Stat 521A
Lecture 2

Outline

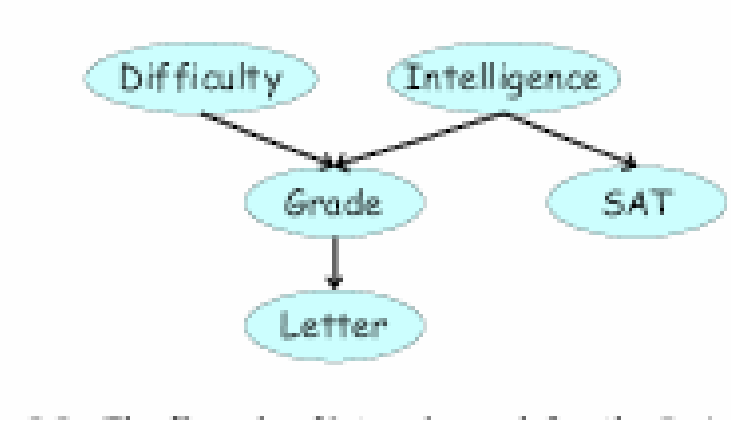
- DAGs
  - global Markov (3.3)
  - deriving graphs from distributions (3.4)
- UGs
  - Global Markov property (4.3.1)
  - Parameterization (4.2)
  - Gibbs distributions, energy based models (4.4.1)
  - Local and pairwise Markov properties (4.3.2)
  - From distributions to graphs (4.3.3)

Active trails

- Whenever influence can flow from to Y via Z, we say that the trail X <-> Y <-> Z is active.
- Causal trail: X -> Z -> Y. Active iff Z not obs.
- Evidential trail: X <- Z <- Y. Active iff Z not obs
- Common cause: X <- Z -> Y. Active iff Z not obs
- Common effect; X -> Z <- Y. Active iff either Z or one of its descendants is observed.
- Def 3.3.1. Let G be a BN structure, and X1 <-> … <-> Xn be a trail in G. Let E be a subset of nodes. The trail is active given E if
- Whenever we have a v-structure X_{i-1} -> X_i <- X_{i+1}, then X_i or one of its desc is in E
- No other node along the trail is in E

Example

- D-> G <- I ->S not active for E={}
- D-> G <- I ->S is active for E={L}
- D-> G <- I ->S not active for E={L,I}
- Non-monotonic



d-separation

- Def 3.3.2, We say X and Y are d-separated given Z, denoted d-sep_G(X;Y|Z), if there is no active trail between any node in X to any node in Y, given Z. The set of such independencies is denoted

$$I(G) = \{X \perp Y | Z : \text{dsep}_G(X; Y|Z)\}$$

- Thm 3.3.3. (Soundness of dsep). If P factorizes according to G, then I(G) $\subseteq$ I(P).
- False thm (completeness of dsep). For any P that factorizes according to G, if X $\perp$ Y | Z in I(P), then desp$_G$(X;Y|Z) (i.e., P is faithful to G)

Faithfulness

- Def 3.3.4. A distribution P is faithful to G if, whenever X ⊥ Y | Z in I(P), we have dsep_G(X;Y|Z) i.e., there are no "non-graphical" independencies buried in the parameters
- A simple unfaithful distribution, with Imap A->B:
-

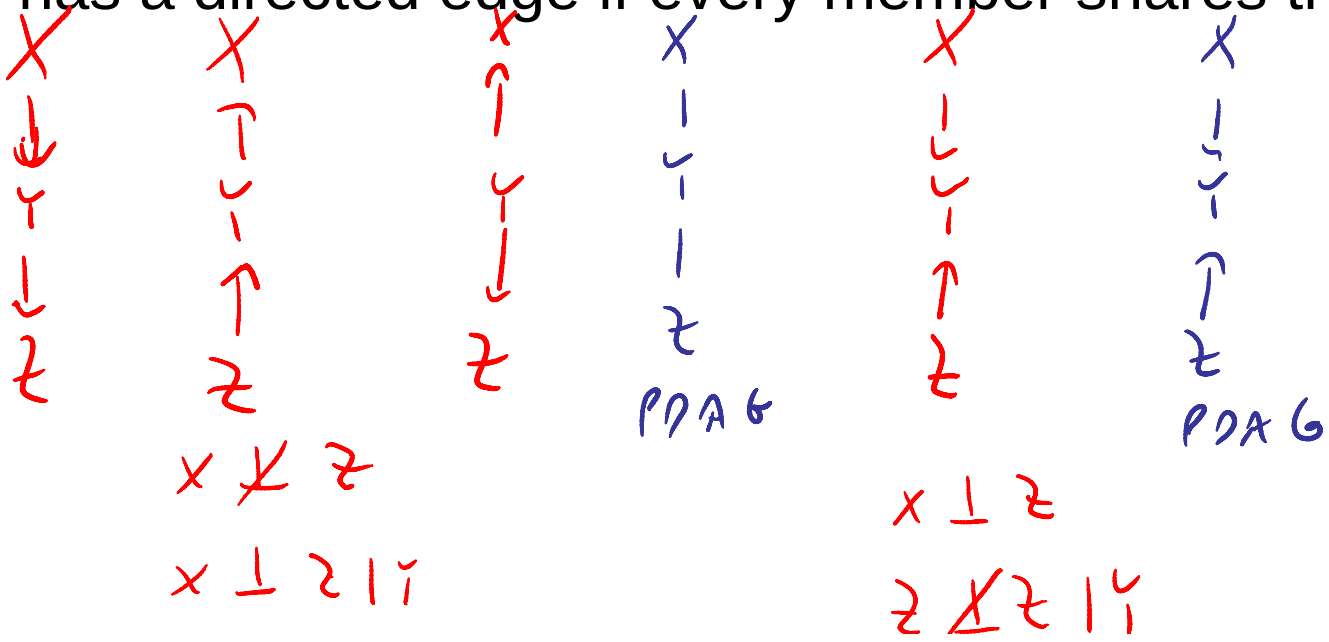
	b^0	b^1
a^0	0.4	0.6
a^1	0.4	0.6

Such distributions are "rare"

