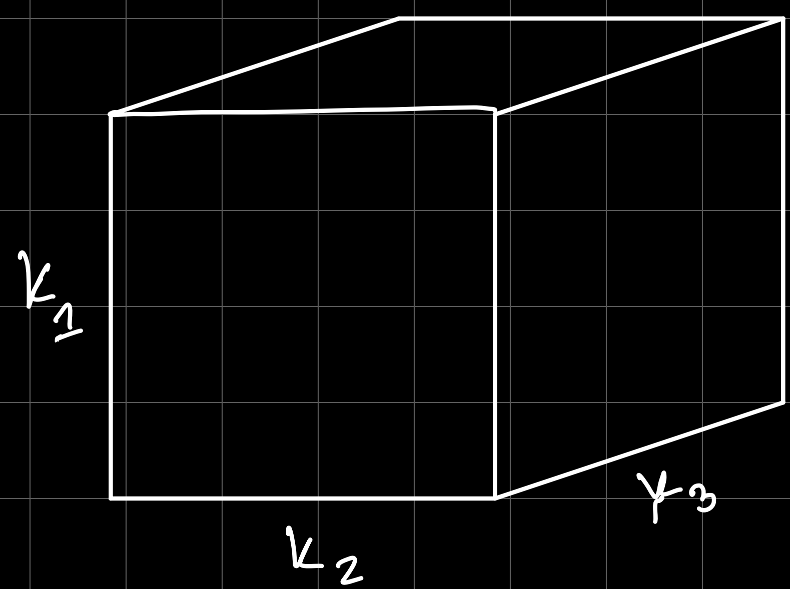
probabilistic graphical models (PGMs)

- model $p(X_1 \ldots X_n)$
- e.g., discrete $X_i \in \{1 \ldots k_i\}$, $i = 1 \ldots n$
- notation:
  $p(x_1 \ldots x_n) \equiv p(X_1 = x_1, \ldots, X_n = x_n)$
- $p(x_1 \ldots x_n)$ stored in a <u>probability table</u>
- e.g., $n = 3$



- $$\sum_{x_1=1}^{k_1} \sum_{x_2=1}^{k_2} \sum_{x_3=1}^{k_3} p(X_1 = x_1, X_2 = x_2, X_3 = x_3) = 1$$
- $k_1 \times k_2 \times k_3$ entries
- <u>exponential</u> in n
- e.g. k^n if $k_1 = k_2 \cdots = k_n = k$

- So what? Consider a conditional distribution (e.g., a posterior)...

$$P(x_1 \cdots x_{n-1} \mid X_n = x_n) = \frac{P(x_1 \cdots x_n)}{\underbrace{\sum_{x_1=1}^{K} \cdots \sum_{x_{n-1}=1}^{K} P(x_1 \cdots x_{n-1}, x_n)}_{K^{n-1} \text{ summands}}}$$

- This becomes <u>intractable</u> quickly

- e.g., a single binary trial $x_i \in \{0, 1\}$ in all $n = 140$ search cells
    - 2^{140} cells of $P(x_1 \cdots x_n)$
    - modern processor can compute $\sim 10^9$ FLOPs/sec
    - so summing 2^{140} cells would take $\sim 10^{33}$ seconds $\approx 10^{25}$ years !! (age of universe: 10^{10} years)

Directed graphical models (DGMs)

- The reason there were so many cells in $P(x_1 \dots x_n)$ is because we did not account for any <u>conditional independence</u> structure.

