

Undirected Graphical Models

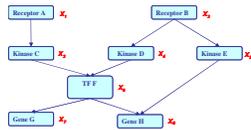
Probabilistic Graphical Models (10-708)

Lecture 2, Sep 17, 2007



Eric Xing

Reading: MJ-Chap. 2,4, and KF-chap5

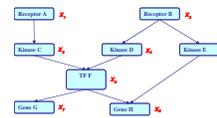


1

Two types of GMs

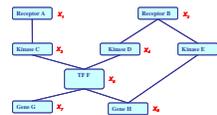
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3/X_1) P(X_4/X_2) P(X_5/X_2) P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6)$$



- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Review: independence properties of DAGs



- Defn: let $I_l(\mathcal{G})$ be the set of local independence properties encoded by DAG \mathcal{G} , namely:

$$\{ X_i \perp\!\!\!\perp \text{NonDescendants}(X_i) \mid \text{Parents}(X_i) \}$$
- Defn: A DAG \mathcal{G} is an **I-map** (independence-map) of \mathcal{P} if $I_l(\mathcal{G}) \subseteq I(\mathcal{P})$
- A fully connected DAG \mathcal{G} is an I-map for any distribution, since $I_l(\mathcal{G}) = \emptyset \subseteq I(\mathcal{P})$ for any \mathcal{P} .
- Defn: A DAG \mathcal{G} is a minimal I-map for \mathcal{P} if it is an I-map for \mathcal{P} , and if the removal of even a single edge from \mathcal{G} renders it not an I-map.
- A distribution may have several minimal I-maps
 - Each corresponding to a specific node-ordering

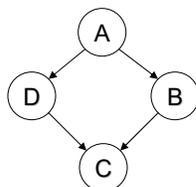
Eric Xing

3

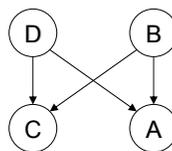
P-maps



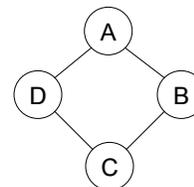
- Defn: A DAG \mathcal{G} is a **perfect map** (P-map) for a distribution \mathcal{P} if $I(\mathcal{P}) = I(\mathcal{G})$.
- Thm: not every distribution has a perfect map as DAG.
 - Pf by counterexample. Suppose we have a model where $A \perp C \mid \{B, D\}$, and $B \perp D \mid \{A, C\}$.
This cannot be represented by any Bayes net.
 - e.g., BN1 wrongly says $B \perp D \mid A$, BN2 wrongly says $B \perp D$.



BN1



BN2

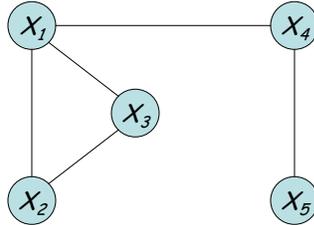


MRF

Eric Xing

4

Undirected graphical models



- Pairwise (non-causal) relationships
- Can write down model, and score specific configurations of the graph, but no explicit way to generate samples
- Contingency constrains on node configurations

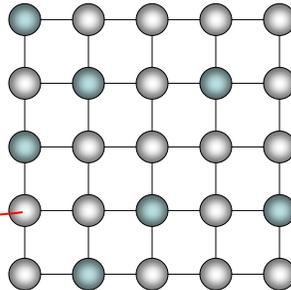
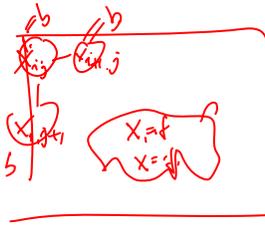
Eric Xing

5

Canonical examples



- The grid model



Ising Model.

$p(X)$

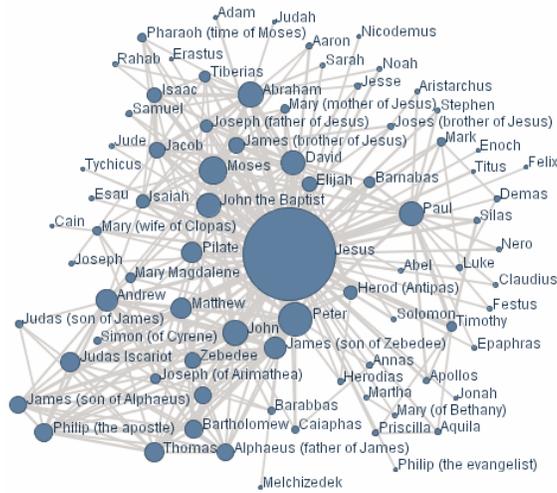
$X_{ij} \quad X_{i+1, j+1}$ } $\begin{matrix} +1 \quad -1 \\ 0 \quad -1 \end{matrix}$

