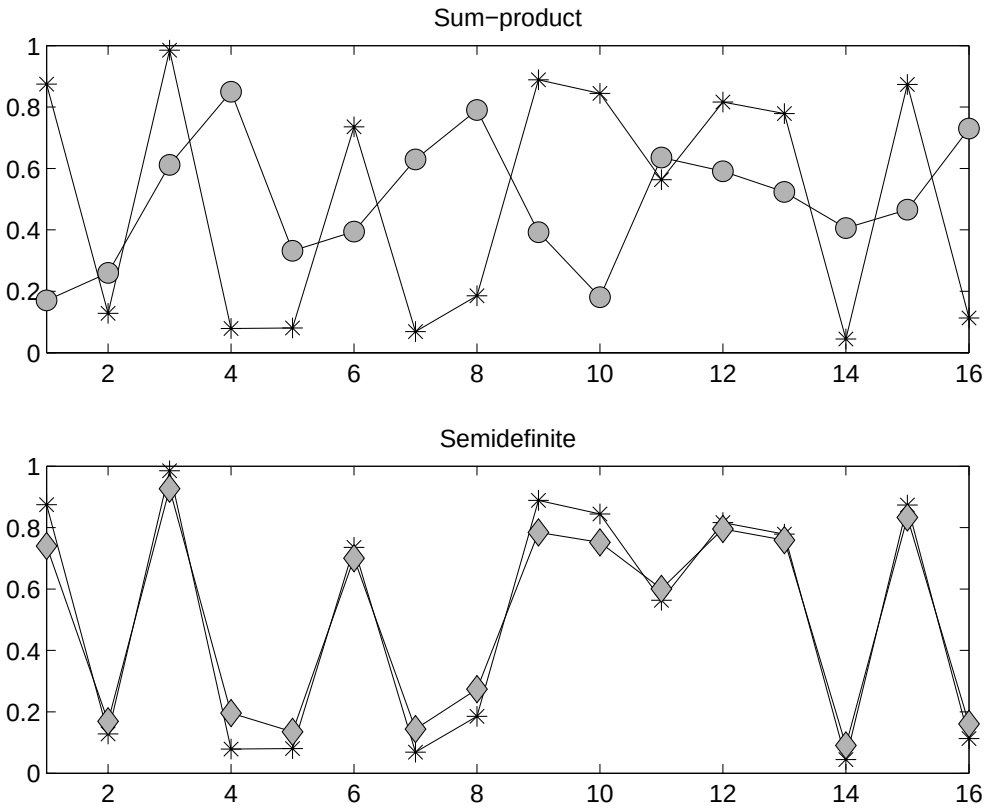


(a) Weak



(b) Medium

Strong Couplings



(c) Strong

Results for Fully Connected Graph

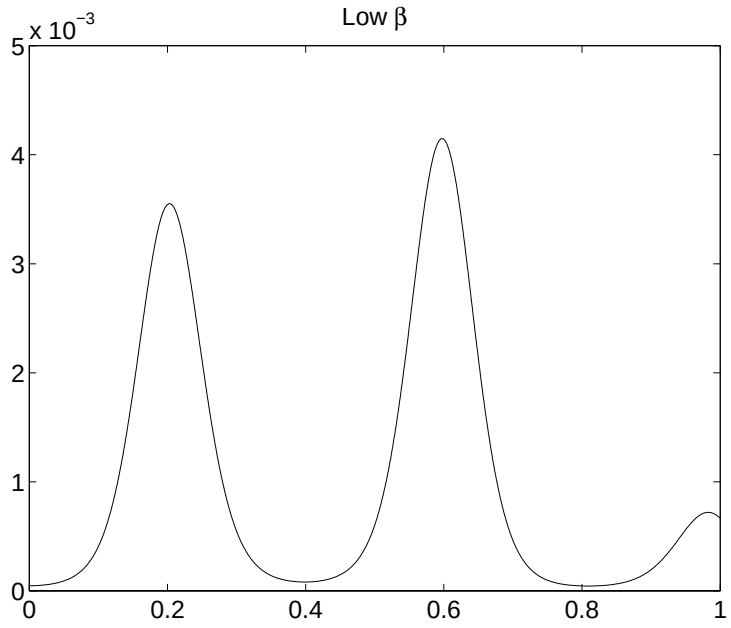
Problem type		Method			
		Sum-product		Log-determinant	
Coup.	Str.	Mean $\pm$ std	Range	Mean $\pm$ std	Range
$-$	Weak	0.037 ± 0.015	$[0.01, 0.10]$	0.020 ± 0.005	$[0.01, 0.03]$
$-$	Strong	0.071 ± 0.032	$[0.03, 0.20]$	0.018 ± 0.005	$[0.01, 0.04]$
$+/-$	Weak	0.004 ± 0.005	$[0.00, 0.04]$	0.020 ± 0.005	$[0.01, 0.03]$
$+/-$	Strong	0.055 ± 0.060	$[0.01, 0.31]$	0.021 ± 0.010	$[0.01, 0.06]$
$+$	Weak	0.024 ± 0.016	$[0.00, 0.08]$	0.027 ± 0.015	$[0.01, 0.06]$
$+$	Strong	0.435 ± 0.196	$[0.08, 0.86]$	0.033 ± 0.019	$[0.01, 0.09]$