- Thm 3.3.7. For almost all distributions P that factorize over G (ie except for a set of measure zero in the space of CPD parameterizations), we have that I(P)=I(G)

Markov equivalence

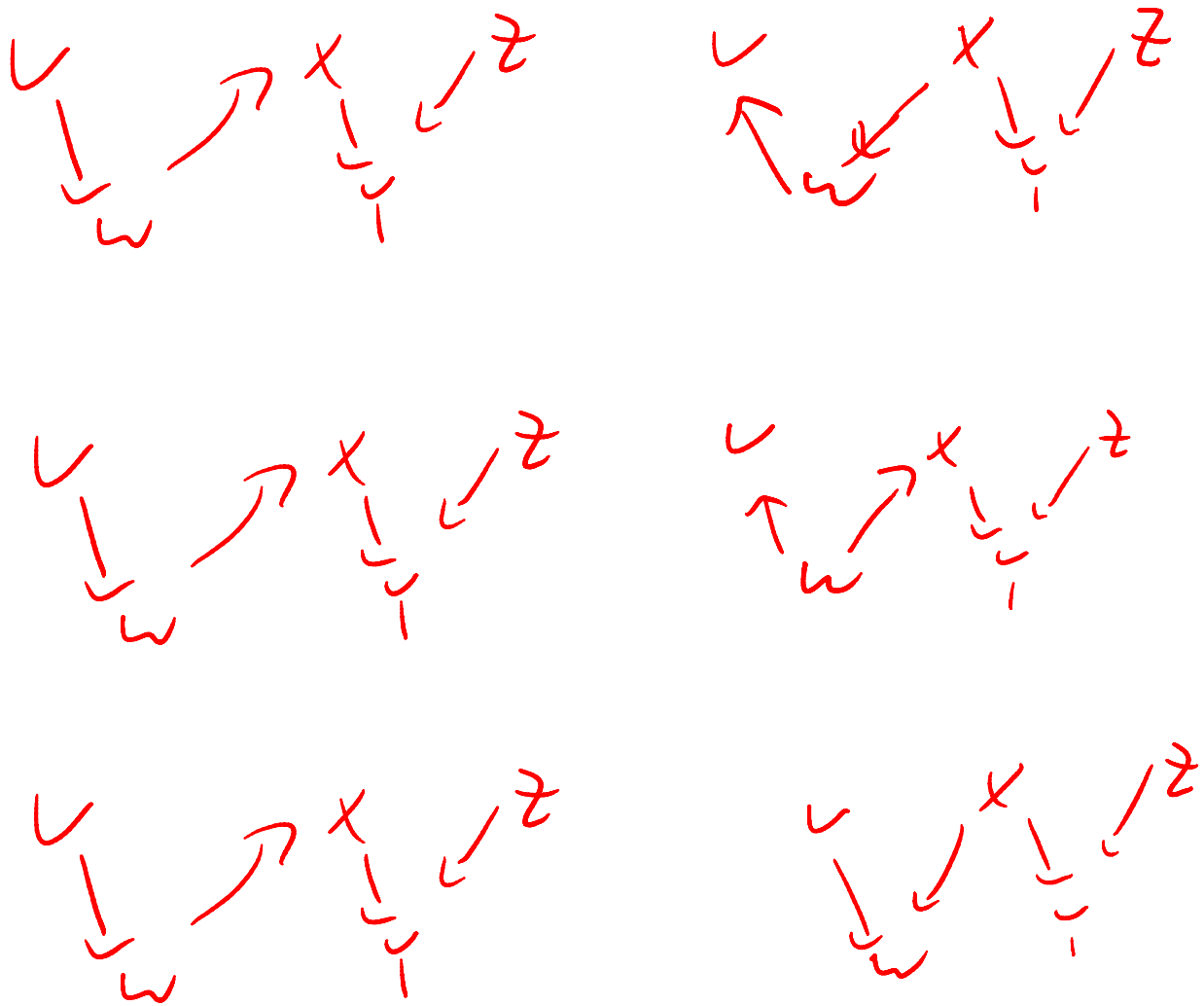
- A DAG defines a set of distributions. Different DAGs may encode the same set and hence are indistinguishable given observational data.
- Def 3.3.10. DAGs G1 and G2 are I-equivalent if I(G1)=I(G2). The set of all DAGs can be partitioned into I-equivalence classes.
- Def 3.4.11. Each can be represented by a class PDAG: only has a directed edge if every member shares that edge.