- Naturally arises in image processing, lattice physics, etc.
- Each node may represent a single "pixel", or an atom
 - The states of adjacent or nearby nodes are "coupled" due to pattern continuity or electro-magnetic force, etc.
 - Most likely joint-configurations usually correspond to a "low-energy" state

Eric Xing

6

Social networks

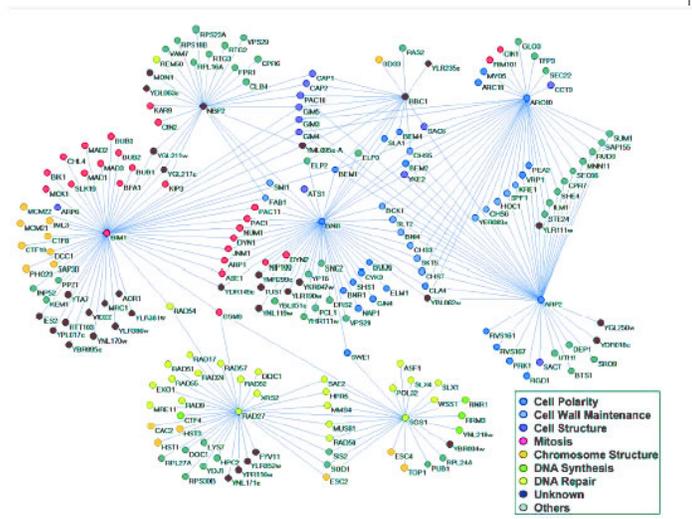


The New Testament Social Networks

Eric Xing

7

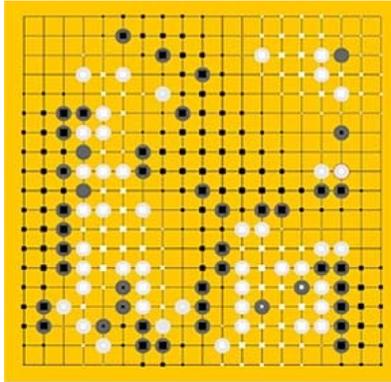
Protein interaction networks



Eric Xing

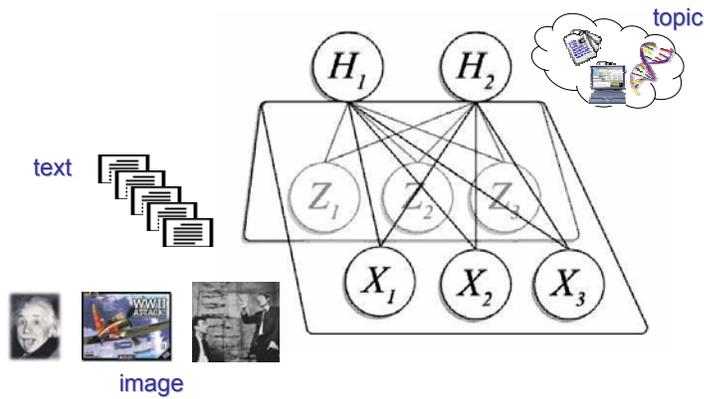
8

Modeling Go



This is the middle position of a Go game.
Overlaid is the estimate for the probability of becoming black or white for every intersection.
Large squares mean the probability is higher.

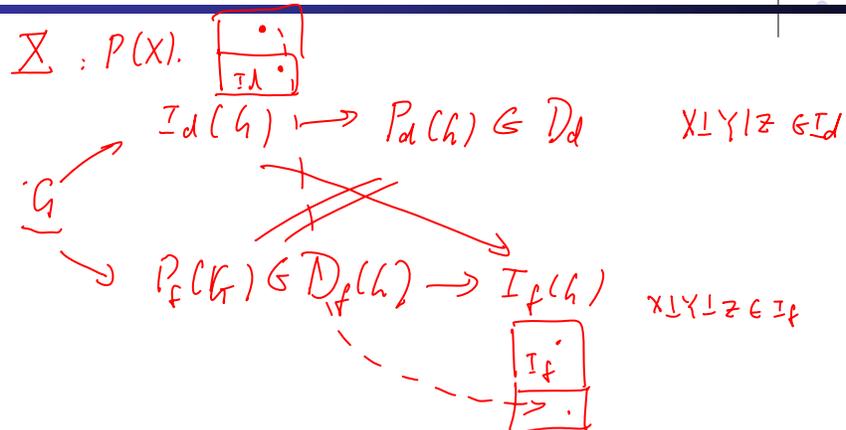
Information retrieval





- First hw out today! Start now!
- Auditing students: please fill out forms
- Recitation:
- questions:

