10708
Probabilistic Graphical Models: Fall 2022

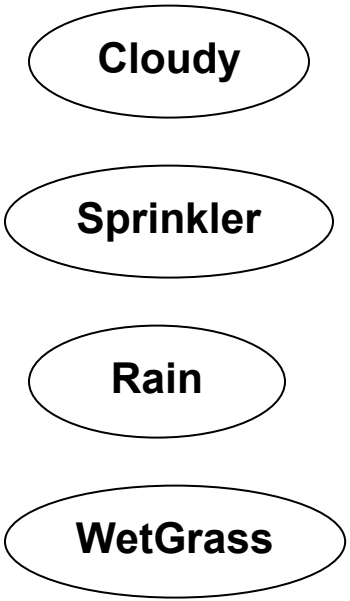
Andrej Risteski

Machine Learning Department

Lecture 2:
Directed and undirected graphical models: definitions and motivations

Basic motivation: representing high-dimensional distributions

How much memory do we need to represent a (general) distribution over 4 variables, each taking 2 possible values?



Basic motivation: representing high-dimensional distributions

More generally, if we have a distribution over d variables, each taking c possible values, we need a table of size c^d : prohibitively large for large d

However, in general, many distributions we are interested will have some *structure*:

(1) Instead of lookup table, we may want to have postulate some compact functional form.

(2) Some variables have some *correlation*: it's much more likely that it's cloudy if it rains.

Basic motivation: representing high-dimensional distributions

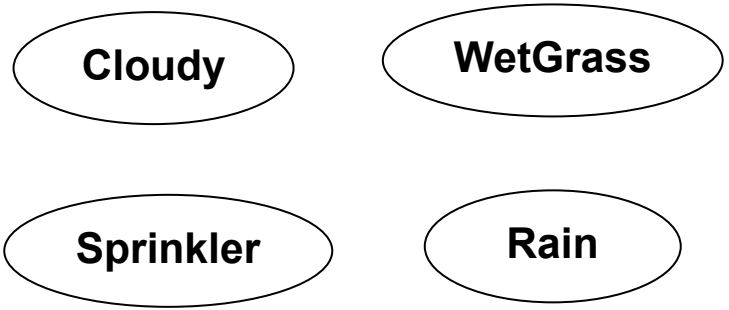
Instead of lookup table, we can parametrize distribution by

$$p_\theta(x) \propto \exp(-E_\theta(x))$$

where $E_\theta(x)$ has some compact form.

$E_\theta(x)$ can be seen as "energy" or "soft constraint": tells us what configurations are "lower energy" and the distribution prefers.

$E_\theta(x)$ can "respect" the structure of a graph:



$$p_\theta(x) \propto \exp\left(\sum_{ij} \phi_{ij}(x_i, x_j)\right)$$

Basic motivation: representing high-dimensional distributions

More generally, if we have a distribution over d variables, each taking c possible values, we need a table of size c^d : prohibitively large for large d

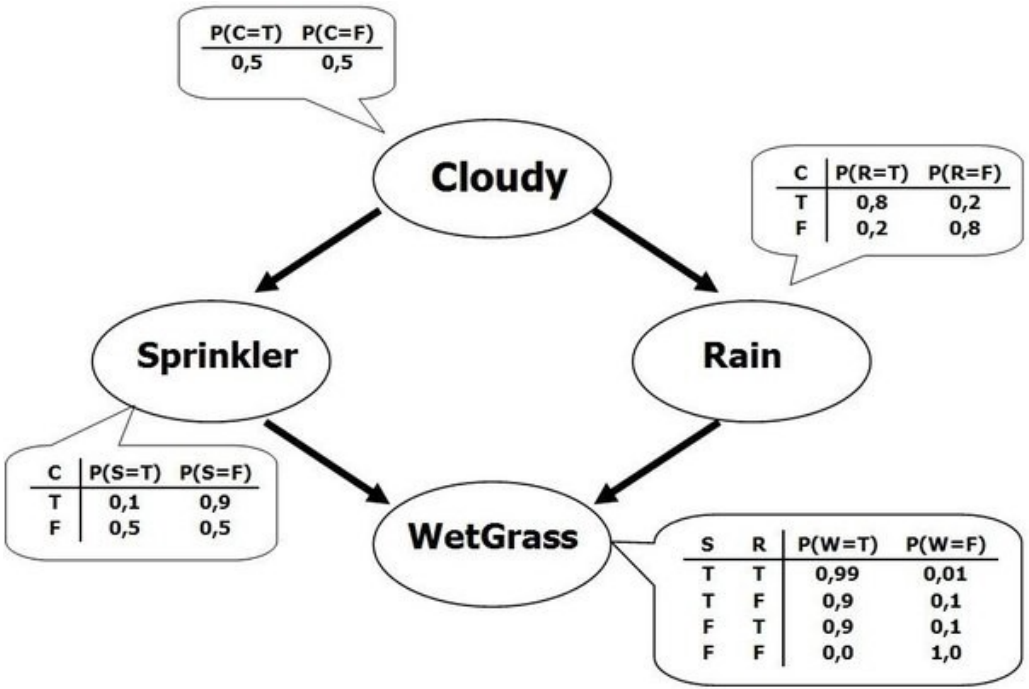
However, in general, many distributions we are interested will have some *structure*:

(1) Instead of lookup table, we may want to have postulate some compact functional form.

(2) Some variables have some *correlation*: it's much more likely that it's cloudy if it rains.

(3) We may have some mechanistic knowledge of domain: a sprinkler *causes* the grass to be wet.

Basic motivation: representing high-dimensional distributions



Basic motivation: representing high-dimensional distributions

More generally, if we have a distribution over d variables, each taking c possible values, we need a table of size c^d : prohibitively large for large d

However, in general, many distributions we are interested will have some *structure*:

(1) Instead of lookup table, we may want to have postulate some compact functional form.

(2) Some variables have some *correlation*: it's much more likely that it's cloudy if it rains.

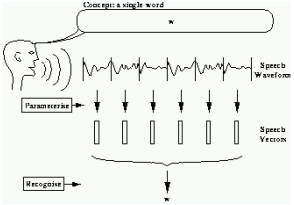
(3) We may have some mechanistic knowledge of domain: a sprinkler *causes* the grass to be wet.

(4) Some variables are *independent*: the sprinkler being on and rain are independent events.