Identifying I-equivalence

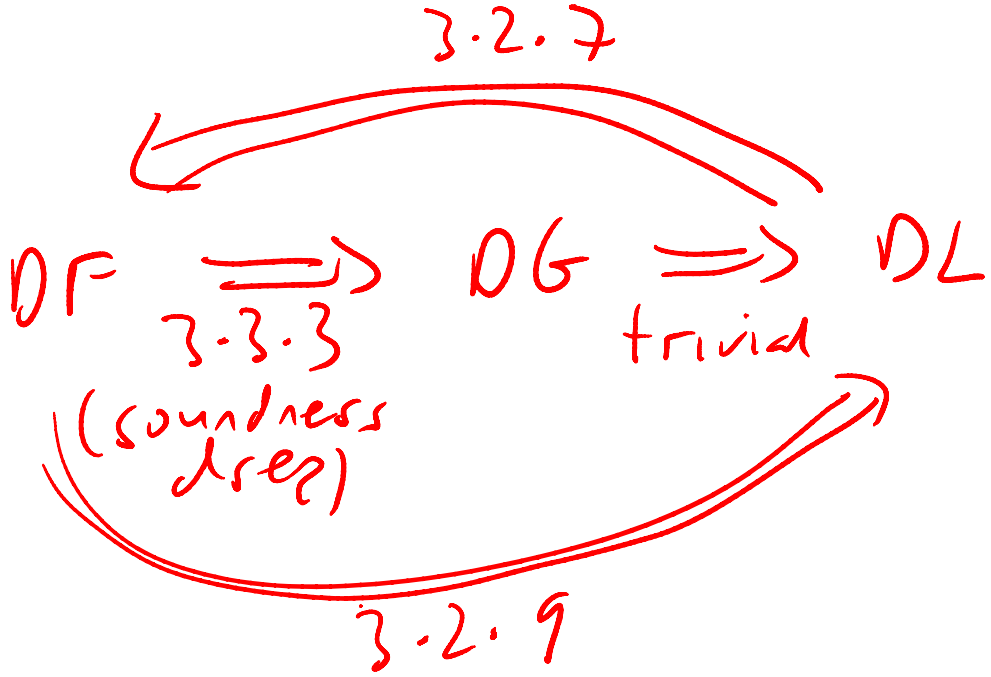
- Def 3.3.11. The skeleton of a DAG is an undirected graph obtained by dropping the arrows.
- Thm 3.3.12. If G1 and G2 have the same skeleton and the same v-structures, they are I-equivalent.
- However, there are structures that are I-equiv but do not have same v-structures (eg fully connected DAG).
- Def 3.3.13. A v-structure X->Z<-Y is an immorality if there is no edge between X and Y (unmarried parents who have a child)
- Thm 3.3.14. G1 and G2 have the same skeleton and set of immoralities iff they are I-equiv.

Examples



Markov properties of DAGs

- DF: F factorizes over G
- DG: $I(G) \subseteq I(P)$
- DL: $I_l(G) \subseteq I(P)$



Based on Jordan ch 4

Deriving graphs from distributions

- So far, we have discussed how to derive distributions from graphs.
- But how do we get the DAG?
- Assume we have access to the true distribution P, and can answer questions of the form

$$P \models X \perp Y | Z$$

- For finite data samples, we can approximate this oracle with a CI test – the frequentist approach to graph structure learning (see ch 18)
- What DAG can be used to represent P?

Minimal I-map

- The complete DAG is an I-map for any distribution (since it encodes no CI relations)
- Def 3.4.1. A graph K is a minimal I-map for a set of independencies I if it is an I-map for I, and if the removal of even a single edge from K renders it not an I-map.
- To derive a minimal I-map, we pick an arbitrary node ordering, and then find some minimal subset U to be X_i's parents, where

$$X_i \perp \{X_1, \ldots, X_{i-1}\} \setminus U | U$$

- (K2 algorithm replace this CI test with a Bayesian scoring metric: sec 18.4.2).

Effect of node ordering

- "Bad" node orderings can result in dense, unintuitive graphs.
- Eg L,S,G,I,D. Add L. Add S: must add L as parent, since $P \not\models L \perp S$. Add G: must add L,S as parents.

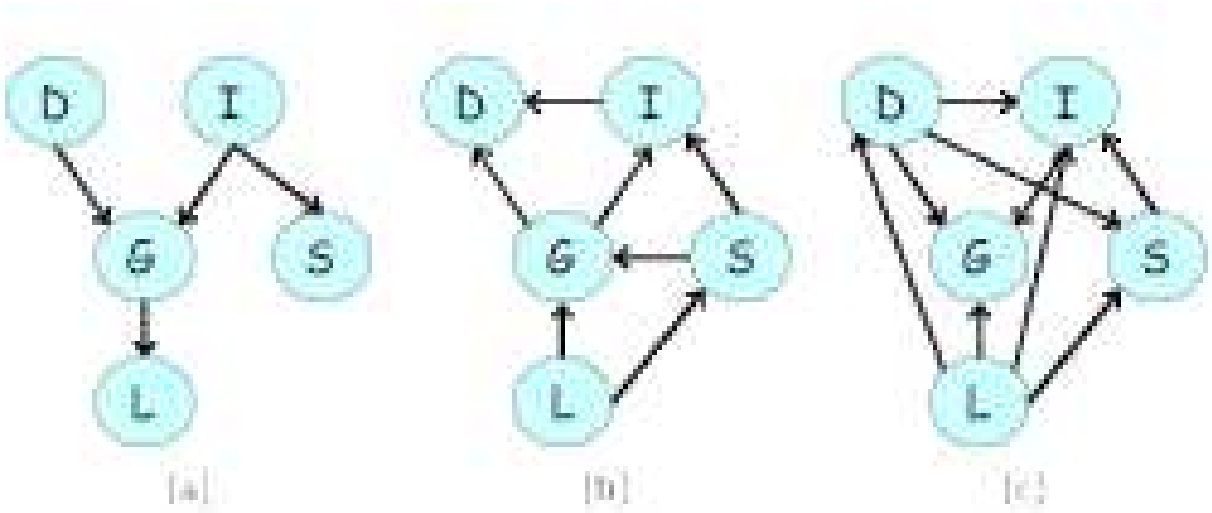
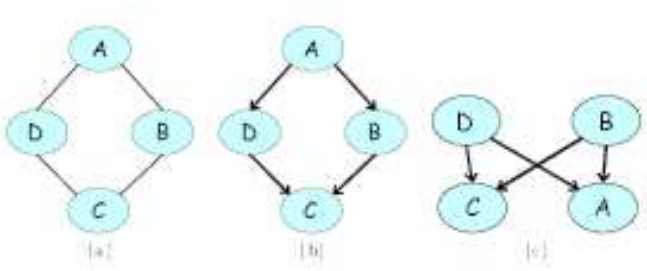