Distributional equivalence and I-equivalence

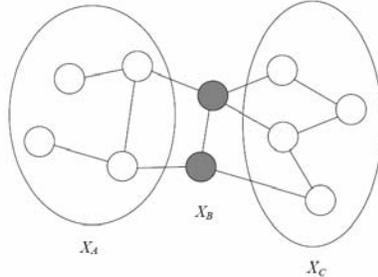


- All independence in $I_d(G)$ will be captured in $I_f(G)$, is the reverse true?
- Are "not-independence" from G all honored in P_f ?

Global Markov Independencies



- Let H be an undirected graph:



- B **separates** A and C if every path from a node in A to a node in C passes through a node in B : $\text{sep}_H(A; C|B)$
- A probability distribution satisfies the **global Markov property** if for any disjoint A, B, C , such that B separates A and C , A is independent of C given B : $I(H) = \{A \perp C | B : \text{sep}_H(A; C|B)\}$

Eric Xing

13

Soundness of separation criterion



- The independencies in $I(H)$ are precisely those that are guaranteed to hold for every MRF distribution P over H .
- In other words, the separation criterion is sound for detecting independence properties in MRF distributions over H .

Eric Xing

14

Local Markov independencies

- For each node $X_i \in \mathbf{V}$, there is *unique Markov blanket* of X_i , denoted MB_{X_i} , which is the set of neighbors of X_i in the graph (those that share an edge with X_i)



- Defn (5.5.4):** The *local Markov independencies* associated with H is:

$$I_{\ell}(H): \{X_i \perp \mathbf{V} - \{X_i\} - MB_{X_i} \mid MB_{X_i} : \forall i\},$$

In other words, X_i is independent of the rest of the nodes in the graph given its immediate neighbors

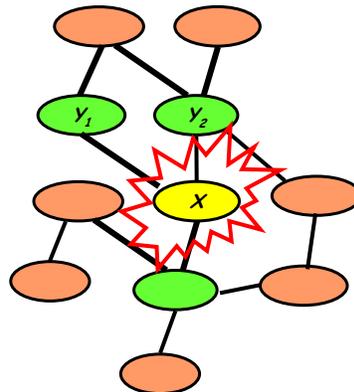
Eric Xing

15

Summary: Conditional Independence Semantics in an MRF

Structure: an *undirected graph*

- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.
- Give **correlations** between variables, but no explicit way to generate samples



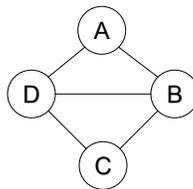
Eric Xing

16

Cliques



- For $G=\{V,E\}$, a complete subgraph (clique) is a subgraph $G'=\{V'\subseteq V,E'\subseteq E\}$ such that nodes in V' are fully interconnected
- A (maximal) clique is a complete subgraph s.t. any **superset** $V''\supset V'$ is not complete.
- A sub-clique is a not-necessarily-maximal clique.



- Example:
 - max-cliques = $\{A,B,D\}, \{B,C,D\}$,
 - sub-cliques = $\{A,B\}, \{C,D\}, \dots \rightarrow$ all edges and singletons

Eric Xing

17

Quantitative Specification



- Defn: an **undirected graphical model** represents a distribution $P(X_1, \dots, X_n)$ defined by an undirected graph H , and a **set** of positive **potential functions** ψ_c associated with cliques of H , s.t.

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \quad (\text{A Gibbs distribution})$$

where Z is known as the partition function:

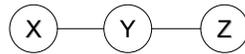
$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$$

- Also known as **Markov Random Fields**, **Markov networks** ...
- The **potential function** can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.