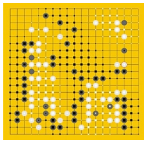
Applications of PGMs

- Computer vision and graphics
- Natural language processing
- Information retrieval
- Robotic control
- Computational biology
- Genetics and medical diagnosis/prognosis
- Finance and economics
- Decision making under uncertainty
- Social network analysis
- …

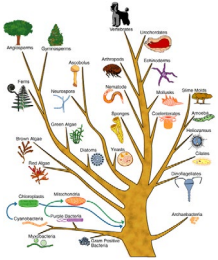
Computer vision



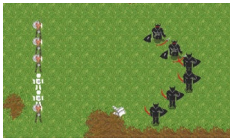
Speech recognition



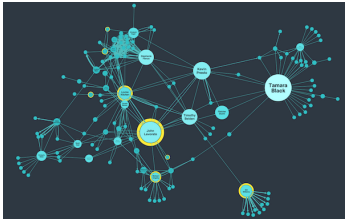
Games



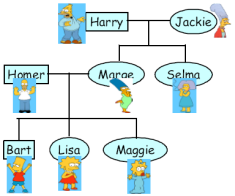
Evolution



Planning



Social network analysis



Pedigree

Robotic control

The three pillars

Representation: is the representation compact and/or captures some structural semantics about distribution of choice.

> The latter is a little open-ended: we'll see things like conditional dependence, maximum-entropy principles, etc.

Inference: can we efficiently draw samples from the model and/or calculate marginals in the model.

> "What's the probability that a patient has a disease, given some symptoms they exhibit?"

"Chess match between Albert Einstein and Abraham Lincoln, dim moody lighting, photorealistic "

The three pillars

Representation: is the representation compact and/or captures some structural semantics about distribution of choice.

> The latter is a little open-ended: we'll see things like conditional dependence, maximum-entropy principles, etc.

Inference: can we efficiently draw samples from the model and/or calculate marginals in the model.

Learning: can the model be fit from data in an efficient manner.

> What loss do we optimize? How do we optimize it? Can we use gradient-based methods to do so?

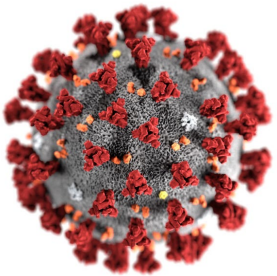
Part I: Directed graphical models

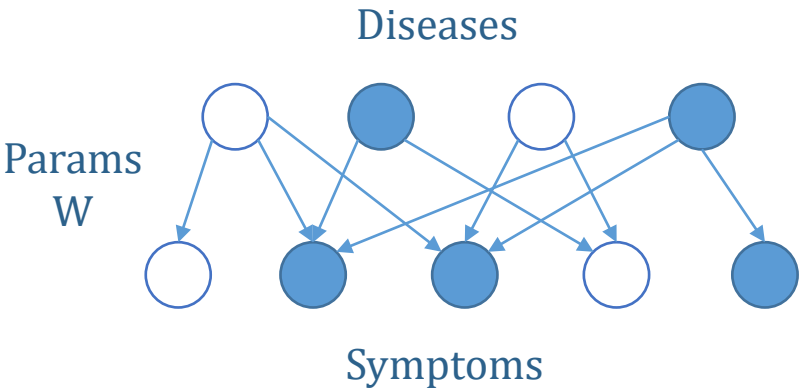
First view of directed graphical models: causal relationships

Directed Graphs are useful for expressing *directional/causal* relationships between random variables.

Your symptoms: fever + loss of taste.

Probability that you have covid?



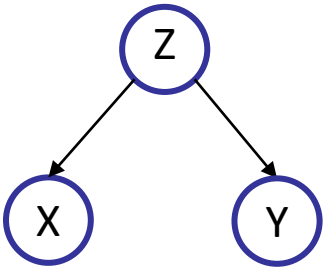


Directed graphical model succinctly describes Pr[symptom| diseases]

First view of directed graphical models: causal relationships

Directed Graphs are useful for expressing *directional/causal* relationships between random variables.

- X = height of a child
- Y = vocabulary of a child
- Z = age of a child



First view of directed graphical models: causal relationships

Directed Graphs are useful for expressing *directional/causal* relationships between random variables.

Is direction somehow “inherent” or “unique” ?

First view of directed graphical models: causal relationships

•Where does the graph come from?

- Prior knowledge of causal relationships
- Prior knowledge of modular relationships
- Assessment from experts
- Learning from data
- We simply prefer a certain structure (e.g., a layered graph)
- …

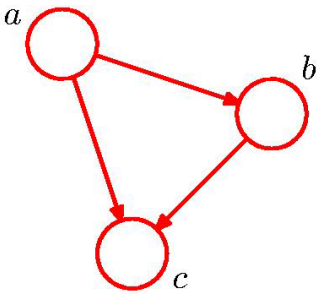
Second view: restricted conditional factorization

"Deriving" Bayesian Networks as restrictions of arbitrary distributions:

An **arbitrary** joint distribution $p(a, b, c)$ over three random variables a,b, and c can be written as

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

Associate a graph with the decomposition:



- Node for each of the random variables.
- Add **directed** links to the graph from the nodes corresponding to the vars on which the distribution is conditioned.

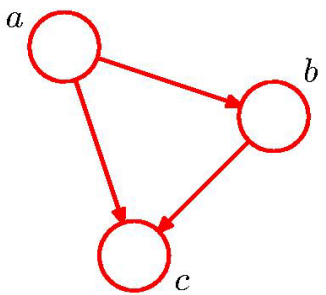
Second view: restricted conditional factorization

"Deriving" Bayesian Networks as restrictions of arbitrary distributions:

An **arbitrary** joint distribution $p(a, b, c)$ over three random variables a,b, and c can be written as

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

Associate a graph with the decomposition:



Different ordering => different graphical representation.