Figure 3.8 Three minimal I-maps for $P_{\{Student\}}$ induced by different orderings: (a) D, I, S, G, L (b) L, S, G, I, D (c) L, D, S, I, G

Perfect maps

- Minimal I-maps can have superfluous edges.
- Def 3.4.2. Graph K is a perfect map for a set of independencies I if I(K)=I. K is a perfect map for P if I(K)=I(P).
- Not all distributions can be perfectly represented by a DAG.
- Eg let Z = xor(X,Y) and use some independent prior on X, Y. Minimal I-map is X -> Z <- Y. However, X $\perp$ Z in I(P), but not in I(G).
- Eg. A $\perp$ C | {B,D} and B $\perp$ D | {A,C}, A dep | B,C, etc



Finding perfect maps

- If P has a perfect map, we can find it in polynomial time, using an oracle for the CI tests.
- We can only identify the graph up to I-equivalence, so we return the PDAG that represents the corresponding equivalence class.
- The method* has 3 steps (see sec 3.4.3)
  - Identify undirected skeleton
  - Identify immoralities
  - Compute eclass (compelled edges)
- This algorithm has been used to claim one can infer causal models from observational data, but this claim is controversial

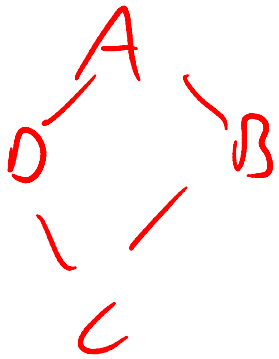
Algorithm due to Verma & Pearl 1991, Spirtes, Glymour, Scheines 1993, Meek 1995

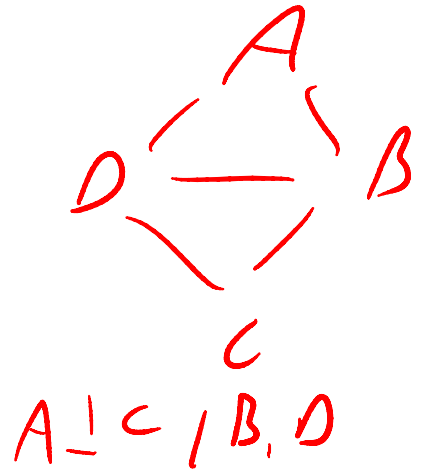
Global Markov property of UGs

- Def 4.3.1. The path X_1 - … - X_k is active given E if none of the nodes on the path are in E.
- Def 4.3.2. The global Markov assumptions associated with a UG H are

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

•eg. A ⊥ C | {B,D} and B ⊥ D | {A,C}





Monotonic, unlike d-separation

$$\text{sep}_H(X; Y | Z) \Rightarrow \text{sep}_H(X; Y | Z') \forall Z \subset Z'$$

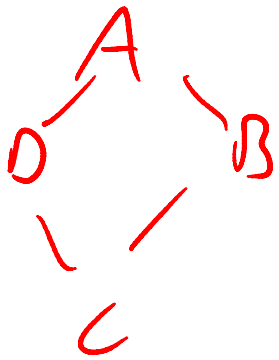
Parameterization

- To specify a specific distribution, we need to associate parameters (local distributions) with the graph.
- CPDs cannot be used because they are not symmetric, and the chain rule need not apply.
- Marginals cannot be used because a product of marginals does not define a consistent joint.
- Instead we multiply a product of **factors (potentials),** one per maximal clique, and then compute a global normalization constant Z (partition function)