- e.g. say $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$

  $\hookrightarrow \; P(x_1 \dots x_n) = \prod_{i=1}^{n} P(x_i)$

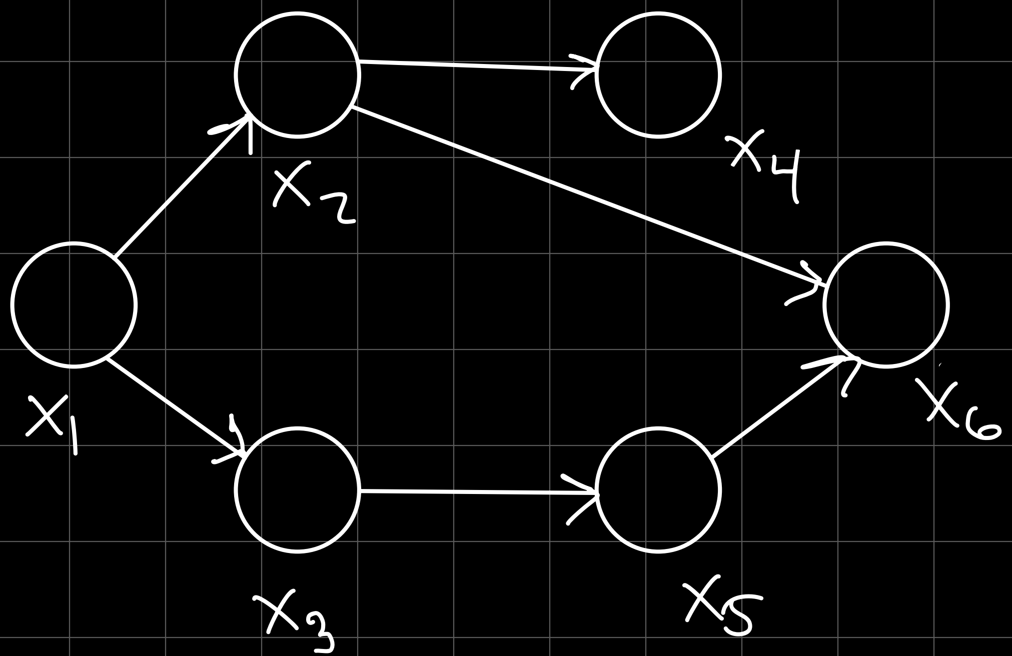
  we only need to store $\sum_{i=1}^{n} K_i$ cells (as opposed to $\prod_{i=1}^{n} K_i$)

- DGMs provide a formal language to describe the set of conditional independencies in a joint distribution.

- We are already used to defining models (i.e. joint distributions) in terms of <u>forward sampling algorithms</u>

- e.g.,
$$
\begin{aligned}
X_1 &\sim P(X_1) \\
X_2 &\sim P(X_2 \mid X_1) \\
X_3 &\sim P(X_3 \mid X_1) \\
X_4 &\sim P(X_4 \mid X_2) \\
X_5 &\sim P(X_5 \mid X_3) \\
X_6 &\sim P(X_6 \mid X_5, X_2)
\end{aligned}
$$

