Graphical Models, Variational Inference and Nonparametric Priors

Michael I. Jordan

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

http://www.cs.berkeley.edu/∼jordan

Joint work with: Martin Wainwright and David Blei

Graphical Models Past, Present and Future

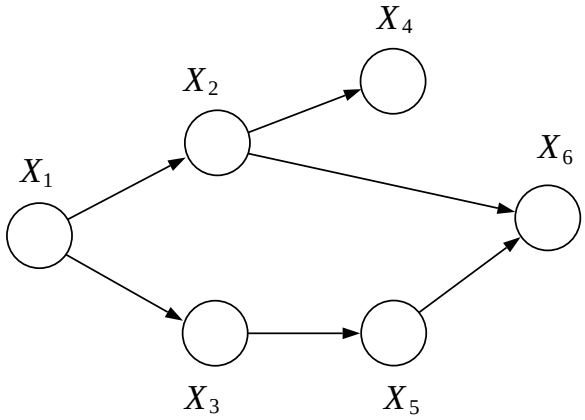
- Some virtues
  - toolbox for modular design of probabilistic systems
  - simple algorithms—often surprisingly effective
  - unifying framework
- Some limitations
  - restriction to parametric models
  - convergence/accuracy of algorithms?
- *Need to strengthen the links with optimization theory and statistics*

Outline

- Some examples
- Exponential families
- Variational representation of exponential families
- Variational relaxations (non-convex and convex)
- Nonparametric methods (Chinese restaurant process)

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:

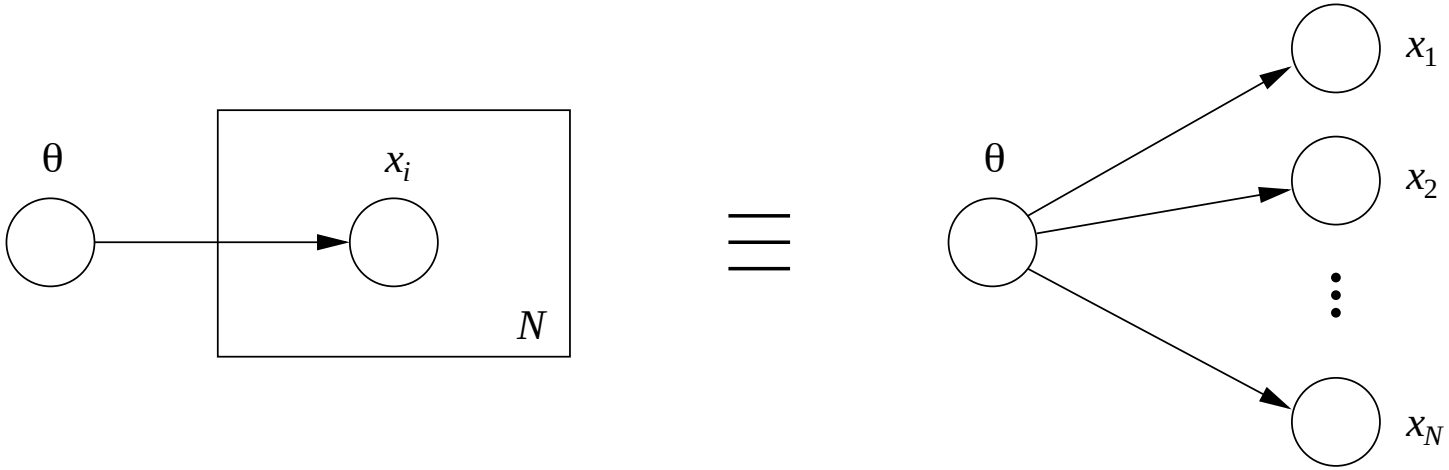


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

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

Plates

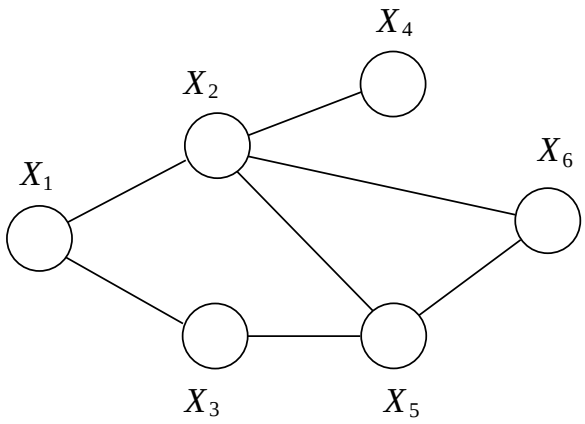
- A *plate* is a "macro" that allows subgraphs to be replicated:



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

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:



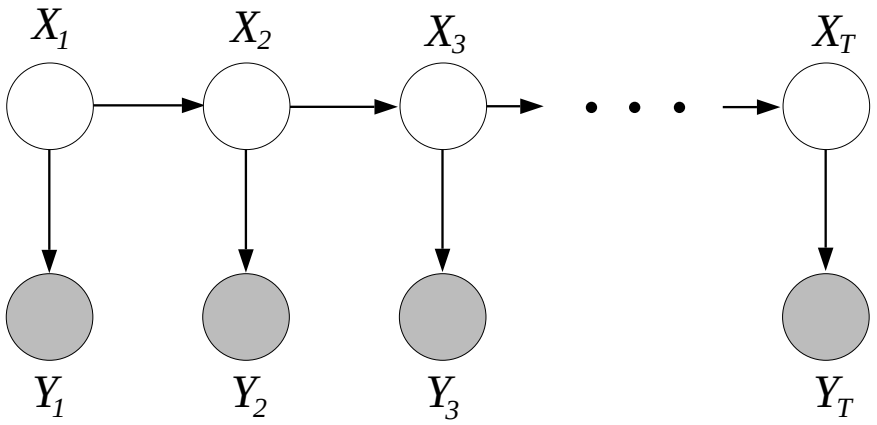
- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the set of cliques defined by the edges $\mathcal{E}$:

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

Examples

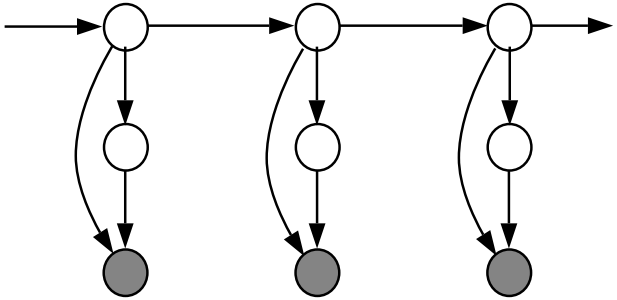
- Hidden Markov models
- Phylogenies
- Hidden Markov phylogenies
- Low-density parity check codes
- Medical diagnosis
- Latent Dirichlet allocation models

Hidden Markov Models

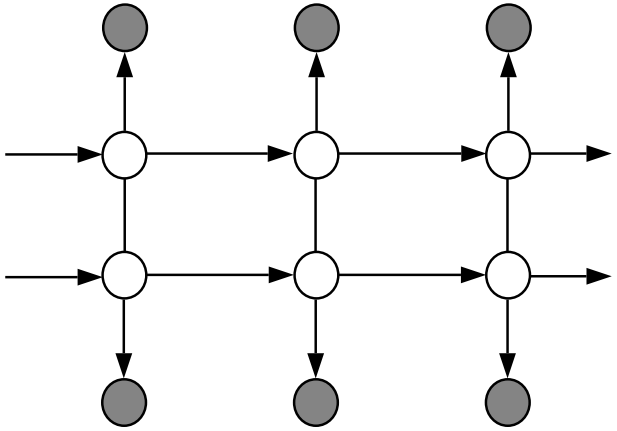


- Generally wish to compute $p(x_i \mid y_1, y_2, \ldots, y_T)$
- For discrete X_i, widely used in speech modeling, bioinformatics, etc., to represent segments of a string
- For continuous X_i, this is the Kalman filter/smoother