Joint distribution over K variables factorizes:

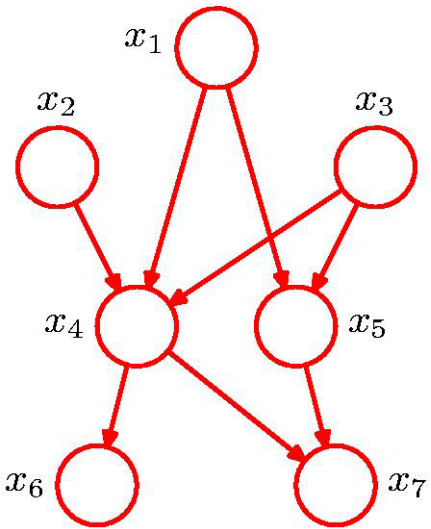
$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

Corresponding undirected graph is fully connected:

(as each lower-numbered node points to each higher-numbered node)

Second view: restricted conditional factorization

A graph that is **not** fully connected conveys information about the conditional factorization of the distribution it encodes.



E.g. consider the graph on the left.

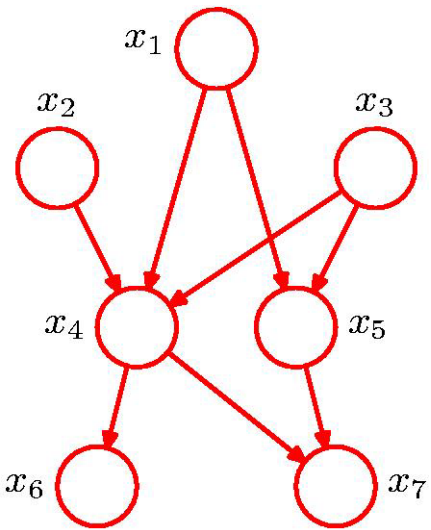
It encodes distributions over $x_1,\ldots,x_7$ that can be written as the product:

$$p(x_1,\ldots,x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)\, p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

Note the change from the previous slide: e.g. x_5 is **not** conditioned on all of x_1, x_2, x_3, x_4 but only on x_1, x_3.

Second view: restricted conditional factorization

The joint distribution defined by the graph is given by the product of a conditional distribution for each node conditioned on its parents:



$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

where pa_k denotes a set of parents for the node x_k.

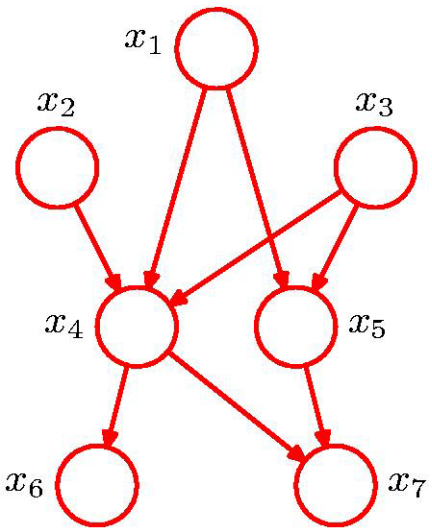
Each of the conditional distributions will typically have some parametric form. (e.g. product of Bernoullis in the noisy-OR case)

Important restriction: There must be **no directed cycles!** (i.e. graph is a DAG)

Third view: generative model

Consider a joint distribution over K random variables $p(x_1, x_2, ..., x_K)$t factorizes as:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$



Suppose each of the conditional distributions are **easy to sample from**. How do we sample from the joint?

Start at the top and sample in order.

$$\hat{x}_1 \sim p(x_1)$$
$$\hat{x}_2 \sim p(x_2)$$
$$\hat{x}_3 \sim p(x_3)$$
$$\hat{x}_4 \sim p(x_4|\hat{x}_1, \hat{x}_2, \hat{x}_3)$$
$$\hat{x}_5 \sim p(x_5|\hat{x}_1, \hat{x}_3)$$

The parent variables are set to their sampled values

To obtain a sample from the marginal distribution, e.g . $p(x_2, x_5)$, sample from the full joint distribution, retain $\hat{x}_2, \hat{x}_5,$ discard the remaining values.

Compactness of representation

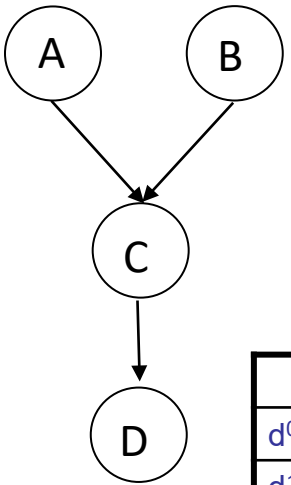
How much memory do we need to represent a DGM over a discrete space?

Simplest representation: probability tables for conditional distributions

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Compactness of representation

Consider random variables $X_1, X_2, \ldots, X_n$ where $X_i \in \mathcal{X}$, where $|\mathcal{X}| = r$

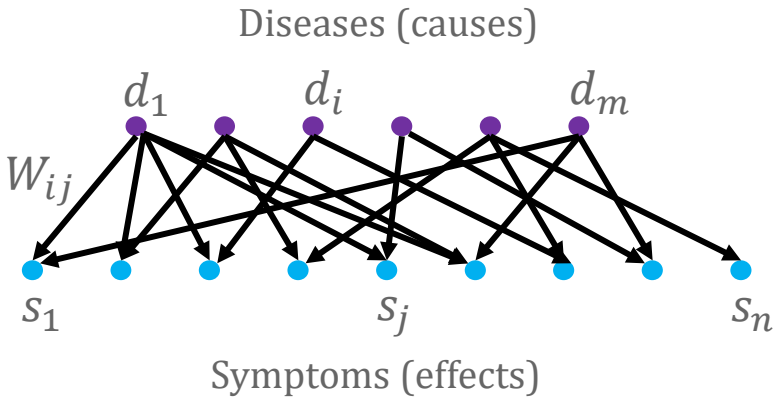
- **No DAG**: To represent an arbitrary distribution P(**X**) via a single joint probability table requires $r^n - 1$ values.
- If the distribution factors according to a graph G with

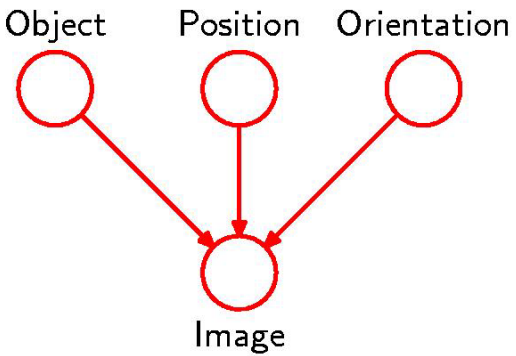
$$\max_{X_i} |\text{parents}(X_i)| \leq D$$

- then each $P(X_i \mid \text{parents}(X_i))$ needs only $r^D(r - 1)$ values
- for a total of only $n(r^D(r - 1))$ values.