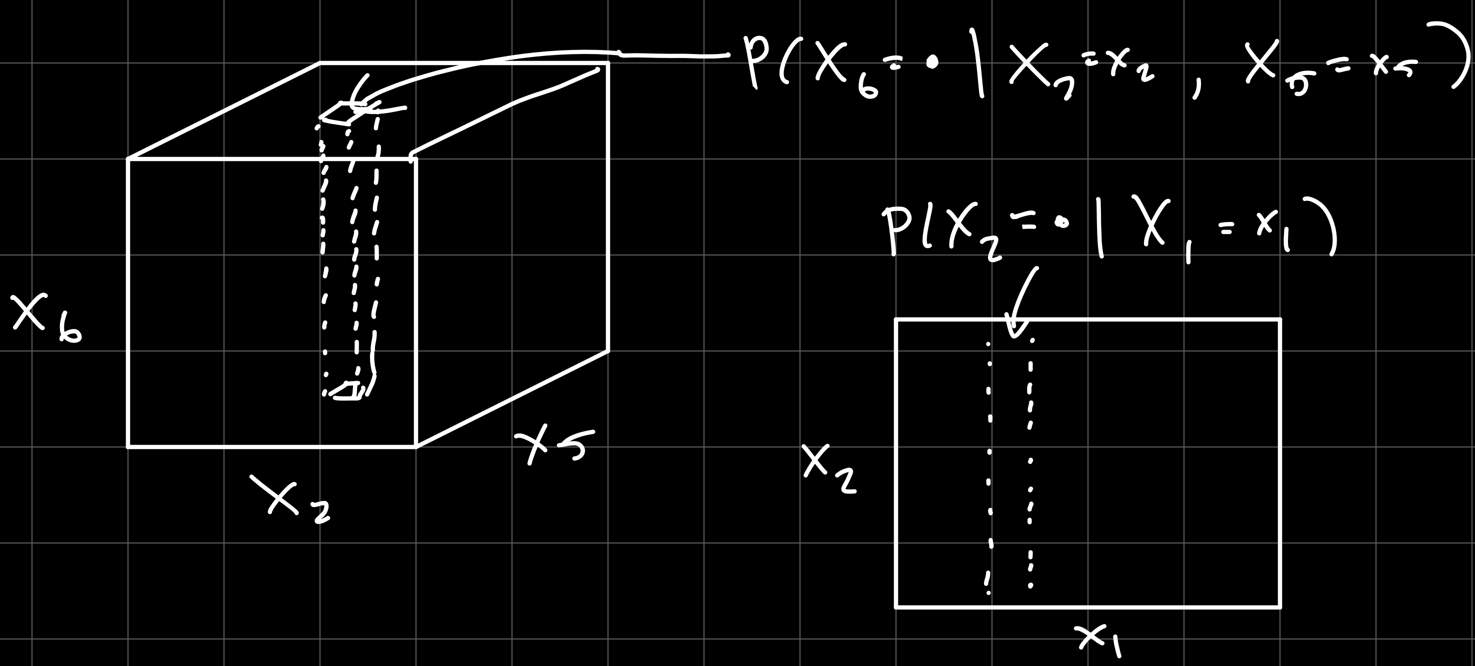
- We can represent this algorithm graphically:



- Both the algorithm and the graph represent a joint dist. that factorizes as

$$P(X_1 \dots X_6) = P(X_1)\,P(X_2 \mid X_1) \dots P(X_6 \mid X_2, X_5)$$

- Each factor e.g. $P(X_6 \mid X_2, X_5)$ comes from a local probability table (LPT)



- DGM is a directed acyclic graph (DAG)
  - nodes ≡ random variables
  - edges ≡ "parenthood"
    - $\pi_i \triangleq \text{parents}(x_i)$
    - e.g. $\pi_6 = \{2, 5\}$
- It is defined WRT a specific topological ordering of the variables. (in this case $x_1, x_2, \dots, x_6$)
- The joint distribution is:

$$P(x_1 \dots x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$$

- Each $p(x_i \mid x_{\pi_i})$ is an LPT
- Total # of cells : $\sum_{i=1}^{n} k_i \prod_{j \in \pi_i} k_j$
- e.g. $k_1 = \dots = k_n = 2$

$\hookrightarrow 2^n$ vs. $\sum_{i=1}^{n} 2^{|\pi_i|}$

(exponential in n vs. exponential in $|\pi_i|$)

Graph Separation

- A <u>subset</u> of the conditional indep. relations are encoded directly by <u>graph seperation</u>
- Define <u>non-parent ancestors</u>:

$$\nu_i \triangleq \text{ancestors}(x_i) \setminus \text{parents}(x_i)$$

- Graph seperation encodes:

$$X_i \perp\!\!\!\perp X_{\nu_i} \mid X_{\pi_i}$$

- e.g., show $X_5 \perp\!\!\!\perp X_1 \mid X_3$

$$P(x_5 \mid x_3, x_1) = \frac{P(x_5, x_3, x_1)}{\sum_{x_5} P(x_5, x_3, x_1)}$$

$$P(x_5, x_3, x_1) = P(x_5 \mid x_3)\, P(x_3 \mid x_1)\, P(x_1)$$

$$= \frac{P(x_5 \mid x_3)\, P(x_3 \mid x_1)\, P(x_1)}{\sum_{x_5} P(x_5 \mid x_3)\, P(x_3 \mid x_1)\, P(x_1)}$$