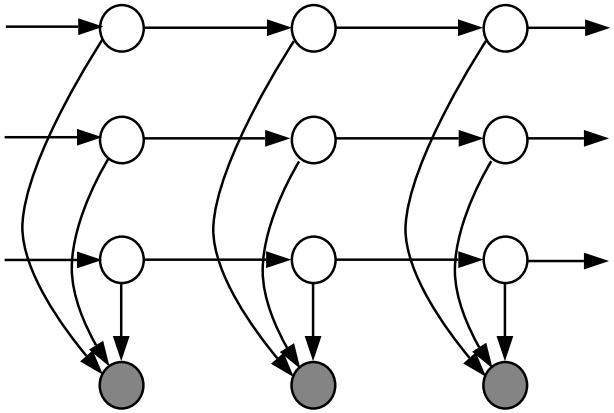
Hidden Markov Model Variations



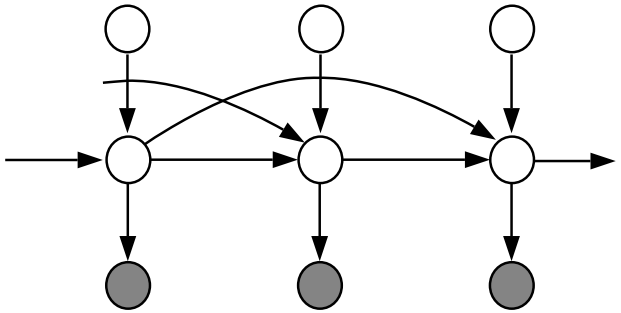
(a)



(b)

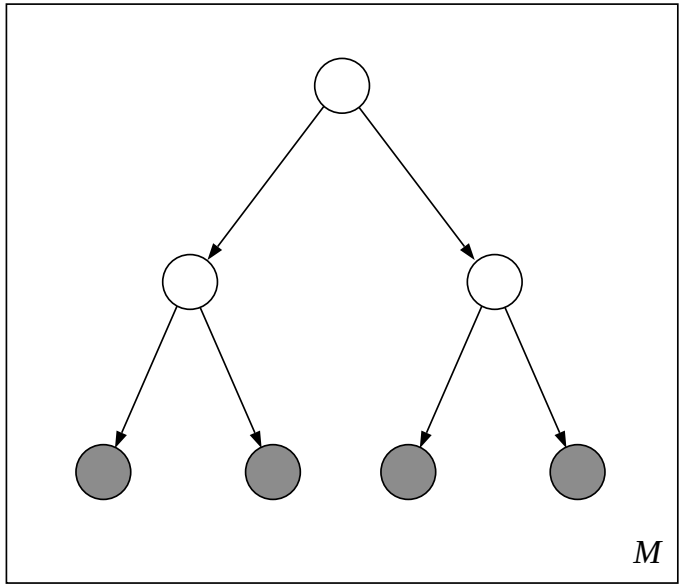


(c)



(d)

Phylogenies



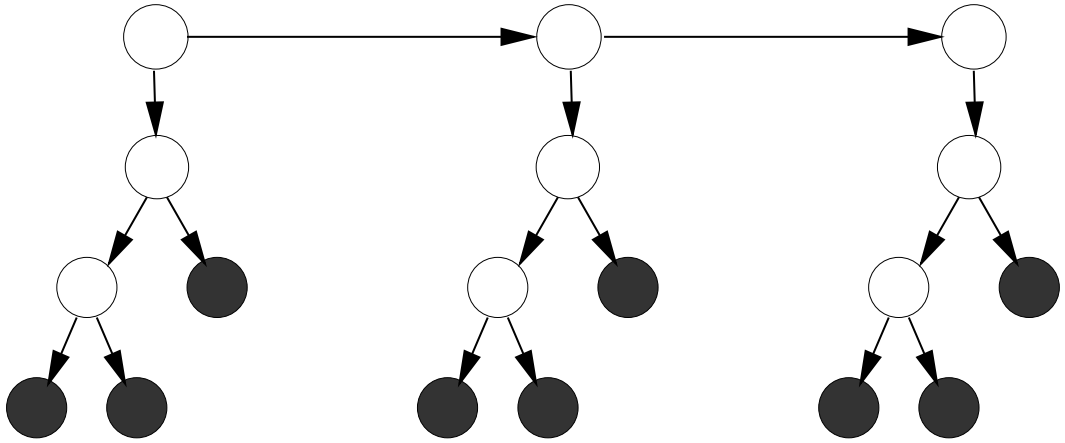
- The shaded nodes represent the observed nucleotides at a given site for a set of organisms
- Site independence model (note the plate)
- The unshaded nodes represent putative ancestral nucleotides
- Computing the likelihood involves summing over the unshaded nodes

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

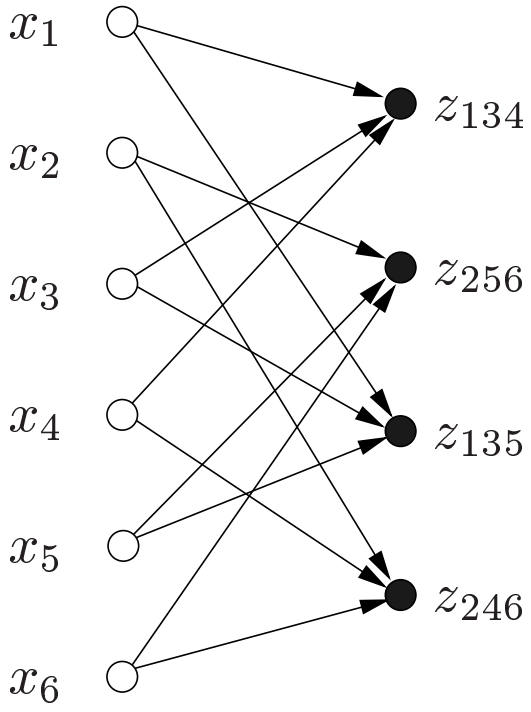
(McAuliffe, Pachter, & Jordan, 2003)



- 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

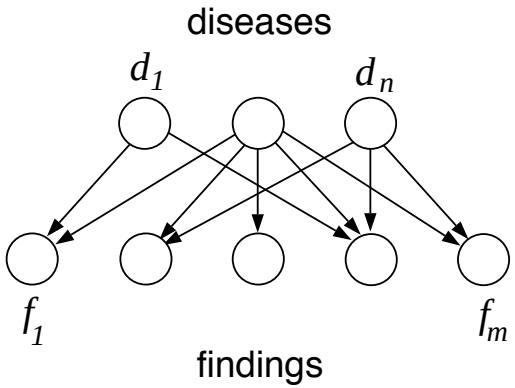


- The x_i denote the unknown message; the z_{ijk} denote the parity checks
- Compute the maximum a posteriori message
  - exact algorithms and MCMC algorithms are not viable
  - a variational algorithm ("max-product algorithm") is used instead, yielding impressive results

Quick Medical Reference (QMR) model

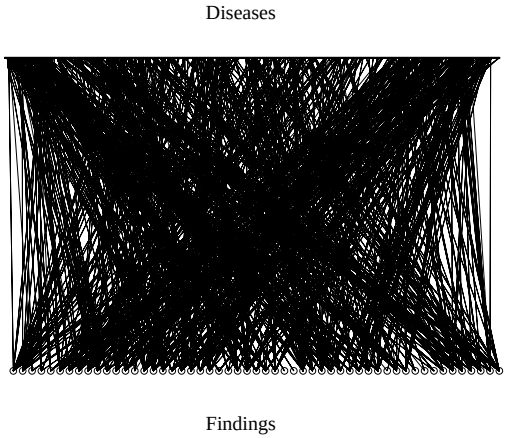
(Jaakkola & Jordan, 1999)