Eric Xing

18

Interpretation of Clique Potentials



$$p(x, y, z) = \frac{1}{Z} \phi(x, y) \phi(y, z)$$

- The model implies $X \perp Z | Y$. This independence statement implies (by definition) that the joint must factorize as:

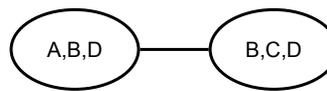
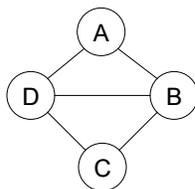
$$p(x, y, z) = p(y)p(x|y)p(z|y)$$

- We can write this as: $p(x, y, z) = p(x, y)p(z|y)$, but $p(x, y, z) = p(x|y)p(z, y)$
 - cannot have all potentials be marginals
 - cannot have all potentials be conditionals
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

Eric Xing

19

Example UGM – using max cliques



$$\psi_c(\mathbf{x}_{124})$$

$$\psi_c(\mathbf{x}_{234})$$



$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

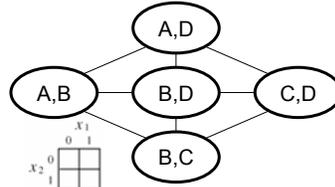
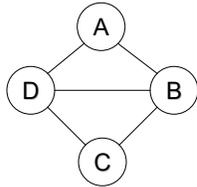
$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

- For discrete nodes, we can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table

Eric Xing

20

Example UGM – using subcliques



$$\begin{aligned}
 P''(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \prod_{ij} \psi_{ij}(x_{ij}) \\
 &= \frac{1}{Z} \psi_{12}(x_{12}) \psi_{14}(x_{14}) \psi_{23}(x_{23}) \psi_{24}(x_{24}) \psi_{34}(x_{34}) \\
 Z &= \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(x_{ij})
 \end{aligned}$$

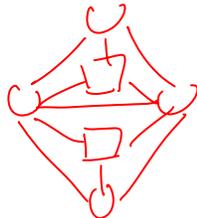
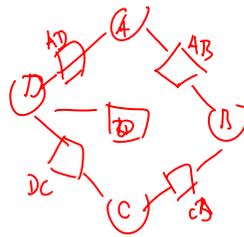
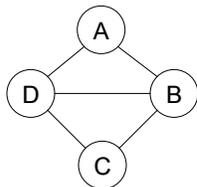
- We can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table
- Pair MRFs, a popular and simple special case

complete $I(P')$ vs. $I(P'')$? *discuss* | $D(P')$ vs. $D(P'')$

Eric Xing

21

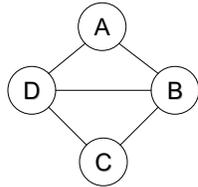
Factored graph



Eric Xing

22

Example UGM – canonical representation



$$\begin{aligned}
 P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234}) \\
 &\quad \times \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34}) \\
 &\quad \times \psi_1(x_1) \psi_2(x_2) \psi_3(x_3) \psi_4(x_4)
 \end{aligned}$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234}) \times \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34}) \times \psi_1(x_1) \psi_2(x_2) \psi_3(x_3) \psi_4(x_4)$$

- Most general, subsume P' and P'' as special cases
- I(P) ~~vs.~~ I(P') ~~vs.~~ I(P'')
- D(P) ~~vs.~~ D(P') ~~vs.~~ D(P'')

Eric Xing

23

Hammersley-Clifford Theorem



- If arbitrary potentials are utilized in the following product formula for probabilities,

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

then the family of probability distributions obtained is exactly that set which **respects** the *qualitative specification* (the conditional independence relations) described earlier

- **Thm (5.4.2):** Let P be a positive distribution over \mathbb{V} , and H a Markov network graph over \mathbb{V} . If H is an I-map for P, then P is a Gibbs distribution over H.

Eric Xing

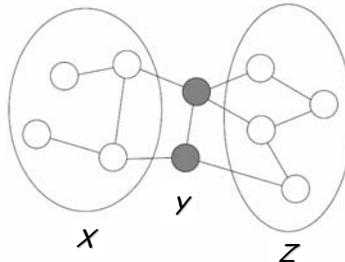
24

Independence properties of UGM



- Let us return to the question of what kinds of distributions can be represented by undirected graphs (ignoring the details of the particular parameterization).
- Defn: the global Markov properties of a UG H are

$$I(H) = \{X \perp Z | Y : \text{sep}_H(X; Z | Y)\}$$



- Is this definition sound and complete?