$$= \frac{P(x_5 \mid x_3)\, \cancel{P(x_3 \mid x_1)}\, \cancel{P(x_1)}}{\cancel{P(x_3 \mid x_1)}\, \cancel{P(x_1)}\, \underbrace{\sum_{x_5} P(x_5 \mid x_3)}_{=1}}$$

$$= P(x_5 \mid x_3) \checkmark$$

- Using graph seperation, we know:
  - $X_4 \perp\!\!\!\perp X_2 \mid X_1$
  - $X_5 \perp\!\!\!\perp X_1 \mid X_3$
  - $X_6 \perp\!\!\!\perp X_5, X_3 \mid X_1$
- Are these the only conditional indepencies among $X_1 \dots X_6$ implied by the graph? No.
- Why? It only reflects graph seperation for a single topological ordering $X_1 \cdots X_6$
- e.g. $P(X_1 \dots X_6) = P(X_6) P(X_4 \mid X_6) \dots$
- Nevertheless, a single DGM implies all conditional independencies via "d-seperation".

D-separation

- "directional" separation - Pearl (1988)

- 3 simple DAGs

  ① chain $\circ \to \circ \to \circ$

  ② tree
  $$\begin{tikzcd} \circ \arrow[r] \arrow[dr] & \circ \\ & \circ \end{tikzcd}$$

  ③ V-structure
  $$\begin{tikzcd} \circ \arrow[dr] & & \circ \arrow[dl] \\ & \circ & \end{tikzcd}$$

- shading $\equiv$ conditioning

  e.g. $\underset{X}{\circ} \to \underset{Y}{\bullet} \to \underset{Z}{\circ} \equiv P(X, Z \mid Y)$

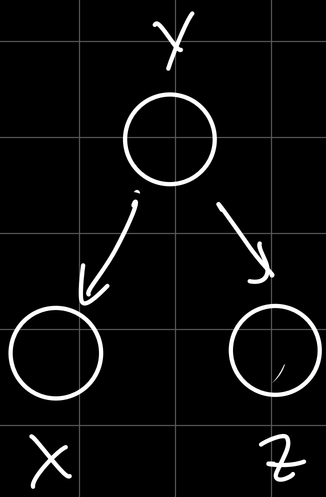
① Chain

$\underset{X}{\circ} \to \underset{Y}{\circ} \to \underset{Z}{\circ}$ $\quad Z \perp\!\!\!\perp X \mid Y$

(example of a Markov assumption)

future $\perp\!\!\!\perp$ past $\mid$ present

② <u>Tree</u>



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

$$X \perp\!\!\!\perp Z \,?$$

$$P(x|z) = \frac{\sum_y p(y)\, p(x|y)\, p(z|y)}{\sum_x \sum_y p(y)\, p(x|y)\, p(z|y)}$$

no.

$$X \perp\!\!\!\perp Z \mid Y \,?$$

$$P(X|z,y) = \frac{p(y)\, p(x|y)\, p(z|y)}{\sum_x p(y)\, p(x|y)\, p(z|y)}$$

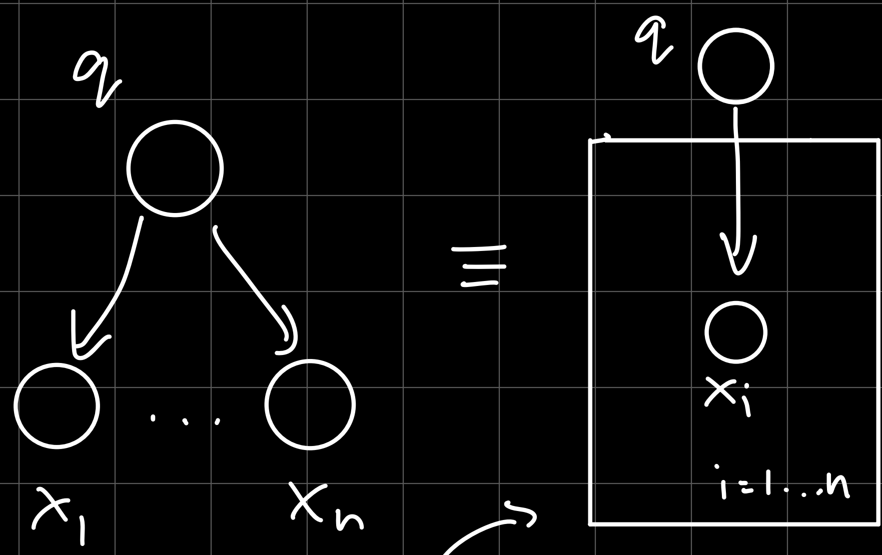
$$= \frac{p(\cancel{y})\, p(x|y)\, p(\cancel{z|y})}{p(\cancel{y})\, p(\cancel{z|y})}$$