- 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

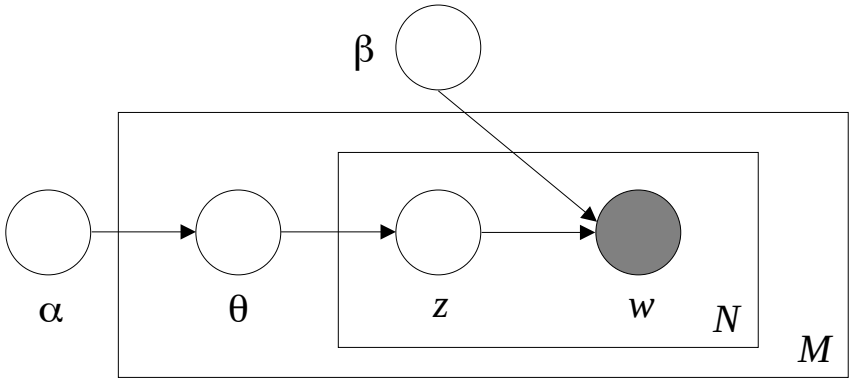
Probabilistic Modeling of Documents

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

Latent Dirichlet Allocation (LDA) Model

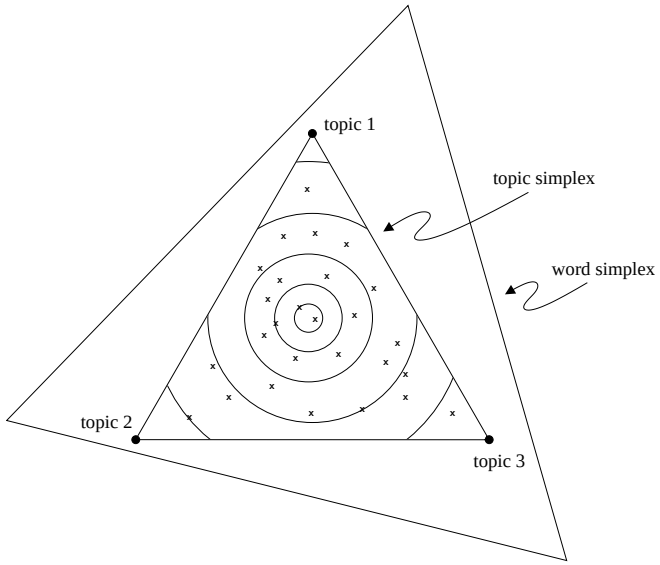
(Blei, Jordan, & Ng 2003)



- *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

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

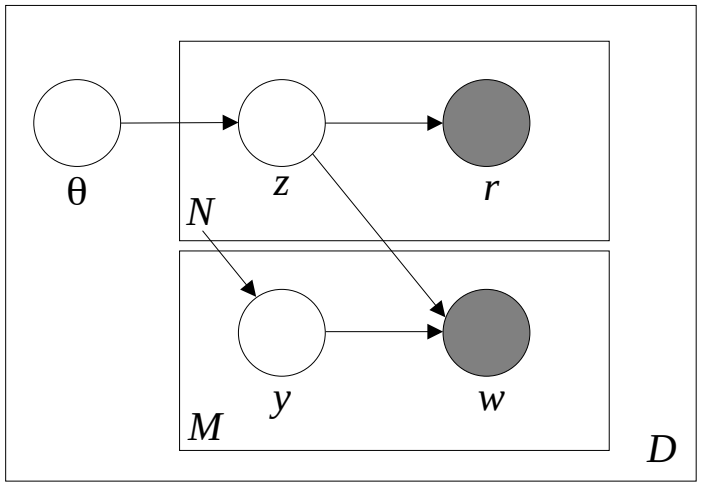
Probabilistic Modeling of Documents/Images

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

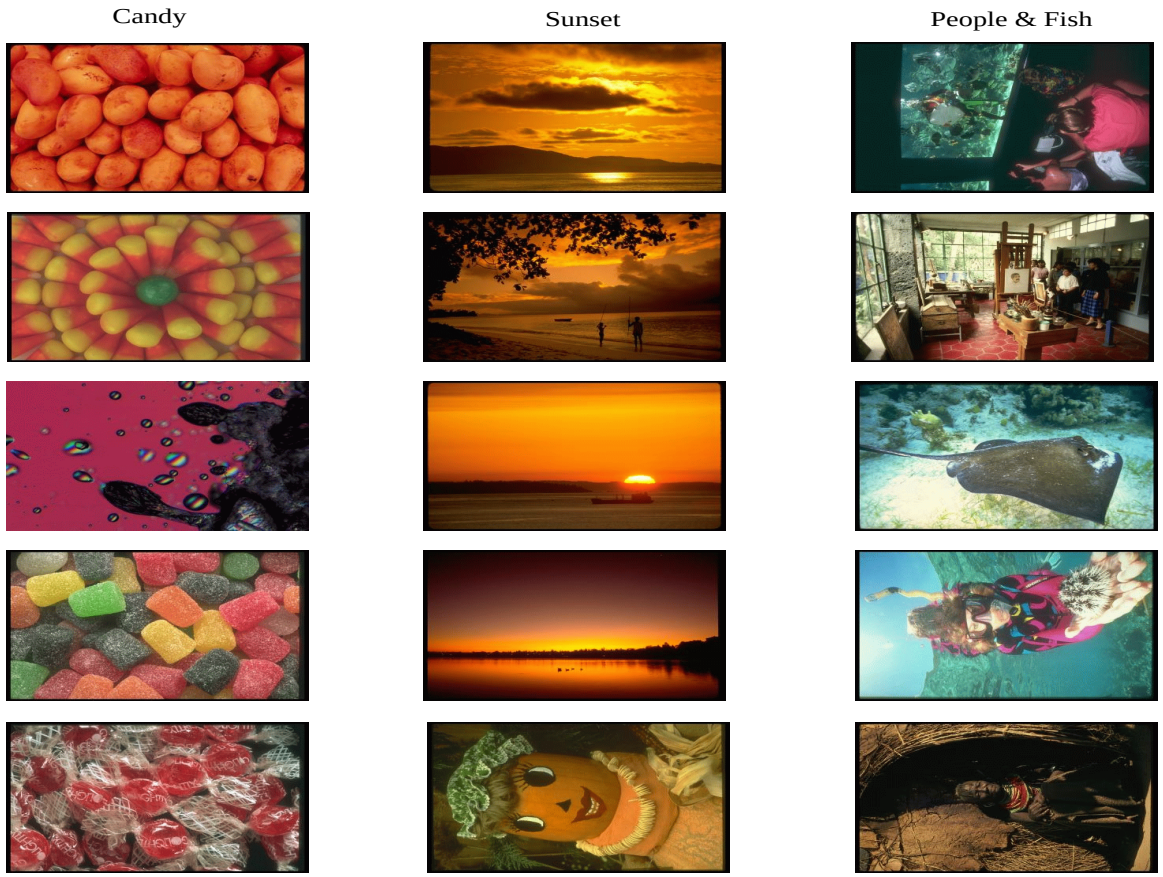
Corr–LDA
scotland water flowers hills tree

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.)