Eric Xing

25

Soundness and completeness of global Markov property



- Defn: An UG H is an I-map for a distribution P if $I(H) \subseteq I(P)$, i.e., P entails $I(H)$.
- Defn: P is a **Gibbs distribution** over H if it can be represented as

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

- Thm 5.4.1 (soundness): If P is a Gibbs distribution over H , then H is an I-map of P .
- Thm 5.4.5 (completeness): If $\neg \text{sep}_H(X; Z | Y)$, then $X \not\perp_P Z | Y$ in **some** P that factorizes over H .

Eric Xing

26

Local and global Markov properties revisit



- For directed graphs, we defined I-maps in terms of local Markov properties, and derived global independence.
- For undirected graphs, we defined I-maps in terms of global Markov properties, and will now derive local independence.
- Defn: The *pairwise Markov independencies* associated with UG $H = (V; E)$ are

$$I_p(H) = \{X \perp Y \mid V \setminus \{X, Y\} : \{X, Y\} \notin E\}$$

- e.g., $X_1 \perp X_5 \mid \{X_2, X_3, X_4\}$



Eric Xing

27

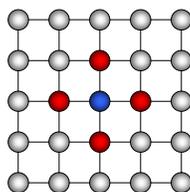
Local Markov properties



- A distribution has the *local Markov property* w.r.t. a graph $H = (V, E)$ if the conditional distribution of variable given its neighbors is independent of the remaining nodes

$$I_l(H) = \{X \perp V \setminus (X \cup N_H(X)) \mid N_H(X) : X \in V\}$$

- **Theorem** (Hammersley-Clifford): If the distribution is **strictly positive** and satisfies the local Markov property, then it factorizes with respect to the graph.
- $N_H(X)$ is also called the **Markov blanket** of X .



$I_g(H)$

$I_p(H)$

$I_l(H)$

Eric Xing

28

Relationship between local and global Markov properties



- Thm 5.5.5. If $P \models I_\lambda(H)$ then $P \models I_p(H)$.
- Thm 5.5.6. If $P \models I(H)$ then $P \models I_\lambda(H)$.
- Thm 5.5.7. If $P > 0$ and $P \models I_p(H)$, then $P \models I(H)$.
 - Pf sketch: $p(a,b|c,d)=p(a|c,d)p(b|c,d)$, then d separate b from {a,c}
 - $p(a,b|c,d)p(c|d)=p(a|c,d)p(b|c,d)p(c|d)=p(a,c|d)p(b|d)$
- **Corollary (5.5.8):** The following three statements are equivalent for a positive distribution P :

$$P \models I_\lambda(H)$$

$$P \models I_p(H)$$

$$P \models I(H)$$

- This equivalence relies on the positivity assumption.
- We can design a distribution locally

Eric Xing

29

I-maps for undirected graphs



- Defn: A Markov network H is a minimal I-map for P if it is an I-map, and if the removal of any edge from H renders it not an I-map.
- How can we construct a minimal I-map from a positive distribution P ?
 - Pairwise method: add edges between all pairs X, Y s.t.

$$P \not\models (X \perp Y | V \setminus \{X, Y\})$$
 - Local method: add edges between X and all $Y \in \text{MB}_P(X)$, where $\text{MB}_P(X)$ is the minimal set of nodes U s.t.

$$P \not\models (X \perp V \setminus \{X\} | U \cup Y)$$
 - Thm 5.5.11/12: both methods induce the unique minimal I-map.
- If $\exists x$ s.t. $P(x) = 0$, then we can construct an example where either method fails to induce an I-map.

Eric Xing

30

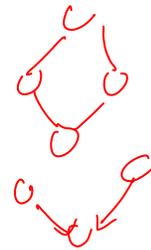
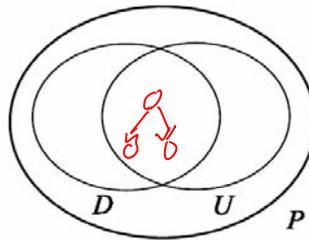
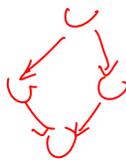
Perfect maps



- Defn: A Markov network H is a perfect map for P if for any X, Y, Z we have that

$$\text{sep}_H(X; Z | Y) \Leftrightarrow P \models (X \perp Z | Y)$$

- Thm: not every distribution has a perfect map as UGM.
 - Pf by counterexample. No undirected network can capture all and only the independencies encoded in a v-structure $X \rightarrow Z \leftarrow Y$.