$$= p(x|y) \checkmark$$

Example: repeated iid sampling

$q \sim \mathrm{Beta}(\alpha, \beta)$

$x_i \overset{iid}{\sim} \mathrm{Bern}(q)$



$P(x_1 \ldots x_n \mid q)$

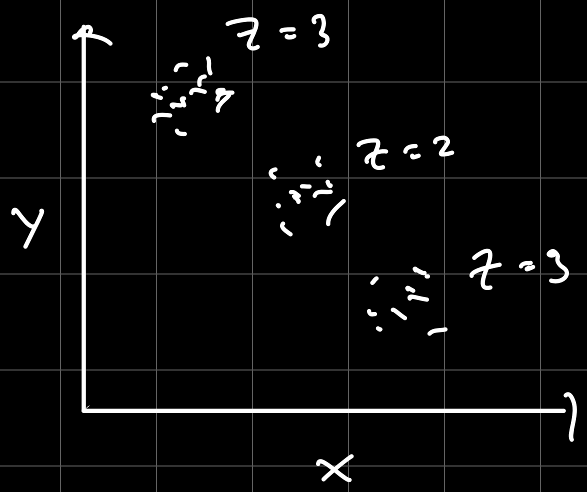
$= \prod_{i=1}^{n} p(x_i \mid q)$

plates denote repetition

$$p(x_1 \ldots x_n) = \underbrace{p(x_1)}_{\mathrm{Bern}\left(\frac{\alpha}{\alpha+\beta}\right)} \underbrace{p(x_2 \mid x_1)}_{\mathrm{Bern}\left(\frac{\alpha + x_1}{\alpha+\beta+1}\right)} \cdots$$

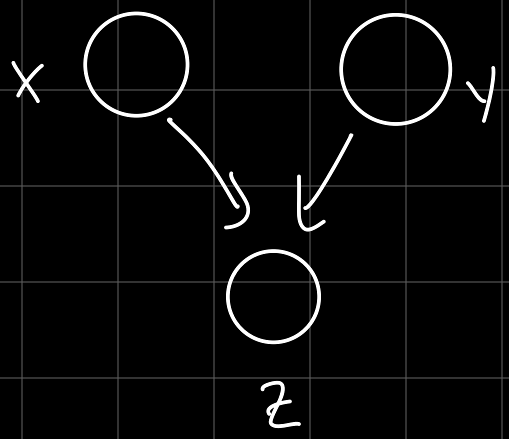
cond. indep $\Rightarrow$ marginal dependence

Example: confounding



$x \perp\!\!\!\perp y \mid z = 3$

③ V-structure

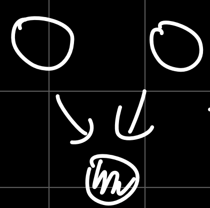


$$P(x, y, z) = P(x)\,P(y)\,P(z \mid x, y)$$

$x \perp\!\!\!\perp y$?

$$p(x \mid y) = \frac{\cancel{\sum_z} P(x)\,\cancel{P(y)}\,\cancel{P(z \mid x, y)}}{\cancel{\sum_x \sum_z P(x)\,P(y)}\,\cancel{P(z \mid x, y)}}$$

$$= p(x) \checkmark$$



$x \perp\!\!\!\perp \mid y \mid z$?

$$p(x \mid y) = \frac{\sum_z P(x)\,P(y)\,P(z \mid x, y)}{\sum_x \sum_z P(x)\,P(y)\,P(z \mid x, y)}$$

No.

e.g. <u>"explaining away"</u>

x = COVID

y = common cold

z = coughing

- $x \perp\!\!\!\perp y \mid z$ does not come from graph seperation...