Text-based image retrieval

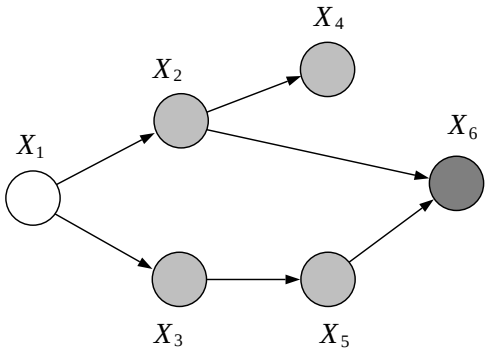


1. Compute $p(\mathbf{w}|\mathbf{r}_g)$ for each image in the test set.

2. Rank the images in order of conditional likelihood.

Inference

- Conditioning



- Marginalization:

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

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

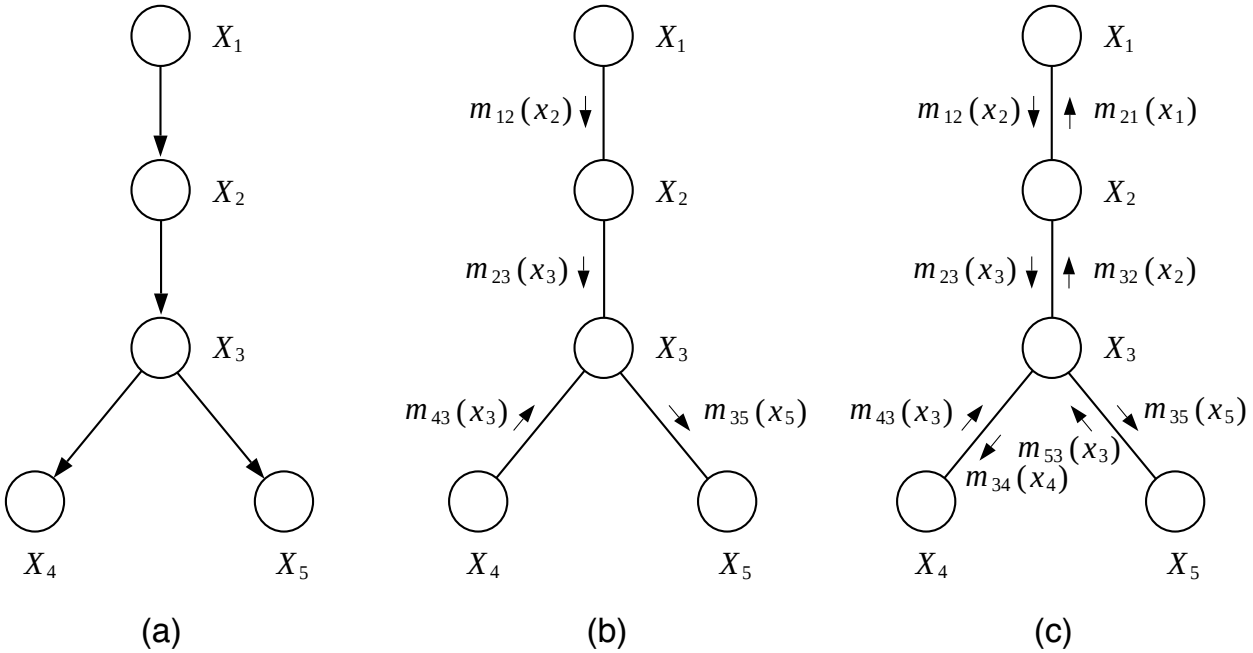
- Posterior probabilities:

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

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)

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

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:

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

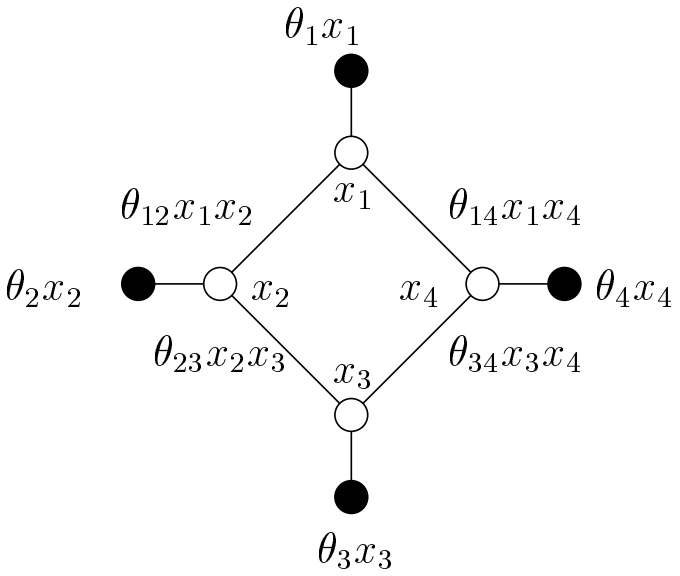
- Useful in a probabilistic setting if there is a concentration in x.
  - the bound is tight for $x \approx \log \mu$
  - turns a nonlinearity into a linearity

- 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:

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

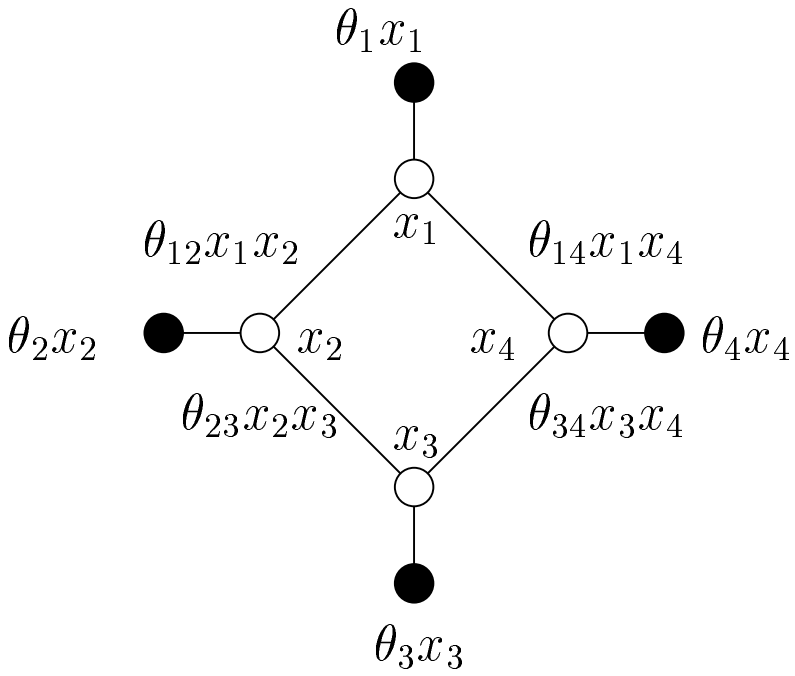
- **Question:** What set do these mean parameters range over?