The latent-variable paradigm

More often than not, we need to model part of the data that is **not observable**. We already saw examples of this:





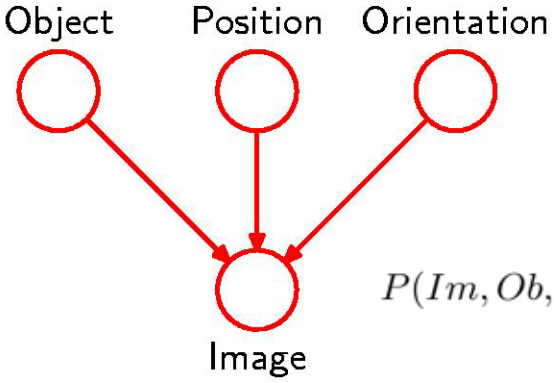
This is also a natural way to extract **features/representation**: the latent variables contain "meaningful" information.

The latent-variable paradigm

Higher-up nodes will typically represent **latent** (hidden) random variables.

The role of latent variables is to allow modeling a **complicated** distribution over observed variables **constructed** from **simpler** conditional distributions.

Latent-variable model of image



Object identity, position, and orientation have independent *prior probabilities*.

Image has probability distr that depends on object identity, position, and orientation (*conditional distribution/likelihood*).

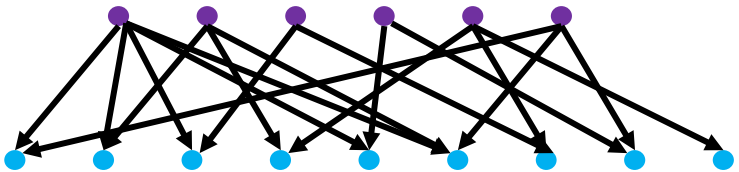
$$P(Im, Ob, Po, Or) = \underbrace{P(Im|Ob, Po, Or)}_{\text{Likelihood}} \underbrace{P(Ob)P(Po)P(Or)}_{\text{Prior}}$$

Likelihood and prior are modeled by parametric distribution whose parameters are fitted throughout training.

Examples: single-layer latent-variable Bayesian networks

Simple, but powerful paradigm:
single-layer Bayesian networks, where top nodes are latent.

Latent variables Z



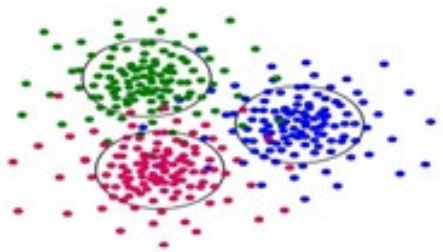
Observable variables X

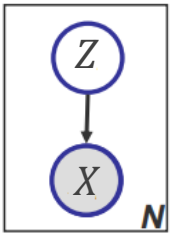
$$p_\theta(X, Z) = p_\theta(Z)\, p_\theta(X|Z)$$

Example 1: Mixture distributions

Mixture models: observables = points; latent = clustering

To draw a sample (X,Z):
Sample Z from a categorial distr. on K components with parameters $\{\pi_i\}$
Sample X from the corresponding component in the mixture.





$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

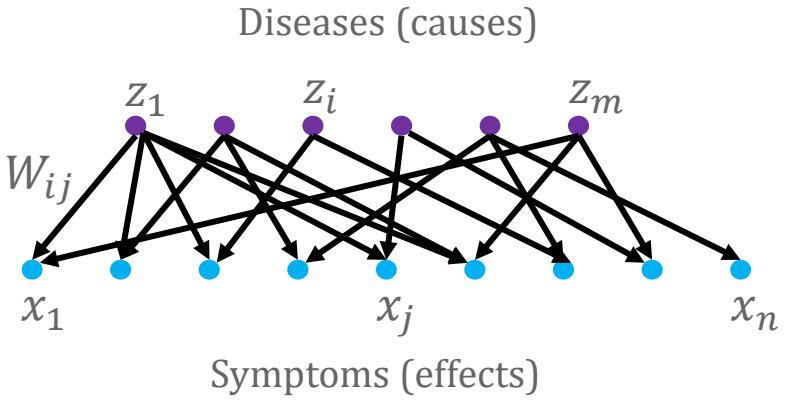
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

Mixing coefficient

Example 2: Noisy-OR networks

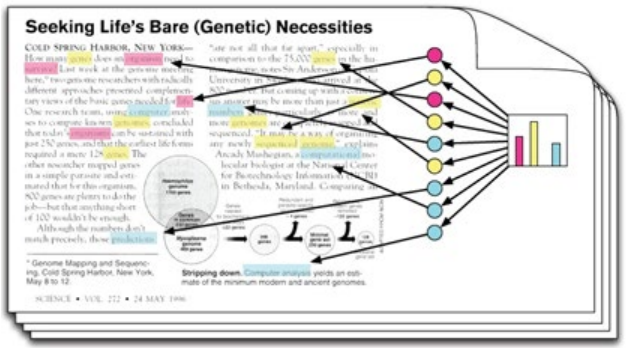
$x_i, z_j \in \{0,1\}$
$W_{ij} \geq 0$



- Sample each z_i is on **independently** with prob. ρ
- When z_i is on, it **activates** x_j with probability $1 - \exp(-W_{ij})$.
- x_j is **on** if one of z_i's **activates** x_j

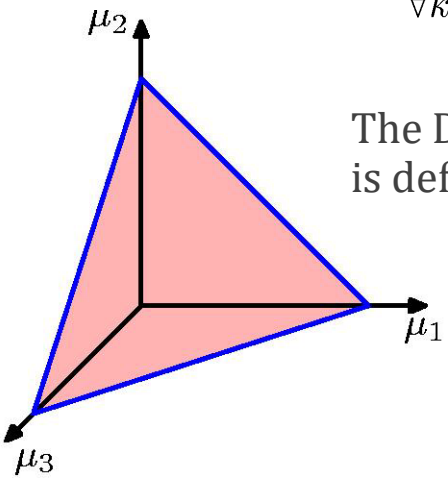
Example 3: Topic models (LDA)