Eric Xing

31

Exponential Form



- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_c(\mathbf{x}_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(\mathbf{x}_c)$:

$$\psi_c(\mathbf{x}_c) = \exp\{-\phi_c(\mathbf{x}_c)\}$$

For convenience, we will call $\phi_c(\mathbf{x}_c)$ a potential when no confusion arises from the context.

- This gives the joint a nice additive structure

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} \phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z} \exp\{-H(\mathbf{x})\}$$

where the sum in the exponent is called the "free energy":

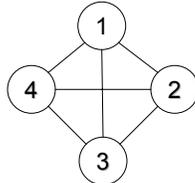
$$H(\mathbf{x}) = \sum_{c \in C} \phi_c(\mathbf{x}_c)$$

- In physics, this is called the "Boltzmann distribution".
- In statistics, this is called a log-linear model.

Eric Xing

32

Example: Boltzmann machines



$$x_i \in [-1, 1]$$

$$x_i x_j = \begin{cases} 1 & \uparrow \\ -1 & \downarrow \end{cases}$$

- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1, +1\}$ or $x_i \in \{0, 1\}$) is called a Boltzmann machine

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp \left\{ \sum_{ij} \phi_{ij}(x_i, x_j) \right\}$$

$$= \frac{1}{Z} \exp \left\{ \sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\}$$

- Hence the overall energy function has the form:

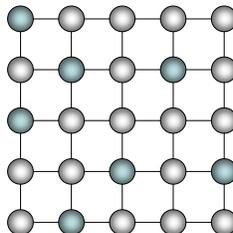
$$H(x) = \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) = (x - \mu)^T \Theta (x - \mu)$$

Eric Xing

33

Example: Ising models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.



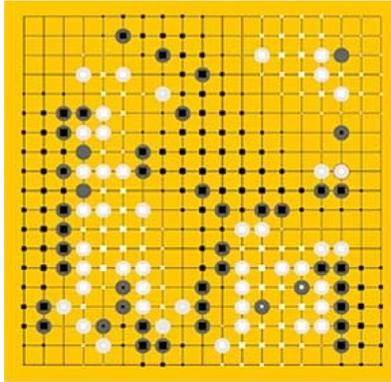
$$p(X) = \frac{1}{Z} \exp \left\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \right\}$$

- Same as sparse Boltzmann machine, where $\theta_{ij} \neq 0$ iff i, j are neighbors.
 - e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.
- Potts model:** multi-state Ising model.

Eric Xing

34

Application: Modeling Go



This is the middle position of a Go game.
Overlaid is the estimate for the probability of becoming black or white for every intersection.
Large squares mean the probability is higher.

Eric Xing

35

Example: multivariate Gaussian Distribution



- A Gaussian distribution can be represented by a fully connected graph with pairwise (edge) potentials over continuous nodes.
- The overall energy has the form

$$H(x) = \sum_{ij} (x_i - \mu) \Sigma_{ij}^{-1} (x_j - \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

where μ is the mean and Θ is the inverse covariance (precision) matrix.

- Also known as Gaussian graphical model (GGM), same as Boltzmann machine except $x_i \in \mathbb{R}$

$p(x) \propto \exp(-H)$

$\sum_{ij} \Theta_{ij} x_i x_j + \sum_i \theta_i x_i$

$\Sigma_{ij} = 0 \rightarrow$ miss edge \rightarrow C-independent

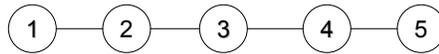
$G(\mu, \Sigma) \rightarrow \Sigma^{-1} \rightarrow$ Gibbs \rightarrow VG

Boltzmann

Eric Xing

36

Sparse precision vs. sparse covariance in GGM



$$\Sigma^{-1} = \begin{pmatrix} 1 & 6 & 0 & 0 & 0 \\ 6 & 2 & 7 & 0 & 0 \\ 0 & 7 & 3 & 8 & 0 \\ 0 & 0 & 8 & 4 & 9 \\ 0 & 0 & 0 & 9 & 5 \end{pmatrix} \Rightarrow \Sigma = \begin{pmatrix} 0.10 & 0.15 & -0.13 & -0.08 & 0.15 \\ 0.15 & -0.03 & 0.02 & 0.01 & -0.03 \\ -0.13 & 0.02 & 0.10 & 0.07 & -0.12 \\ -0.08 & 0.01 & 0.07 & -0.04 & 0.07 \\ 0.15 & -0.03 & -0.12 & 0.07 & 0.08 \end{pmatrix}$$

$$\Sigma_{15}^{-1} = 0 \Leftrightarrow X_1 \perp X_5 \mid X_{nbrs(1) \text{ or } nbrs(5)}$$