- 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

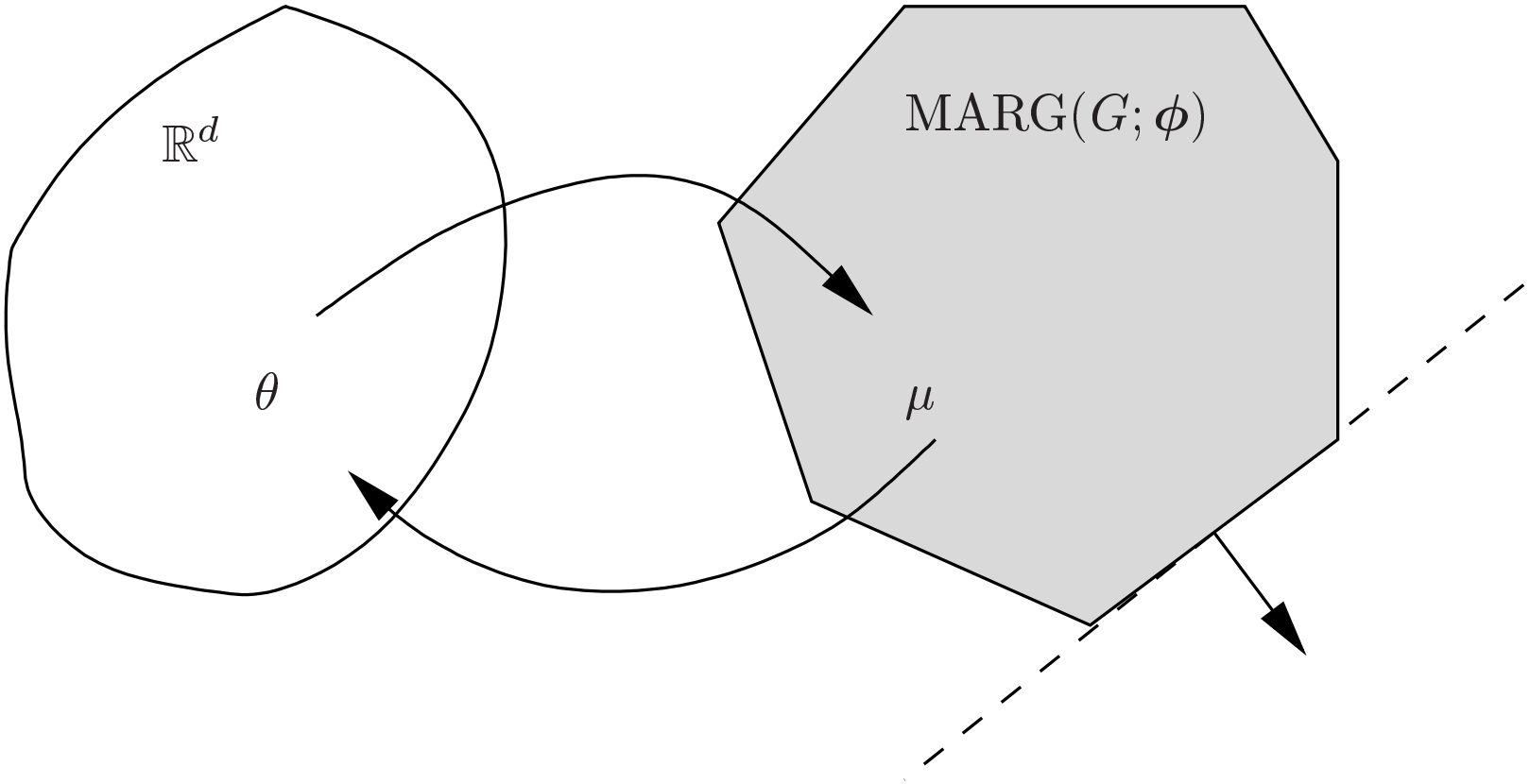


Potentials $\quad \phi = \{x_s \mid s \in V\} \cup \{x_s x_t \mid (s,t) \in E\}$

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:

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

Variational Principle in Terms of Marginals

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

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

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

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

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:

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

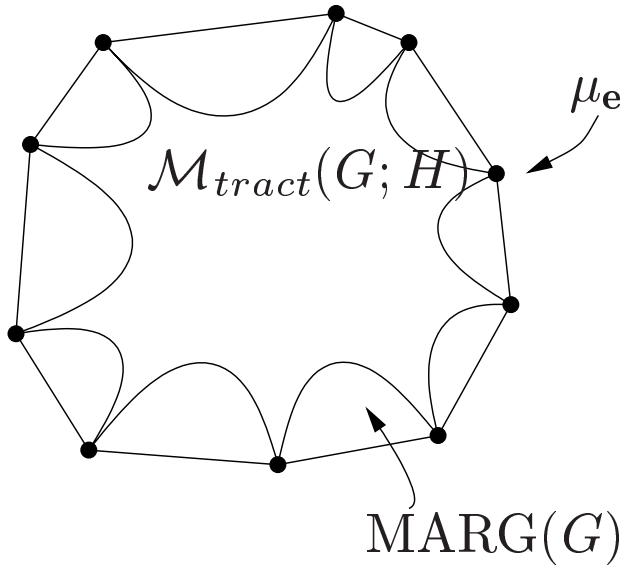
  where $\mathcal{I}(H)$ is the subset of indices associated with cliques in H.

- 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:

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

- Taking derivatives with respect to μ_s yields the naive mean field updates presented earlier.

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:

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

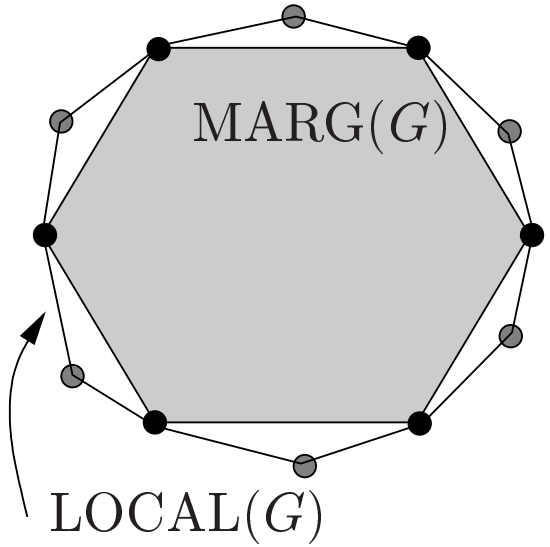
- Denote the corresponding marginals as $\mu_s(x_s)$ and $\mu_{st}(x_s,x_t)$.

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 $-\Phi^*(\mu)$.

- **Solution:**
  - The covariance matrix must be positive semidefinite, thus the cone of all positive semidefinite matrices provides an outer bound for the marginal polytope
  - The differential entropy of any random vector is upper-bounded by the covariance-matched Gaussian
  - So use the log determinant as an upper bound for the entropy

Log-determinant Relaxation

- Let $M_1(\mu) \in \mathrm{OUT}$, where OUT is contained in the semidefinite cone

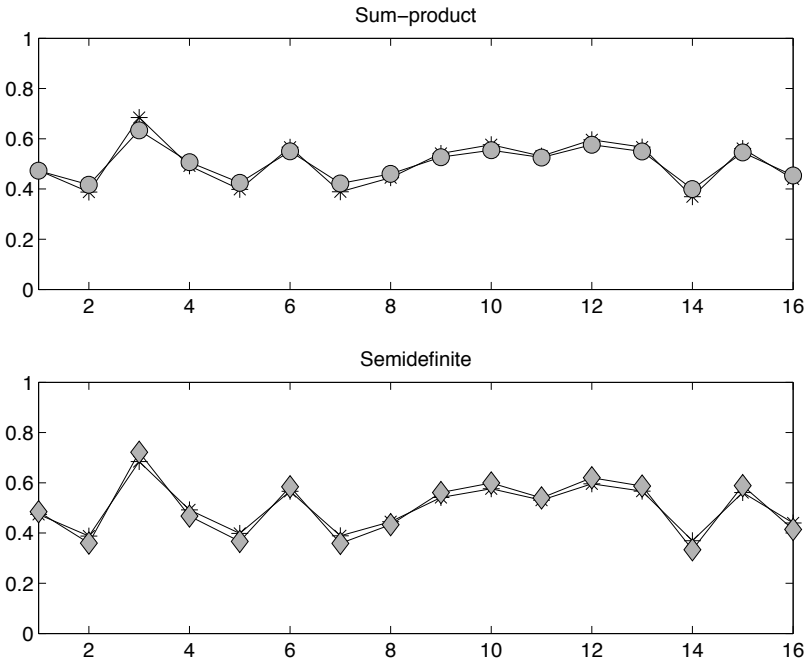
- **Log-det relaxation:** For any such OUT, $\Phi(\theta)$ is upper bounded by:

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

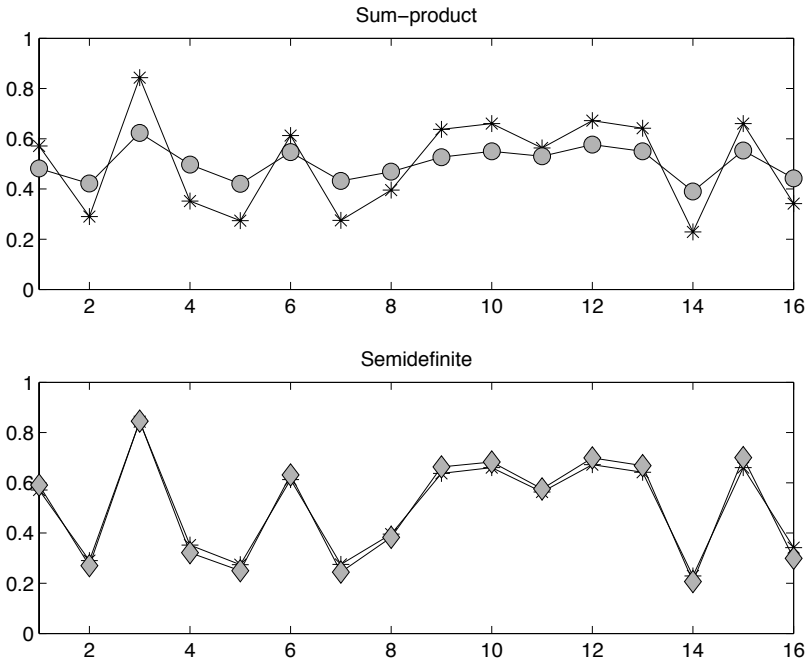
- **Note:** Such a log-det problem with LMI constraints can be solved efficiently by an interior-point method. (Vandenberghe, Boyd, & Wu, 1998)