Latent Dirichlet Allocation: famous model for modeling topic structure of documents of text. (Blei, Ng, Jordan '03)



Side-remark: Dirichlet Distribution

Consider a distribution over simplex, namely over points $\{\mu_i\}_{i=1}^K$



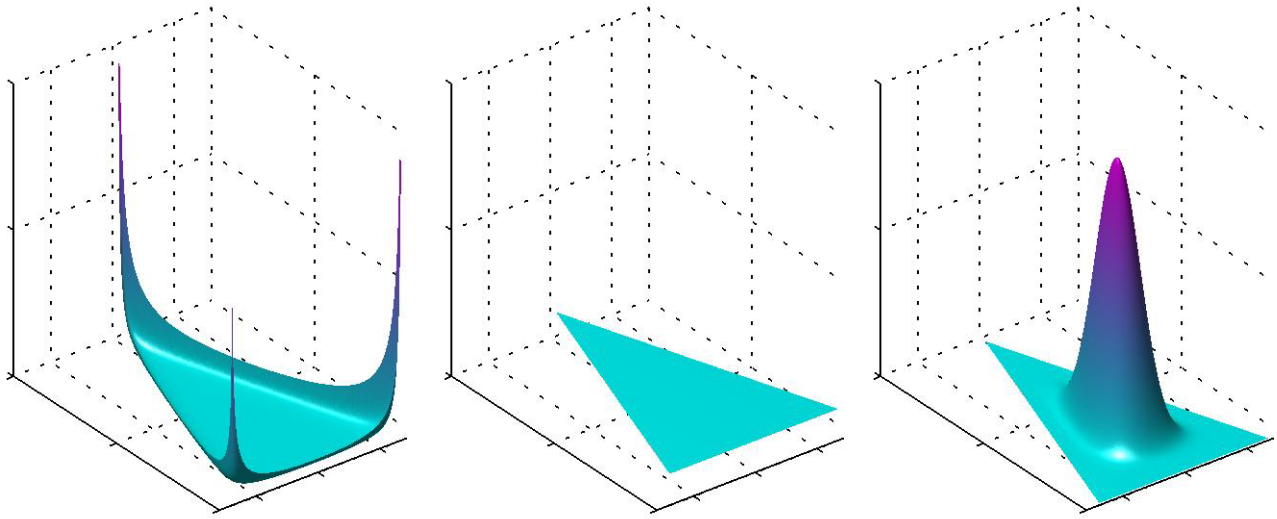
$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

The Dirichlet distribution (with params $\{\alpha_i \geq 0\}_{i=1}^K$) is defined as:

$$\mathrm{Dir}(\mu|\alpha) \propto \Pi_{k=1}^K \mu_k^{\alpha_k - 1}$$

Side-remark: Dirichlet Distribution

Plots of the Dirichlet distribution over three variables.



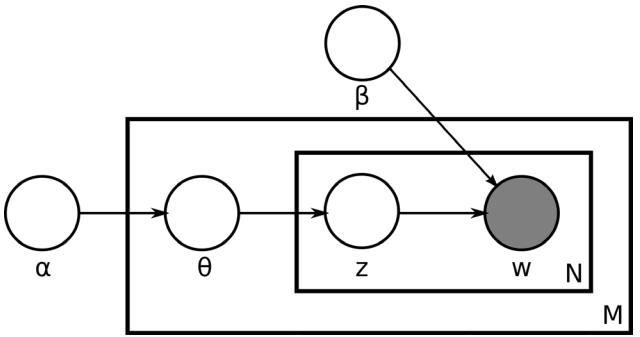
$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

Example 3: Topic models (LDA)

Defines a distribution over documents, involving K topics.

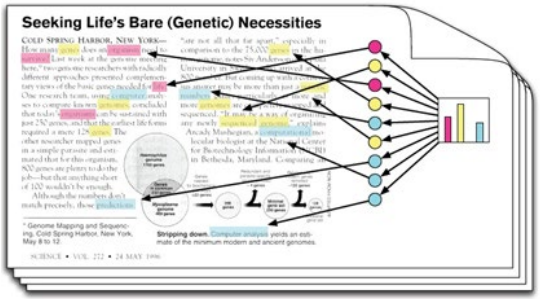
The **parameters** are: $\{\alpha_i\}_{i=1}^K$ (Dirichlet parameters) and matrix $\beta \in \mathbb{R}_+^{N\times K}$, where N is the size of the vocabulary.

The columns of β satisfy $\sum_{j=1}^N \beta_{ij} = 1$ (the distribution of words in a topic i)



To produce document:

- First, sample $\theta \sim \mathrm{Dir}(\cdot\,|\alpha)$: this will be the topic proportion vector for the document.
- Each word in the document is generated in order, independently.
- To generate word i:
  - Sample topic z_i with categorical distribution with parameters θ
  - Sample word w_i with categorical distribution with parameters β_{z_i}



Example 3: Topic models (LDA)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

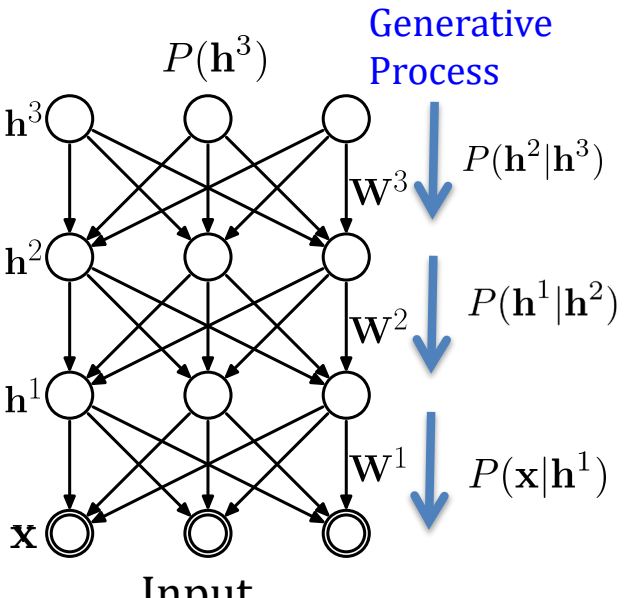
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

Example 4: Variational Autoencoder

Directed Bayesian network with Gaussian layers

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{h}^1,\ldots,\mathbf{h}^L} p(\mathbf{h}^L|\boldsymbol{\theta})p(\mathbf{h}^{L-1}|\mathbf{h}^L,\boldsymbol{\theta})\cdots p(\mathbf{x}|\mathbf{h}^1,\boldsymbol{\theta})$$

Each term may denote a complicated nonlinear relationship



Layers are parametrized as:

$$p(\boldsymbol{h}^{L-1}|\boldsymbol{h}^L,\boldsymbol{\theta}) = \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{h}^L), \Sigma_{\theta}(\boldsymbol{h}^L))$$

Gaussians, means/covariances functions (e.g. neural net) of previous layer and model parameters θ.