$$P(A,B,C,D) = 1/Z\ \phi(A,B,D)\ \phi(B,C,D)$$

$$Z = \sum\_\{A,B,C,D\}\ \phi(A,B,D)\ \phi(B,C,D)$$

Misconception network



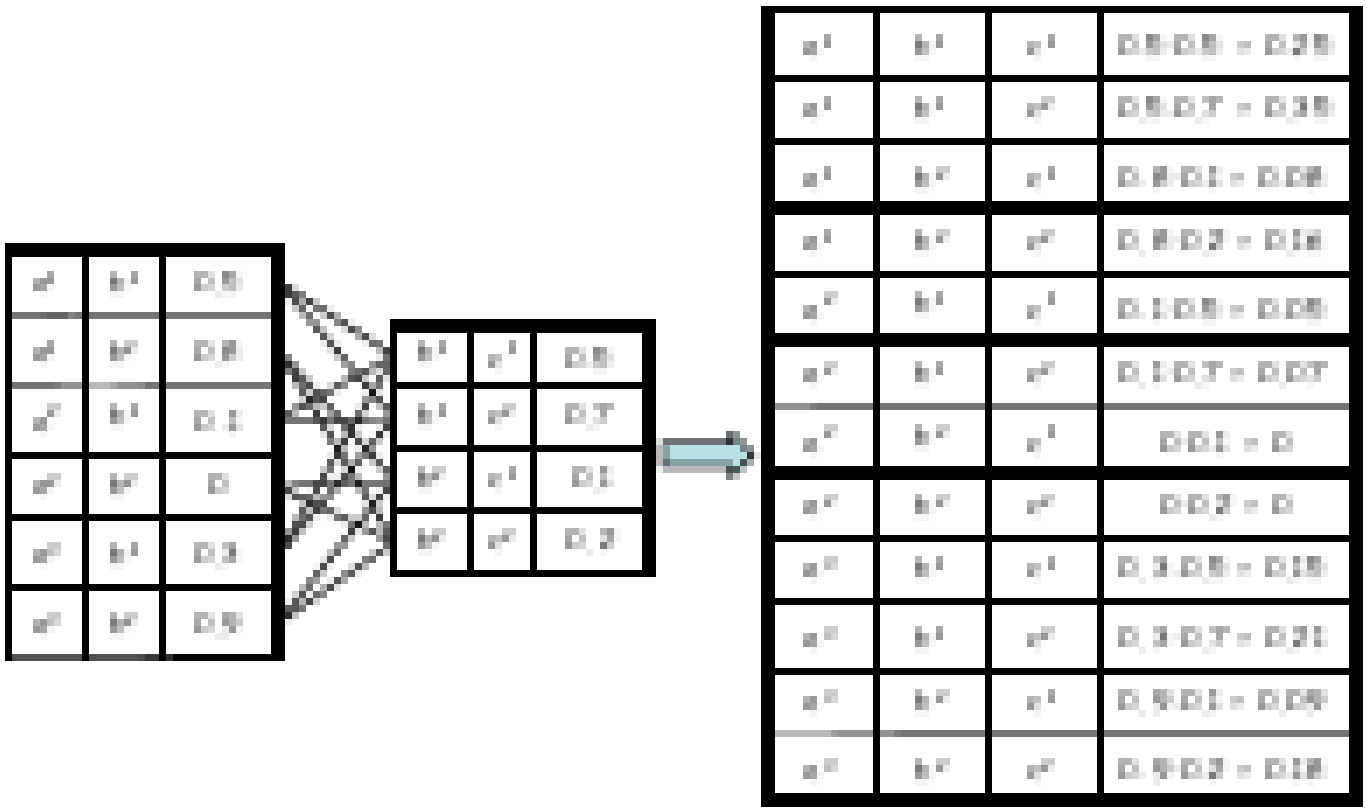
$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

P(A,B,C,D) = 1/Z ϕ(A,B) ϕ(A,D) ϕ(C,D) ϕ(C,B)

Multiplying factors

- Def 4.2.2. We multiply factors by matching up corresponding dimensions
  $\Psi(X,Y,Z) = \phi_1(X,Y) \cdot \phi_2(Y,Z)$



Factors are not marginals

- In the misconception network, the marginal on A,B is

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

- But the local clique potential is

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

- Factors are local affinities or preferences, but get combined with other terms in a non-local way

Factorization and I-maps

- Thm 4.3.3. If P factorizes over H, then H is an I-map for P, ie. I(H) $\subseteq$ I(P). (Soundness of separation.)
- Proof. Suppose Z separates X from Y. Then we can partition the factors such that

$$p(\mathbf{x}) = (1/Z) f(X, Z) g(Y, Z)$$

  QED.
- Def 2.1.11. A distribution is positive if P(x)>0 for all x.
- Thm 4.3.4 (Hammersley Clifford). If P is positive, and H is an I-map for P, then P factorizes over H:

$$p(\mathbf{x}) = (1/Z) \prod_c \phi_c(\mathbf{x}_c)$$

Gibbs distributions

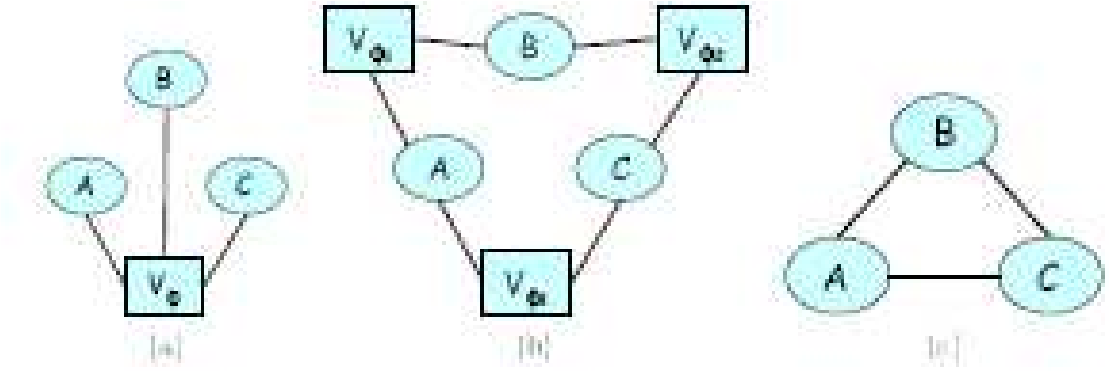
- Def 4.2.3. A Gibbs distribution is defined as

$$p(X_1, \ldots, X_n) = \frac{1}{Z}\phi_1(D_1) \times \cdots \times \phi_m(D_m)$$

- The D_i are the domains or scopes of the factors. We can infer the graph by connecting up all nodes in the same domain. If the D_i are on pairs of nodes (edges), we call it a pairwise Markov random field.
- For a complete graph, we could have one factor per edge or a single clique potential for the whole graph.

Factor graphs

- For a complete graph, we could have one factor per edge or a single clique potential for the whole graph.
- Factor graphs can distinguish these cases.
- Def 4.4.1. Square nodes = factors, ovals = rv's.