Simple Illustration

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

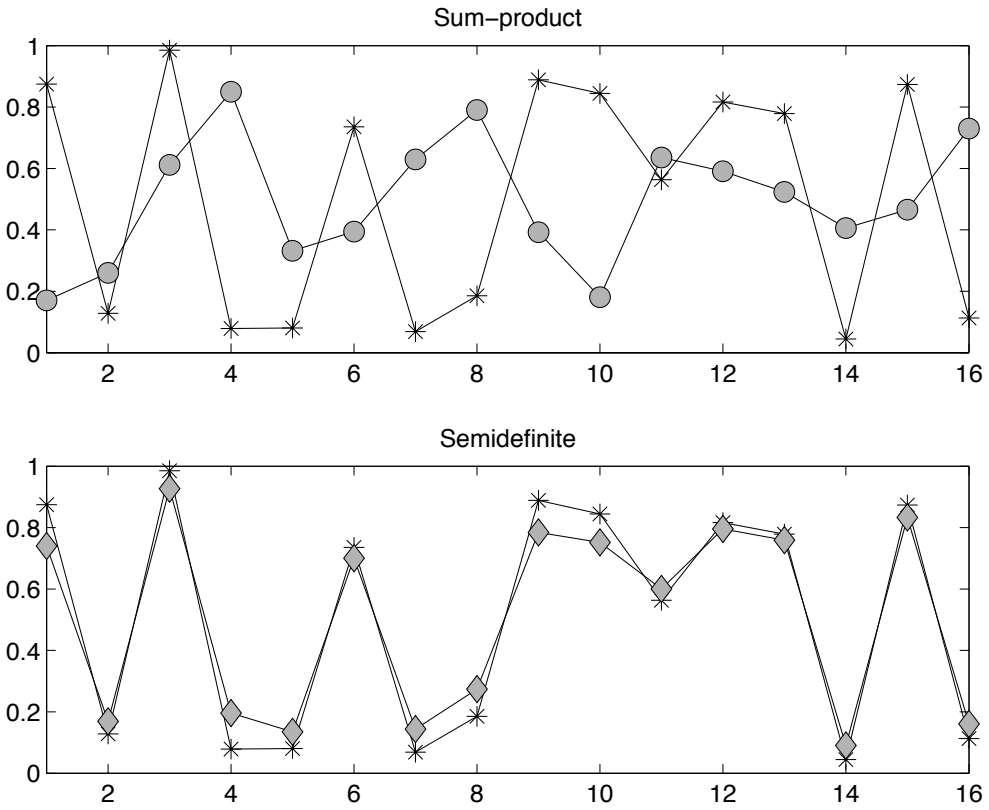


(a) Weak



(b) Medium

Strong Couplings

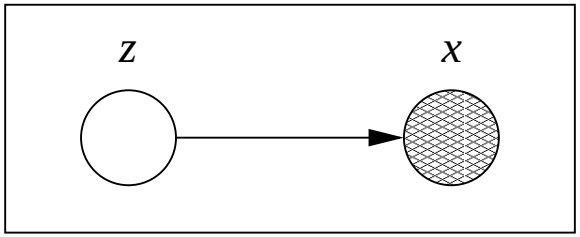


(c) Strong

Bayesian Nonparametrics

- Dirichlet processes, Pólya trees, tail-free distributions
- Urn models, Chinese restaurant process
- Measures on measures—provide flexible representations for structural and parametric uncertainty

Example: Finite Mixture Models



- Probabilistic model for clustering:

$$p(x|\theta) = \sum_{z=1}^{k} p(z|\pi) f(x|z,\beta),$$

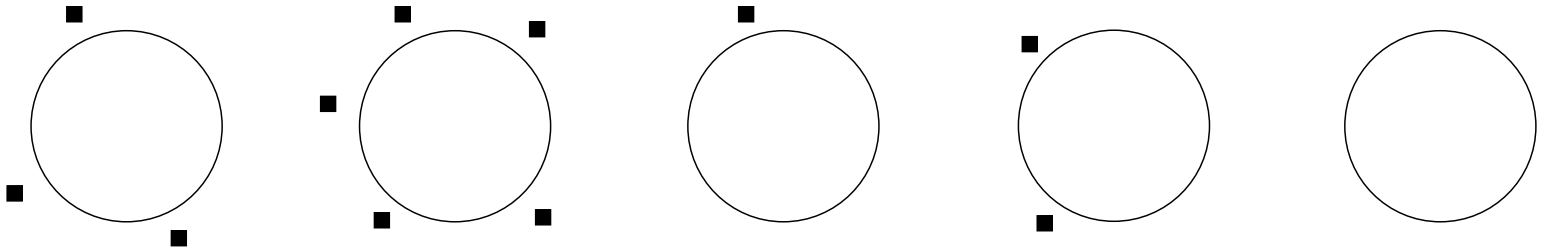
  where $f(x|z,\beta)$ are the *mixture components*, and π are the *mixing proportions*

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

The 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:

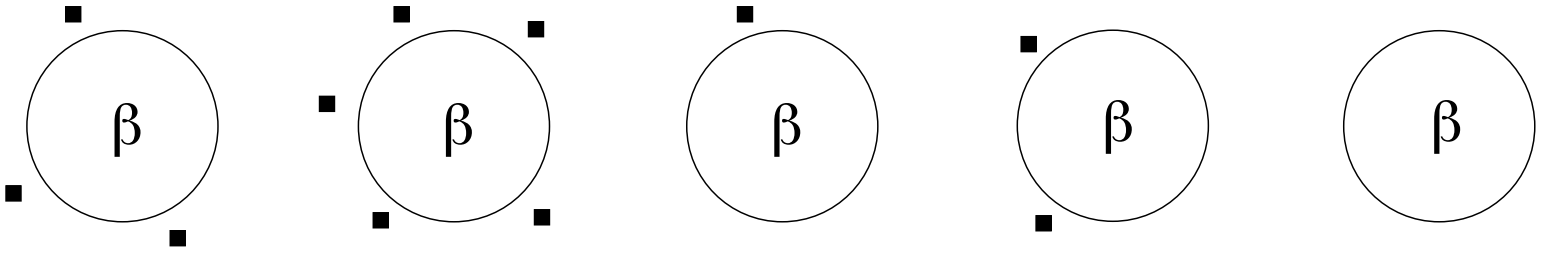
$$\begin{aligned} p(\text{previously occupied table } i) &= \frac{m_i}{\gamma+m-1} \\ p(\text{the next unoccupied table}) &= \frac{\gamma}{\gamma+m-1}, \end{aligned} \tag{3}$$



- Defines an exchangeable distribution on partitions of integers

The CRP and Mixture Models

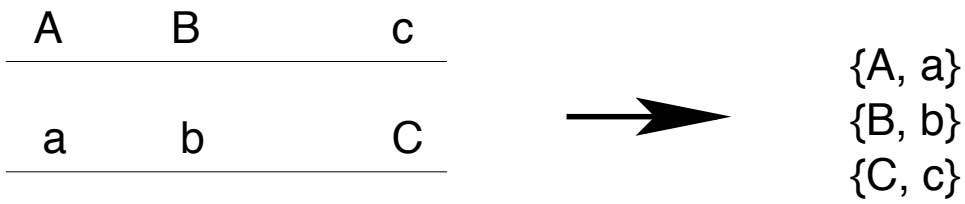
- The customers around a table form a cluster
  - associate a mixture component with each table
  - e.g., for Gaussian mixtures, sample $\beta = (\mu, \Sigma)$ at each table to obtain a mean and covariance matrix for that mixture component



- With this likelihood, and Eq. (3), the CRP yields a posterior probability distribution on the number of mixture components (and on all of the other parameters)

Haplotype modeling with the CRP prior

(Xing, Sharan, & Jordan, 2003)



- 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* of a sample from a population (pairs of alleles with the association of alleles to chromosomes unknown):
  - estimate the underlying haplotypes
  - restore the association of alleles to chromosomes (the *haplotype phase*
- This is a mixture modeling problem

Haplotype Modeling with the CRP Prior (cont.)