Easy to sample!

Example 4: Variational Autoencoder (VQ-VAE, Oord et al ‘17, Razavi et al ‘19)

Figure from Razavi et al ‘19

Example 4: Variational Autoencoder (DALL-E, Ramesh et al ’21)

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT

an armchair in the shape of an avocado [...]

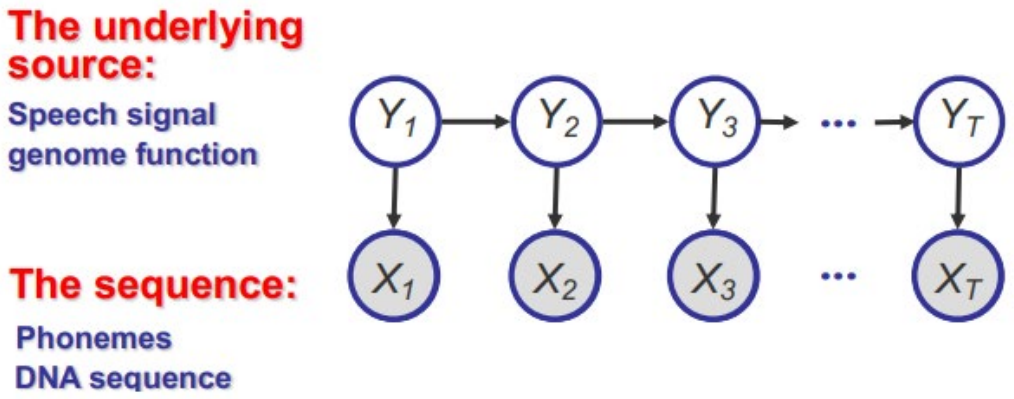
AI-GENERATED IMAGES

https://openai.com/blog/dall-e/

Example 4: Variational Autoencoder (NVAE, Vahdat-Kautz ’21)

Figure from Vahdat-Kautz ‘21

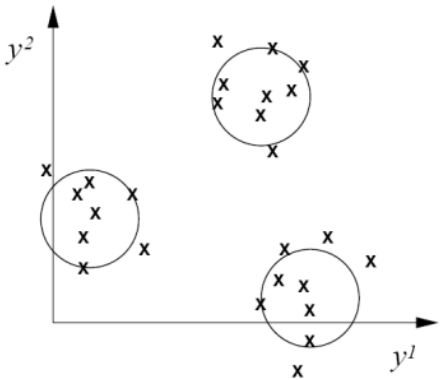
Example 4: Hidden Markov Models

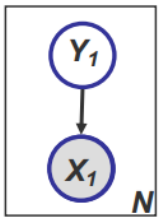


Need to only parametrize: $p(Y_t|Y_{t-1}), p(X_t|Y_t)$

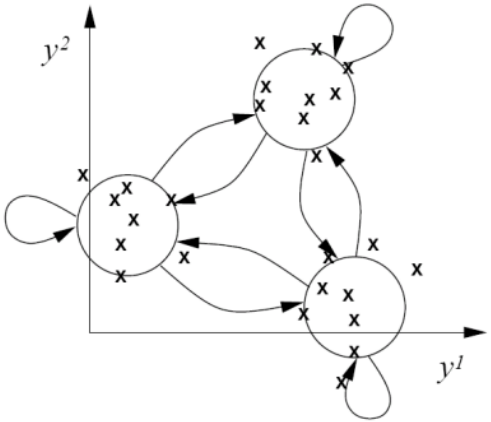
Example 4: Hidden Markov Models

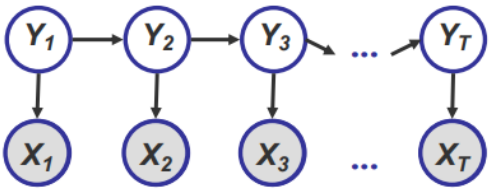
Static mixture





Dynamic mixture





Part II: Undirected graphical models

First view of undirected graphical models: soft constraints/energy

A frequent paradigm is for probabilistic models to have the form:

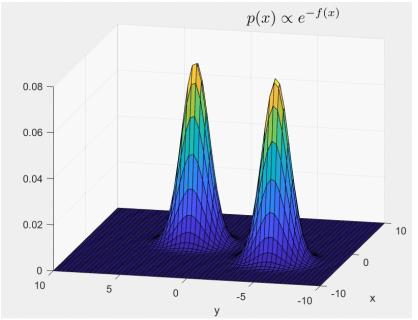
$$p_\theta(x) \propto \exp(-E_\theta(x))$$

where $E_\theta(x)$ has some easy to evaluate form.

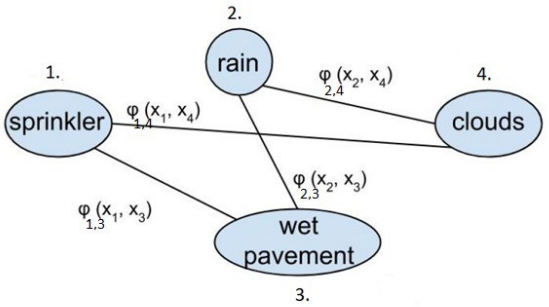
It's an easy way to convert an **energy** E_θ to a probability distribution.

$E_\theta(x)$ can be seen as "energy" or "soft constraint": tells us what configurations are "lower energy" and the distribution prefers.

Furthermore, by scaling $E_\theta(x)$ you can regulate the "sharpness" of the distribution.)