Energy based models

- It is common to work with energies = negative log factors/ potentials (low energy = more probable)

$$\phi(D) = \exp(-\epsilon(D)) \qquad p(x_1, \ldots, x_n) = 1/Z \exp[-\sum_{i=1}^{m} \epsilon_i(D_i)]$$

$\epsilon_1[A, B]$

a^0	b^0	-3.4
a^0	b^1	-1.61
a^1	b^0	0
a^1	b^1	-2.3

$\epsilon_2[B, C]$

b^0	c^0	-4.61
b^0	c^1	0
b^1	c^0	0
b^1	c^1	-4.61

$\epsilon_3[C, D]$

c^0	d^0	0
c^0	d^1	-4.61
c^1	d^0	-4.61
c^1	d^1	0

$\epsilon_4[D, A]$

d^0	a^0	-4.61
d^0	a^1	0
d^1	a^0	0
d^1	a^1	-4.61

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

Ising model

- X_i = +1 if atom is spin up, X_i = -1 if spin down
- Define edge energy as

$$\epsilon_{i,j}(x_i, x_j) = -w_{i,j} x_i x_j$$

- If spins equal (aligned), product is +1, else -1.
- w_{i,j} = 0.5 (E(anti-aligned)-E(aligned)). If +ve, model aligns atoms (ferromagnetic). If –ve, spins should be different (anti-ferromagnetic).
- Define local node energy (external field) as

$$\epsilon_i(x_i) = -u_i x_i$$

- Overall distribution

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left(\sum_{i<j} w_{i,j} x_i x_j + \sum_i u_i x_i \right)$$

Ising models capture pairwise correlation

- Energy can be written as

$$
\begin{aligned}
\epsilon(\mathbf{x}) &= -\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i \\
&= -\tfrac{1}{2}\mathbf{x}^T \mathbf{W} \mathbf{x} - \mathbf{u}^T \mathbf{x} \\
&= -\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{W} (\mathbf{x} - \boldsymbol{\mu}) + c \\
\boldsymbol{\mu} &= -\mathbf{W}^{-1}\mathbf{u} \\
c &= \tfrac{1}{2}\boldsymbol{\mu}^T \mathbf{W} \mu
\end{aligned}
$$

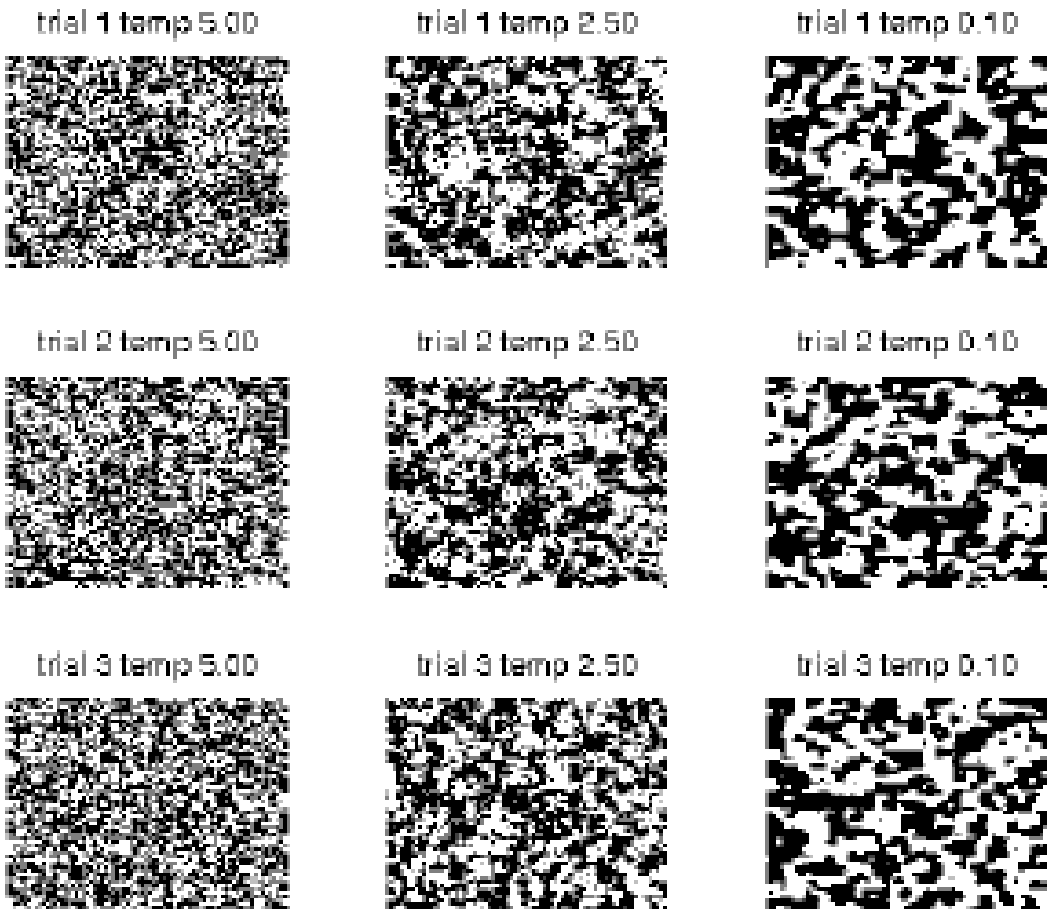
Phase transition

- The strength of the interactions is modulated by a global temperature parameter T

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\epsilon(\mathbf{x})/T\right)$$

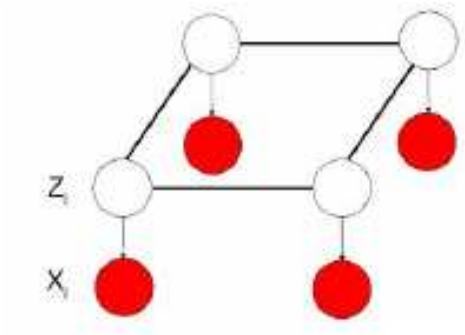
- Large temperature “flattens” the energy landscape and makes the uniform distribution most probable
- Small temperature makes the distribution “peaky”
- One can compute the density of pure vs mixed state configurations as a function of T (as the number of atoms -> ∞). There is often a phase transition: as T exceeds a critical temperature, there is a sudden regime change.
- This has computational analogs in the mixing time of Markov chains.