- The genotype is a mixture over the population haplotypes:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1)p(h_2)1(h_1 \oplus h_2 = g),$$

  where $1(h_1 \oplus h_2 = g)$ is the indicator function of the event that haplotypes h_1 and h_2 are consistent with g.

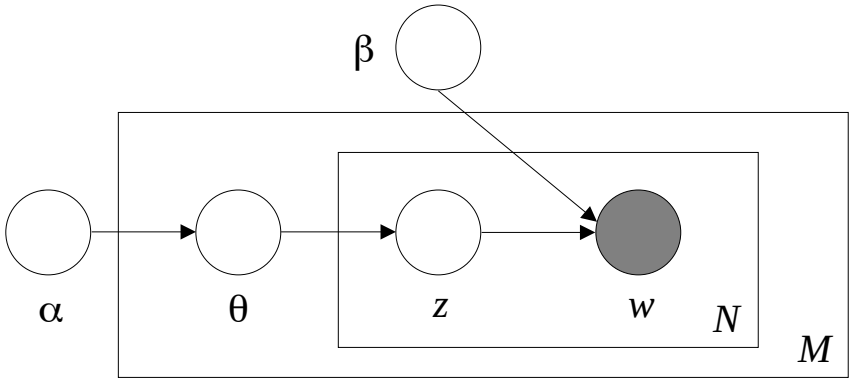
- The number and identity of the haplotypes is unknown
  - use the CRP prior

- Performance on the data of Gabriel, et al (2002):

		CRP			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

Latent Dirichlet Allocation (LDA) Model

(Blei, Jordan, & Ng 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

Issues

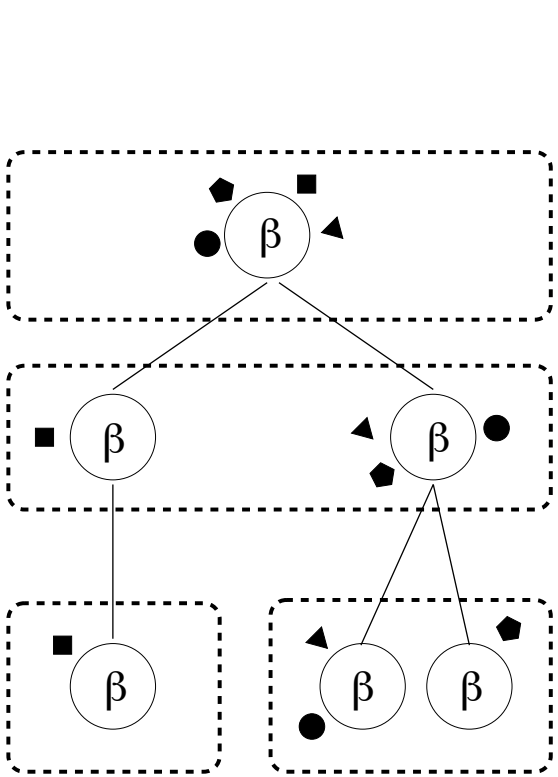
- Can we extend to a hierarchical topic model?
- How to choose the hierarchy?

Nested Chinese Restaurants

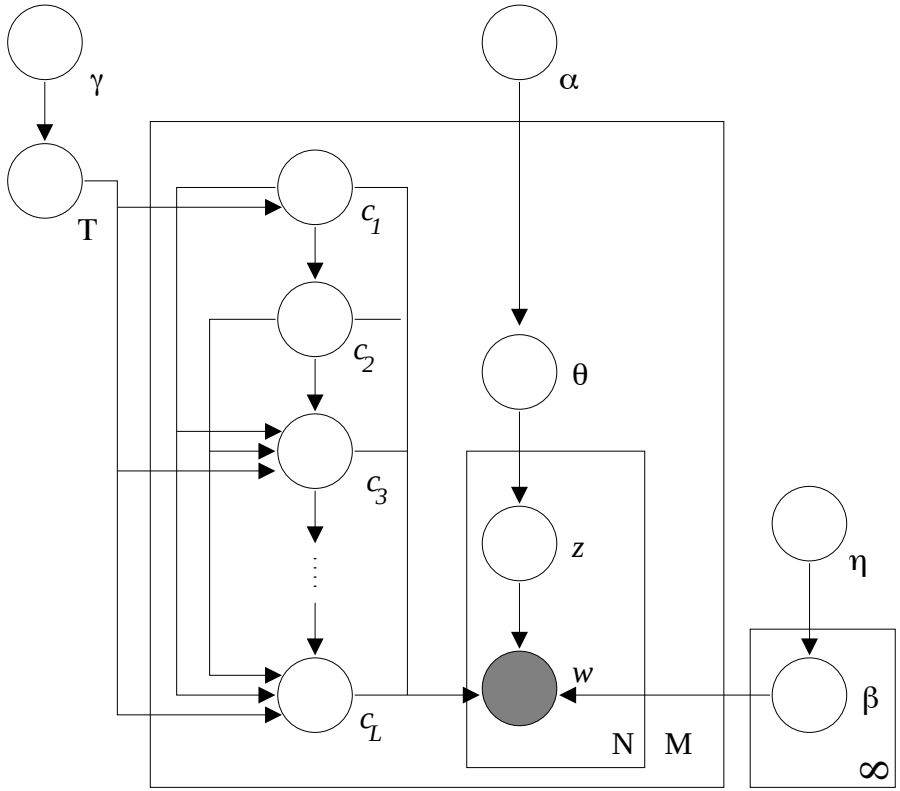
(Blei, Griffiths, Jordan, & Tenenbaum, 2004)

- Let there be an infinite number of infinite-table Chinese restaurants in a city
- One restaurant is determined to be the root restaurant and on each of its infinite tables is a card with the name of another
- On each of the tables in those restaurants are cards that refer to other restaurants, and this structure repeats infinitely
- Thus, the restaurants in the city are organized into an infinitely-branched tree

The Hierarchical Topic Model



(a)



(b)