Bayes ball

- a general algorithm to determine if

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

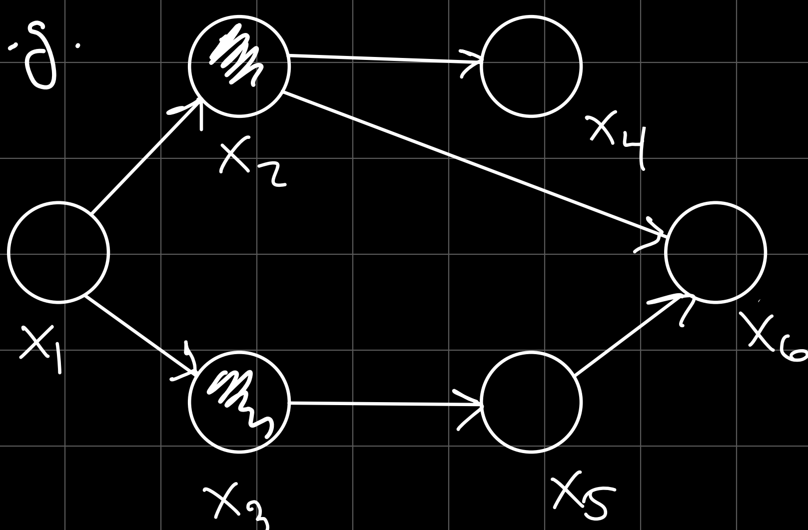
for subsets A, B, C of nodes in a DGM.

- based on <u>reachability</u>: if a ball <u>cannot</u> bounce from X_A to X_B when X_C is observed, then $X_A \perp\!\!\!\perp X_B \mid X_C$

- <u>The Rules</u>

✓	$\rightarrow \bigcirc \rightarrow$	x	$\rightarrow \bullet \rightarrow$
✓	$\leftarrow \bigcirc \rightarrow$	x	$\leftarrow \bullet \rightarrow$
x	$\rightarrow \bigcirc \leftarrow$	✓	$\rightarrow \bullet \leftarrow$

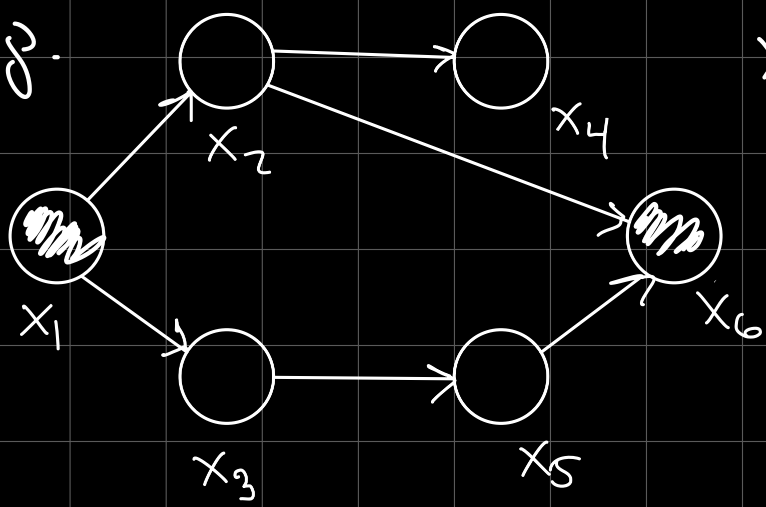
e.g.



$X_1 \perp\!\!\!\perp X_6 \mid X_2, X_3$?

yes

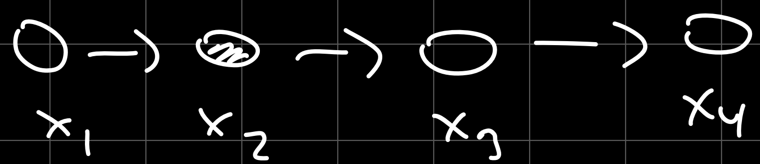
e.g.