Samples from an Ising model



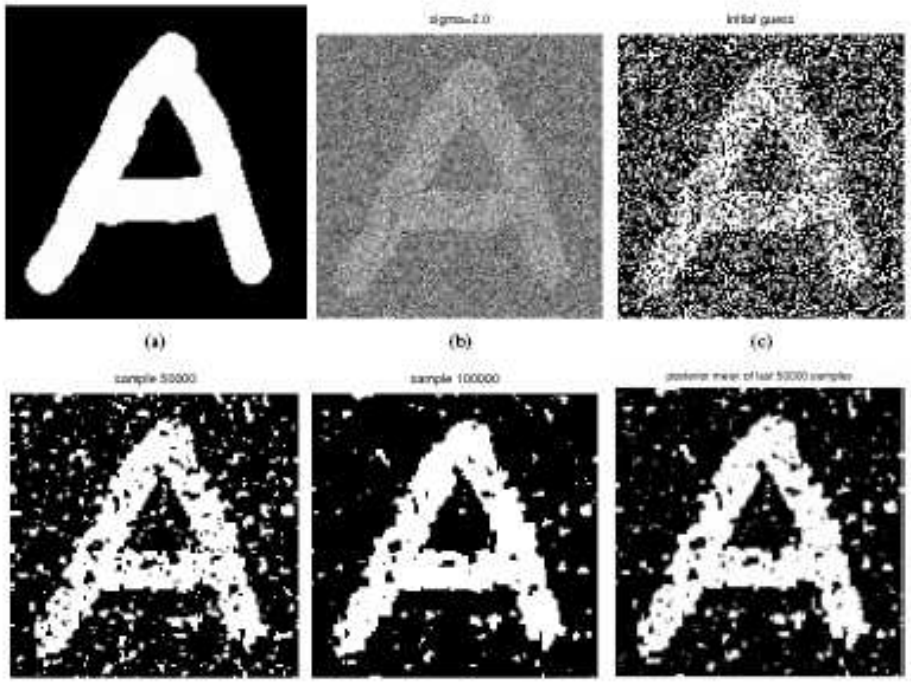
See GibbsDemoIsing in PMTK/bookCode

Image denoising



$$p(\mathbf{x},\mathbf{y}) \quad = \quad p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{<ij>}\phi_{ij}(x_i,x_j)\prod_i p(y_i|x_i)$$

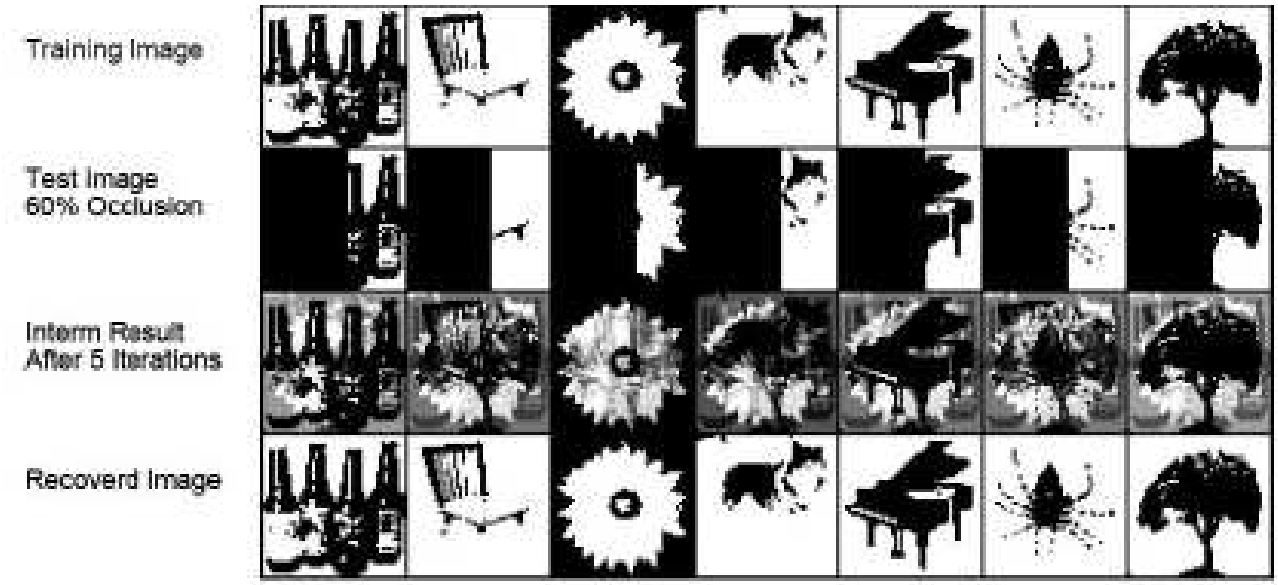
argmax_x P(x|y) is best guess of denoised image