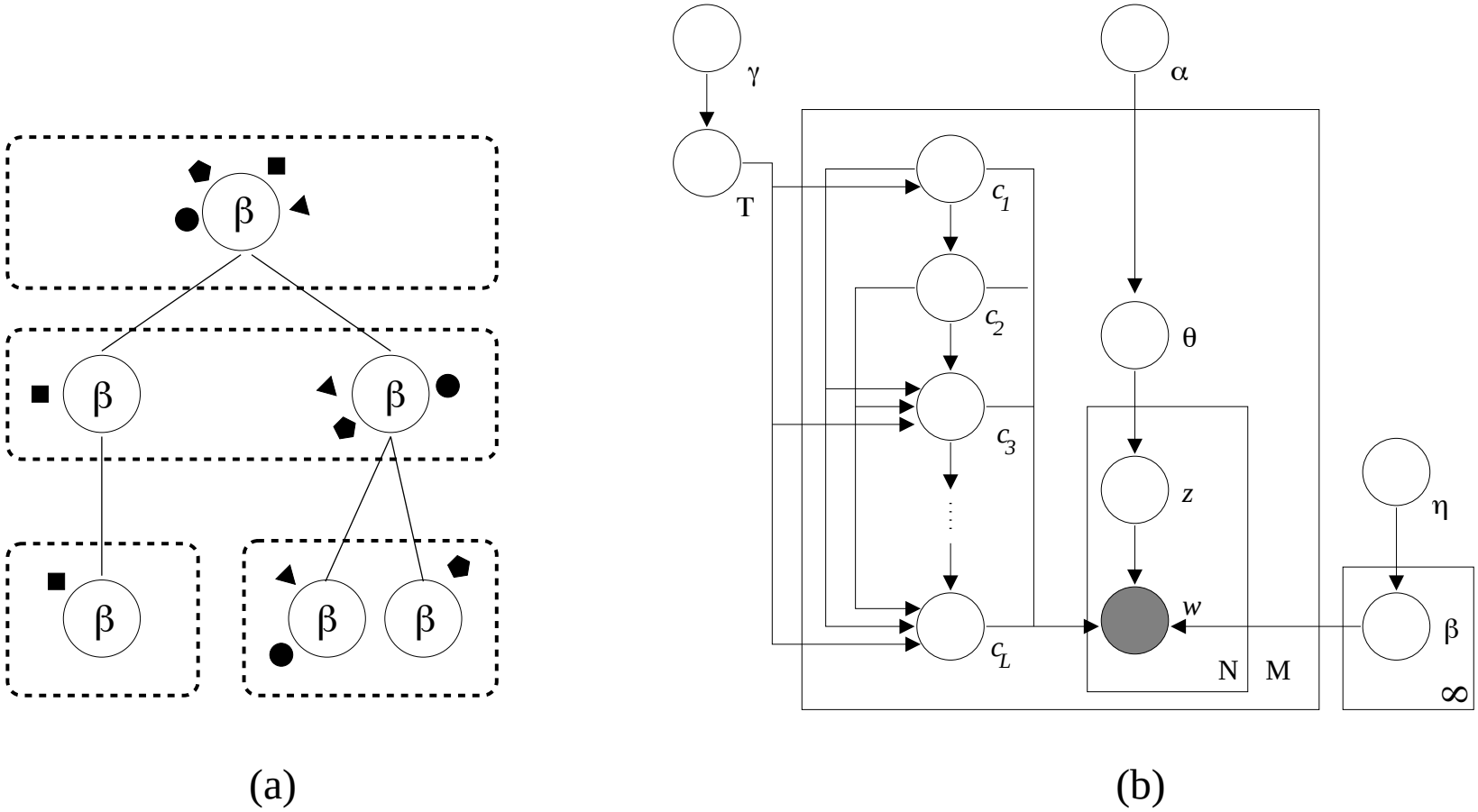
- (a) A sample path of the nested Chinese restaurant process
- (b) The hierarchical latent Dirichlet allocation model

The Nested Chinese Restaurant Process

- A tourist arrives in the city for a culinary vacation
  - on the first evening, he enters the root Chinese restaurant and selects a table using Eq. (3)
  - on the second evening, he goes to the restaurant identified on the first night's table and chooses another table, again using Eq. (3)
  - repeat this process for L days
  - at the end of the trip, the tourist has sat at L restaurants which constitute a path from the root to a restaurant at the Lth level in the infinite tree

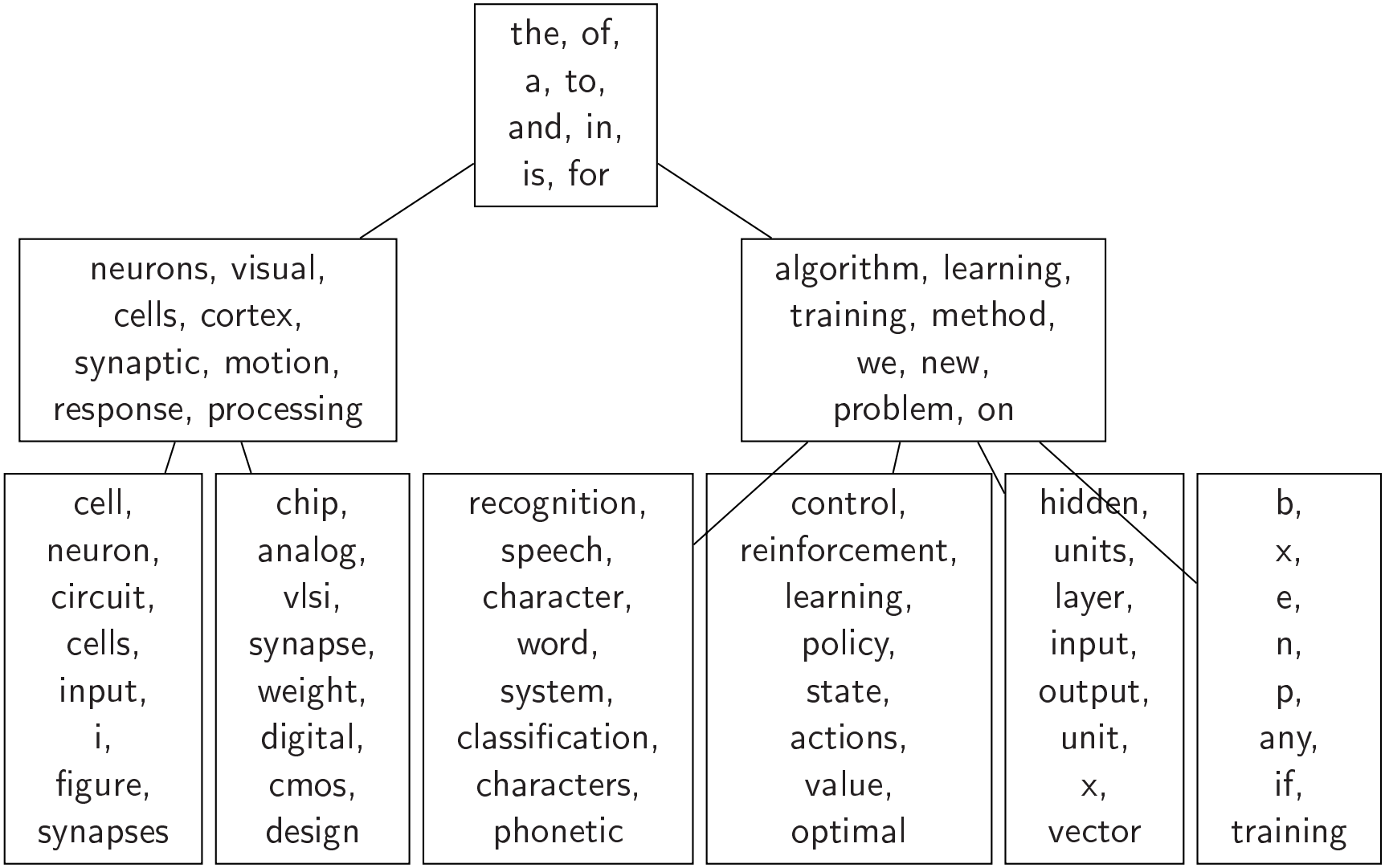
- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree

The Hierarchical Topic Model



- (a) A sample path of the nested Chinese restaurant process
- (b) The hierarchical latent Dirichlet allocation model

Results on the NIPS Abstracts



Summary

- Graphical models provide a general formalism for the design and analysis of complex probabilistic systems

- Variational algorithms convert a marginalization problem into an optimization problem

  – a methodology that is complementary to MCMC
  – many new "relaxations" to be explored

- Much work still to do

  – variational inference algorithms for nonparametric models
  – hierarchical nonparametric models
  – frequentist properties of graphical model algorithms

- For more details, see **http://www.cs.berkeley.edu/∼jordan**