$x_2 \perp\!\!\!\perp x_3 \mid x_1, x_6$?

no

e.g.

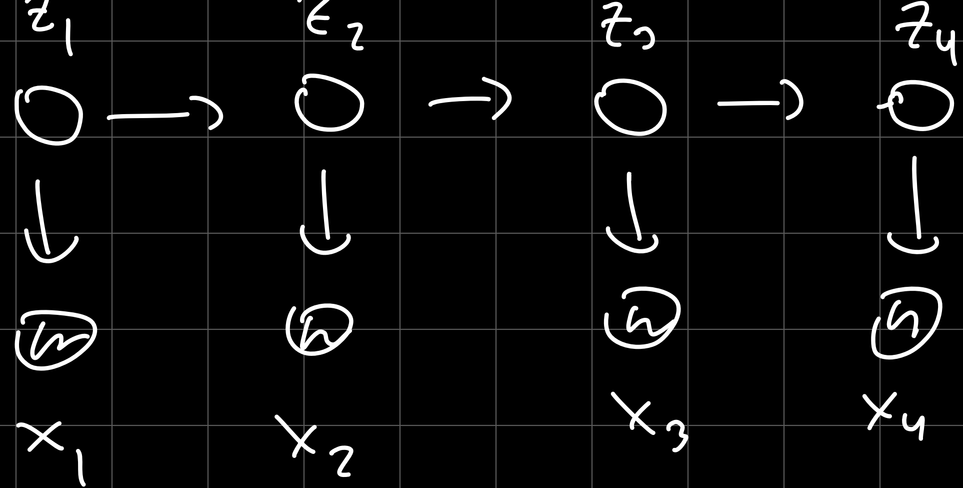


Markov chain

$x_3 \perp\!\!\!\perp x_1 \mid x_2$ (by design)

$x_4 \perp\!\!\!\perp x_1 \mid x_2$ (by Bayes ball)

e.g.



Hidden Markov model (HMM)

$x_1 \perp\!\!\!\perp x_4 \mid x_2, x_3$?

no.

Theorem (Hammersley - Clifford)

- $G = (V, E)$ is a DAG over nodes $V = \{x_1 \cdots x_n\}$

- $\mathcal{S}_1 = \{ p : p \text{ respects } G \}$

all joint dists $p = p(x_1 \cdots x_n)$ that respect all cond. independencies implied by G

- $\mathcal{S}_2 = \{ p_{G,\Phi} \text{ for all } \Phi \}$

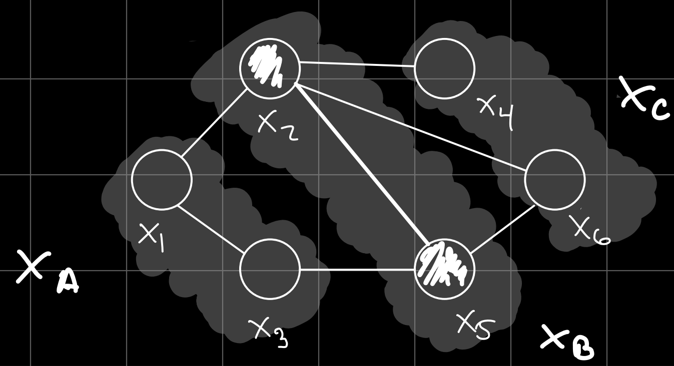
where Φ is a value for the CPTs in G

- Thm : $\mathcal{S}_1 = \mathcal{S}_2$

Undirected graphical models

- "Markov random fields" (MRFs)

- $X_A \perp\!\!\!\perp X_C \mid X_B$ iff X_B graph-separates X_A, X_C



- DGMs are acyclic and have an ordering; they therefore define the joint via LPTs and the chain rule

- How do UGMs parametrize the joint?