See GibbsDemoIsing in PMTK/bookCode

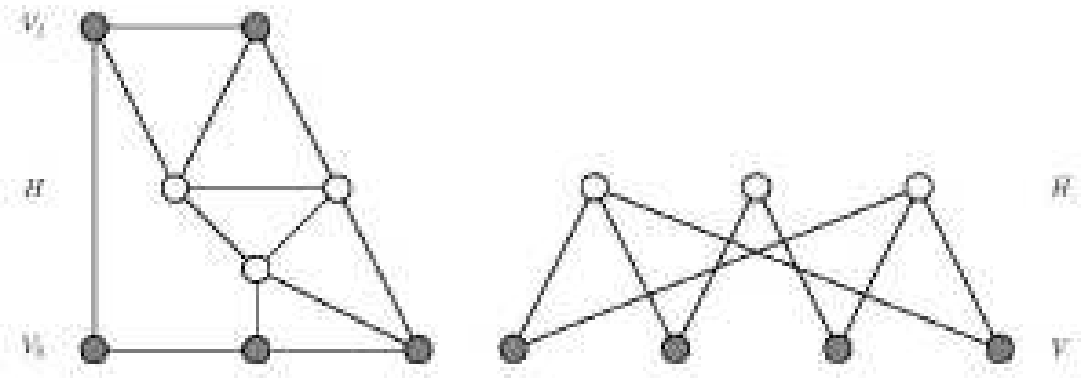
Hopfield network

- A Hopfield network is a stochastic, recurrent neural network.
- It is equivalent to a fully connected Ising model.
- Weights are learned.
- Often used for associative memory/ pattern completion.



Boltzmann machine

- A Boltzmann machine is a Hopfield network (Ising model) with hidden nodes.
- A restricted Boltzmann machine (RBM) is a bipartite BM. This supports efficient block Gibbs sampling (see ch 12).

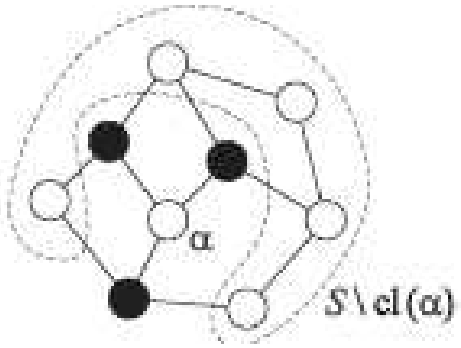


Local Markov assumption

- So far, we have defined the global Markov assumptions using simple graph separation.
- We now consider some variants.
- The boundary of a node α, bd(α), is all nodes which are directly connected to it.
- The closure is cl(α) = bd(α)U α.
- Def 4.3.9. The local Markov properties of H are

$$I_l(H) = \{\alpha \perp S \setminus cl(\alpha) | bd(\alpha)\}$$

- i.e. a is indep of rest given its Markov blanket bd(a).



Pairwise Markov assumption

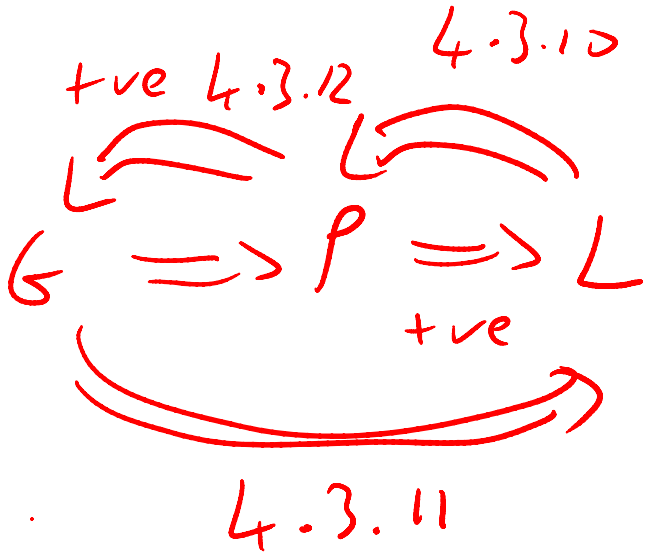
- Def 4.3.7. The pairwise Markov independencies associated with H are

$$I_p(H) = \{\alpha \perp \beta | S \setminus \{\alpha, \beta\} : \alpha - \beta \notin H\}$$

- i.e., a is independent of b given rest if not directly connected.

Markov properties

- G: $I(G) \subseteq I(P)$
- L: $I_l(G) \subseteq I(P)$
- P: $I_p(G) \subseteq I(P)$
- If P is positive, all are equivalent.



Based on Jordan ch 4, thm numbers refer to Koller&Friedman

Problems caused by determinism

- If the distribution is not positive, pairwise indep does not imply local or global indep.
- Ex 4.3.15. Let P be any distribution over $(X_1,\ldots,X_n)$. Make 3 identical copies of each variable, X_i, X_i', X_i''. Let H be the empty MRF on this expanded state space. This satisfies the pairwise Markov properties eg X_i and X_i' are independent, because the remaining nodes contain X_i''. Also, X_i and X_j are independent, because the remaining nodes contain X_i'. However, H does not satisfy local or global indep.

From distributions to graphs

- How do we derive a graph from a distribution?
- For positive distributions, there are two approaches, based on pairwise and local prop.
- Thm 4.3.17. Let P be a +ve dist. Let H be an MRF in which we add an edge X-Y for all X,Y which cannot be made independent when conditioned on any other set:

$$P \not\models (X \perp Y | \mathcal{X} \setminus \{X, Y\})$$

Then H is the unique minimal I-map for P.

From distributions to graphs

- Thm 4.3.18. Let P be a +ve dist. For each node X, let MB_P(X) be a minimal set of nodes U rendering X indep of the rest:

$$X \perp \mathcal{X} \setminus \{X\} \setminus U | U \in I(P)$$

Add an edge X-Y for all Y in MB_P(X). Then H is a unique minimal I-map for P.