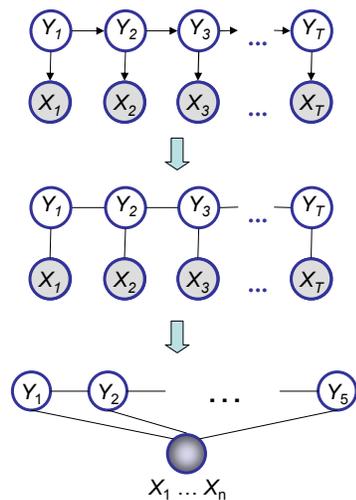
$$\nRightarrow$$

$$X_1 \perp X_5 \Leftrightarrow \Sigma_{15} = 0$$

Eric Xing

37

Example: Conditional Random Fields



- Discriminative

$$p_{\theta}(y|x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Doesn't assume that features are independent
- When labeling X_i , future observations are taken into account

Eric Xing

38

Conditional Models



- Conditional probability $P(\text{label sequence } \mathbf{y} \mid \text{observation sequence } \mathbf{x})$ rather than joint probability $P(\mathbf{y}, \mathbf{x})$
 - Specify the probability of possible label sequences given an observation sequence
- Allow arbitrary, non-independent features on the observation sequence \mathbf{X}
- The probability of a transition between labels may depend on **past** and **future** observations
- Relax strong independence assumptions in generative models

Eric Xing

39

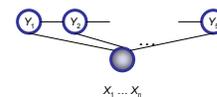
Conditional Distribution



- If the graph $G = (V, E)$ of \mathbf{Y} is a tree, the conditional distribution over the label sequence $\mathbf{Y} = \mathbf{y}$, given $\mathbf{X} = \mathbf{x}$, by the Hammersley Clifford theorem of random fields is:

$$p_{\theta}(\mathbf{y} \mid \mathbf{x}) \propto \exp \left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x}) \right)$$

- \mathbf{x} is a data sequence
- \mathbf{y} is a label sequence
- v is a vertex from vertex set $V = \text{set of label random variables}$
- e is an edge from edge set E over V
- f_k and g_k are given and fixed. g_k is a Boolean vertex feature; f_k is a Boolean edge feature
- k is the number of features
- $\theta = (\lambda_1, \lambda_2, \dots, \lambda_n; \mu_1, \mu_2, \dots, \mu_n)$; λ_k and μ_k are parameters to be estimated
- $\mathbf{y}|_e$ is the set of components of \mathbf{y} defined by edge e
- $\mathbf{y}|_v$ is the set of components of \mathbf{y} defined by vertex v



Eric Xing

40

Conditional Distribution (cont'd)

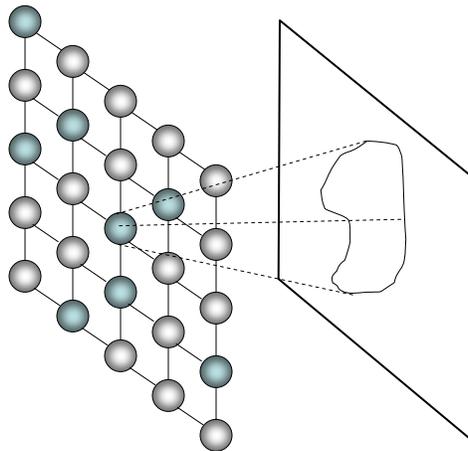


- CRFs use the observation-dependent normalization $Z(\mathbf{x})$ for the conditional distributions:

$$p_{\theta}(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{e \in E, k} \lambda_k f_k(e, y|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, y|_v, \mathbf{x})\right)$$

- $Z(\mathbf{x})$ is a normalization over the data sequence \mathbf{x}

Conditional Random Fields



$$p_{\theta}(y|\mathbf{x}) = \frac{1}{Z(\theta, \mathbf{x})} \exp\left\{\sum_c \theta_c f_c(\mathbf{x}, y_c)\right\}$$

- Allow arbitrary dependencies on input
- Clique dependencies on labels
- Use approximate inference for general graphs