- Recall: $P(x_1 \dots x_n) = \dfrac{f(x_1 \dots x_n)}{Z}$ "kernel", "normalizer"

- $f(x_1 \dots x_n) = \prod_{c \in C} \varphi_c(x_c)$

- C are all maximal cliques in G

- $\varphi_c(\cdot)$ is the potential function for X_c

- $\varphi_c(x_c) > 0$ is the non-negative potential for configuration $X_c = x_c$

- These are like the LPTs, but they are not normalized

- <u>Maximal Cliques</u>:
  - connected components
  - $C = \{ \overset{c=1}{\{1,3\}}, \overset{2}{\{1,2\}}, \overset{3}{\{2,5,6\}}, \overset{4}{\{2,4\}} \}$
  - notice x_1 appears in $c=1$ and $c=2$
  - $\varphi_1(x_1, x_3)$, $\varphi_2(x_1, x_2)$, ...
  - φ_c measures the agreement of a clique

- $Z = \sum_{x_1} \cdots \sum_{x_n} \prod_{c} \varphi(x_c)$ "Partition function"
- Hard to compute
- "Energy" $\varphi(x_c) \triangleq \exp(-\underbrace{H_c(x_c)})$

$$P(x_1 \dots x_n) \triangleq \frac{1}{Z} \prod_{c \in C} \exp(-H_c(x_c))$$

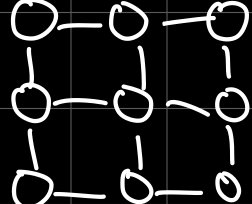
$$= \exp(-\underbrace{\sum_c H(x_c)}_{= H(x_1 \dots x_n)} - \log Z)$$

"Boltzmann distribution"

- low-energy configurations are more probable.

e.g., Ising model $H(x_1 \dots x_n) = \sum_{ij} \underbrace{x_i x_j c_{ij}}_{\varphi(x_i, x_j)}$

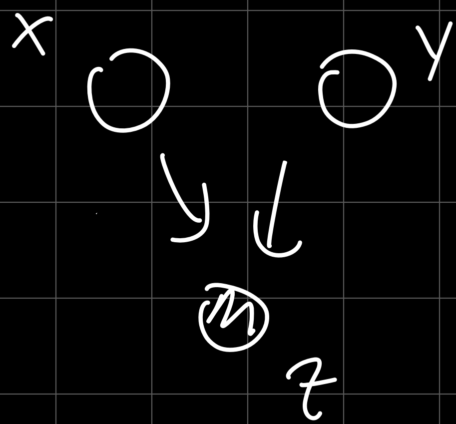
$x_{ij} \in \{-1, 1\}$ spin



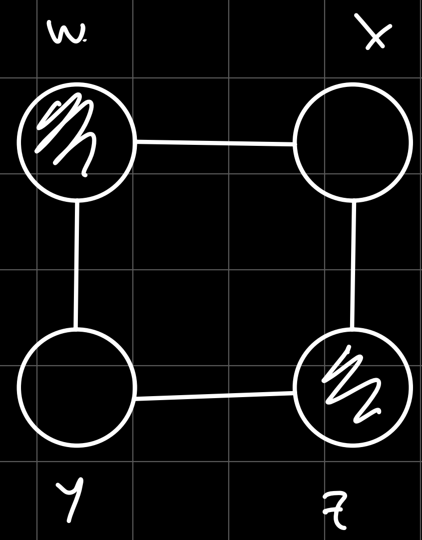
DGMs vs. UGMs

- graph → independence — dependence → graph
- good for generative — good for complex dependence

- not all <u>families</u> can be expressed by both DGMs and UGMs



vs.



$X \perp\!\!\!\perp Y$ and $X \not\!\perp\!\!\!\perp Y \mid Z$

(no way to express this with an UGM)

$Y \perp\!\!\!\perp X \mid W, Z$

$W \perp\!\!\!\perp Z \mid X, Y$

(no way to express with DGM)

Note for a <u>given</u> CPT:

$$P(X=1 \mid Y=1, Z) = p$$

$$P(X=1 \mid Y=0, Z) = p \qquad \forall Z$$

$$\Downarrow$$

$$X \perp\!\!\!\perp Y \mid Z$$