Zero Temperature Limit

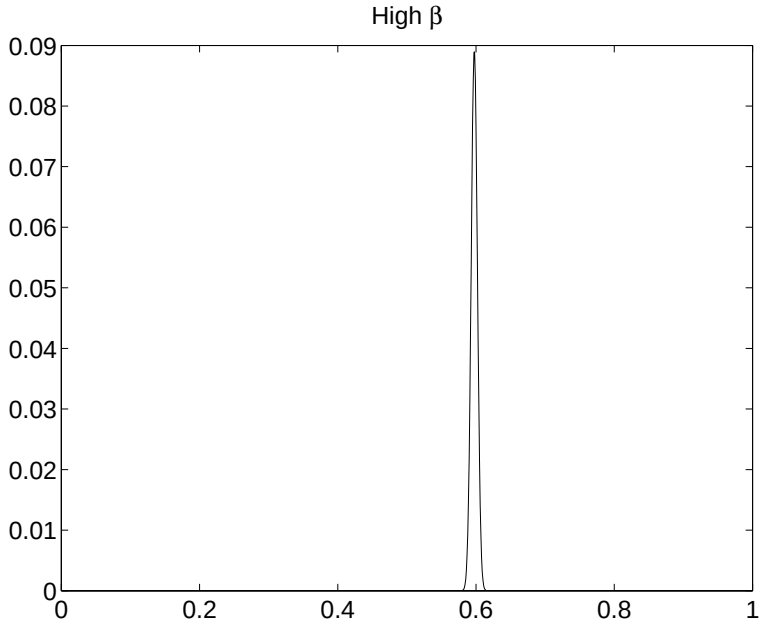
- For fixed θ, consider the 1-parameter family of distributions:

$$p(\mathbf{x}; \beta\theta) \quad = \quad \exp\left\{\beta\langle\theta,\, \boldsymbol{\phi}(\mathbf{x})\rangle - \Phi(\beta\theta)\right\}$$

- Here β should be viewed as inverse "temperature".



(a) Low β



(b) High β

Link to SDP Relaxation for Integer Programming

- For all $\beta > 0$, $\frac{1}{\beta}\Phi(\beta\theta)$ is upper bounded by the following:

$$\frac{1}{\beta} \max_{\mu \in \mathrm{OUT}(K_n)} \left\{ \langle \beta\theta,\, \mu \rangle + \frac{1}{2} \log\det \left[M_1(\mu) + \frac{1}{3}\,\mathrm{blkdiag}[0, I_n] \right] \right\} + C$$

- Taking limits as $\beta \to \infty$ corresponds to computing a recession function. (Rockafellar, 1970)

- Result is a well-known SDP relaxation for integer programming:

$$\max_{\mathbf{x} \in \mathcal{X}^n} \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle \quad \leq \quad \max_{\mu \in \mathrm{OUT}(K_n)} \langle \theta,\, \mu \rangle$$

- For strong coupling, behavior of log-det relaxation (for inference) approaches that of a SDP relaxation for integer programming.

Conclusions

- For details: **http://www.cs.berkeley.edu/∼jordan**
- See the article "**Graphical Models**" for a gentle introduction that covers much of the material discussed in Monday's lecture
- See the technical report "**Hierarchical Dirichlet Processes**" for the material covered in Tuesday's lecture
- See the technical report "**Graphical Models, Exponential Families and Variational Inference**" for the material covered in Wednesday's lecture