First view of undirected graphical models: soft constraints/energy



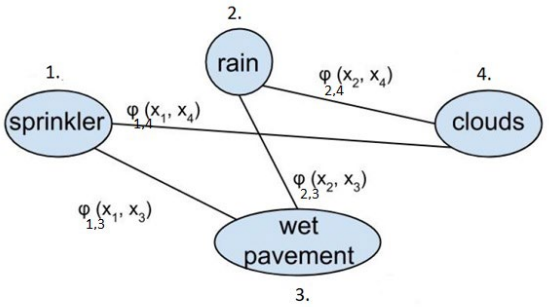
A *pairwise* **undirected graphical model** expresses a distribution as product of local potentials ϕ_{ij} (interactions), for example

$$p(x) \propto \exp\left(\sum_{ij} \phi_{ij}(x_i, x_j)\right)$$

"Soft constraint": the distribution tries to find a *good balance* in satisfying the (possibly conflicting) influences of the potentials.

Not clear how to draw samples efficiently…
(It will turn out to be hard to do so in general.)

First view of undirected graphical models: soft constraints/energy



A *pairwise* **undirected graphical model** expresses a distribution as product of local potentials ϕ_{ij} (interactions), for example

$$p(x) \propto \exp\left(\sum_{ij} \phi_{ij}(x_i, x_j)\right)$$

"Soft constraint": the distribution tries to find a *good balance* in satisfying the (possibly conflicting) influences of the potentials.

Partition function: $Z := \sum_x \exp\left(\sum_{ij} \phi_{ij}(x_i, x_j)\right)$

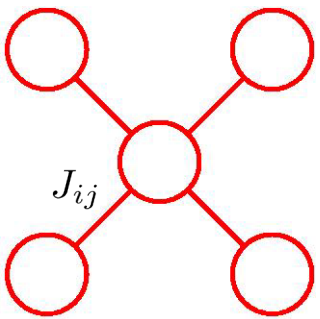
Naively calculating above quantity takes time $O(2^d)$

Example 1: multivariate Gaussian

Recall the multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The term inside the exponential is **quadratic**: namely, we can write



$$P(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{g}^T \mathbf{x}),$$

where $J = \Sigma^{-1}, \qquad \mu = J^{-1}\mathbf{g}.$

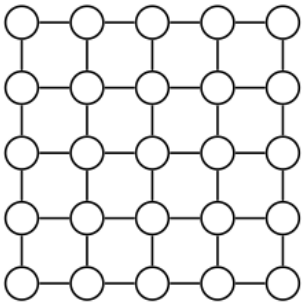
$$\mathbf{x}^T J \mathbf{x} = \sum_i J_{ii} x_i^2 + 2 \sum_{ij \in E} J_{ij} x_i x_j,$$

Thus, the interactions are given by the precision matrix J.

(Note: precision mx being sparse **does not** imply the covariance mx is sparse.)

Example 2: Ising models

MRFs with binary variables are sometimes called Ising models in statistical mechanics, and Boltzmann machines in machine learning literature.



Denoting the binary valued variable at node j by $x_j \in \{\pm 1\}$, the **Ising model** for the joint probabilities is given by:

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \Big)$$

The conditional distribution is given by logistic (*only depends on nbrhood*!):

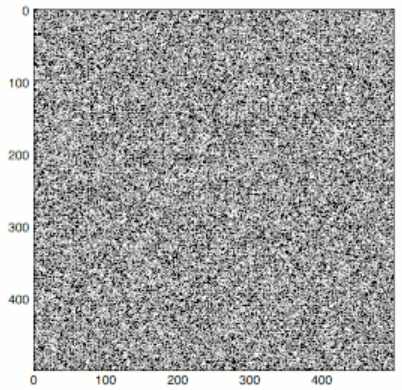
$$P_\theta(x_i = 1 | \mathbf{x}_{-i}) = \frac{1}{1 + \exp(-\theta_i - \sum_{ij \in E} x_j \theta_{ij})},$$

where x_{-i} denotes all nodes except for i.

If $\theta_{ij} \geq 0$: the nodes i,j, prefer to be the same. If $\theta_{ij} \leq 0$: they prefer to be different.

Example 3: Ferromagnetic Ising models

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \Big)$$

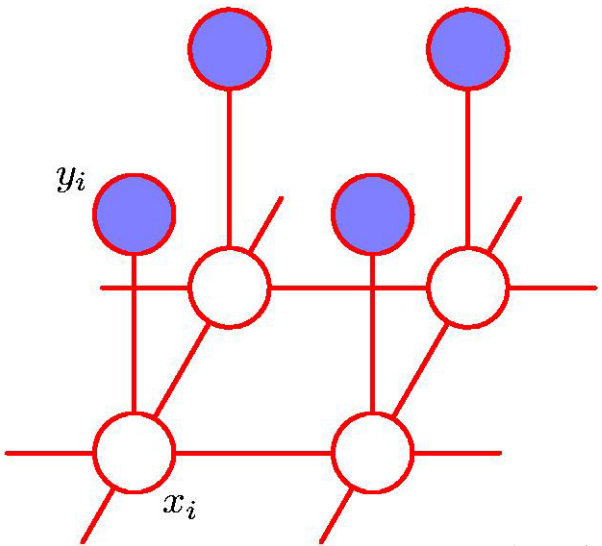


If $\theta_{ij} \geq 0$: the model is called ferromagnetic, and is used in physics to model the atomic structure (spins) of iron.

Example 4: Image Denoising

Noise removal from a binary image:

Let the observed noisy image be described by an array of binary pixel values: $y_j \in \{-1, +1\}$, i=1,...,D.



We take a noise-free image $x_j \in \{-1, +1\}$, and randomly flip the sign of pixels with some small probability.

Bias term

Neighboring pixels are likely to have the same sign

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

Noisy and clean pixels are likely to have the same sign

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

Example 5: Continuous-space energy-based models (EBMs)

If the distribution $p_\theta(x) \propto \exp(-E_\theta(x))$ has as domain $\mathbb{R}^d$, an easy choice is E_θ is a neural net of some kind.

These have scaled up *only very recently* to real-life data, e.g. images.

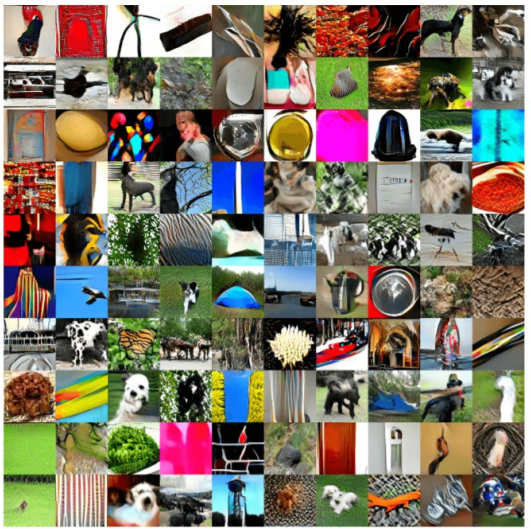


Figure 18: MCMC samples from conditional ImageNet128x128 models

Figure from (Du, Mordatch '20)

Second view: maximum entropy principle

Suppose we know, for a distribution p, only some statistics:

$\mathbb{E}_p \phi_i(x) = \mu_i, i \in N$ where ϕ_i is some function.

Example: $\phi_i(x) = x_i$ is the mean of the i-th coordinate.

What is the distribution that "assumes the least" other than p matching these statistics?

Principle of maximum entropy (Jaynes, '57): pick the p that maximizes H(p), subject to matching these statistics.

(Aka Occam's razor.)

What is this distribution p?

Second view: maximum entropy principle

We are trying to solve the optimization problem:

$$\max_p H(p), \text{s.t.} \quad \mathbb{E}_p \phi_i(x) = \mu_i, i \in N$$

(The variables are values of $p(x), x \in \mathcal{D}$ for a discrete-space distribution.)

Can be rewritten in the Lagrangian form:

$$\max_{p,\lambda_i,\lambda_0} \left\{ H(p) + \sum_i \lambda_i \left(\mathbb{E}_p \phi_i(x) = \mu_i \right) + \lambda_0 \left(\sum_i p_i - 1 \right) \right\}$$

Second view: maximum entropy principle

$$\max_{p,\lambda_i,\lambda_0} \left\{ H(p) + \sum_i \lambda_i \left(\mathbb{E}_p \phi_i(x) = \mu_i \right) + \lambda_0 \left(\sum_i p_i - 1 \right) \right\}$$

Taking derivatives:

$$\frac{\partial}{\partial \lambda_0} = 0: \quad \sum_i p_i = 1 \qquad \frac{\partial}{\partial \lambda_i} = 0: \quad \mathbb{E}_p \phi_i(x) = \mu_i$$

$$\frac{\partial}{\partial p(x)} = 0: \; -\log p(x) - 1 + \sum_i \lambda_i \phi_i(x) + \lambda_0 = 0$$

$$\Rightarrow p(x) \propto \exp\left(\sum_i \lambda_i \phi_i(x) \right)$$

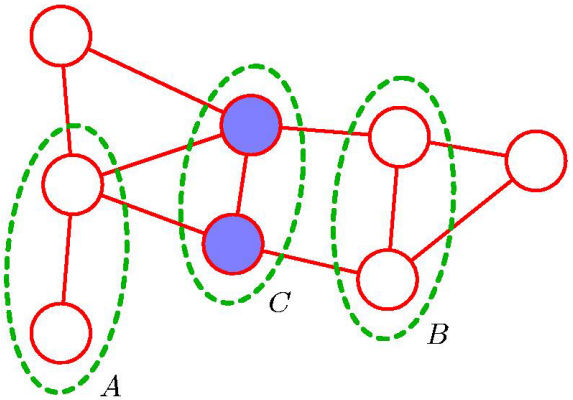
Second view: maximum entropy principle

$$p(x) \propto \exp\left(\sum_i \lambda_i \phi_i(x)\right)$$

In English: The distribution with potentials ϕ_i appropriately weighed, has maximum entropy given the values of the expectations of the potentials.

The potentials $\{\phi_i\}$ are also called sufficient statistics, and the above family of distributions an exponential family (w/ sufficient statistics $\{\phi_i\}$)

Third view: conditional independence



$x_A \perp\!\!\!\perp x_B | x_C$

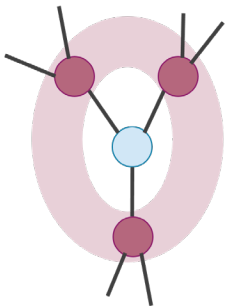
Global Markov Property: Consider pairwise UGM. The independence structure of variables is "captured" by the graph.

Nodes in A, B are independent, given a set of nodes C separating A, B

$$p(x_A \mid x_C, x_B) = p(x_A|x_C)$$

Equivalently:

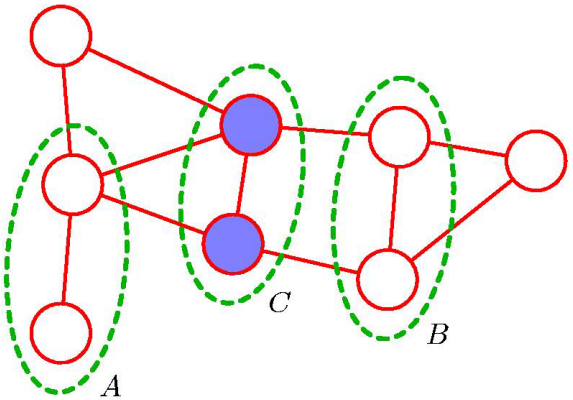
$$p(x_A, x_B \mid x_C) = p(x_A|x_C)\, p(x_B|x_C)$$



Special case (Local Markov Property): node is independent of the rest of the graph, given values of the neighbors

$$p(x_v \mid x_{N(v)}, x_{V/\{N(v),v\}}) = p(x_v|x_{N(v)})$$

Third view: conditional independence



$$x_A \perp\!\!\!\perp x_B | x_C$$

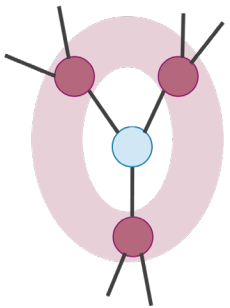
Global Markov Property: Consider pairwise UGM. The independence structure of variables is "captured" by the graph.

Nodes in A, B are independent, given a set of nodes C separating A, B

$$p(x_A \mid x_C, x_B) = p(x_A | x_C)$$

Equivalently:

$$p(x_A, x_B \mid x_C) = p(x_A | x_C)\, p(x_B | x_C)$$



The neighbors of v are the (minimal) **Markov Blanket** of v: the smallest set of nodes S, s.t. v is conditionally independent of all other nodes, given S.

Third view: conditional independence

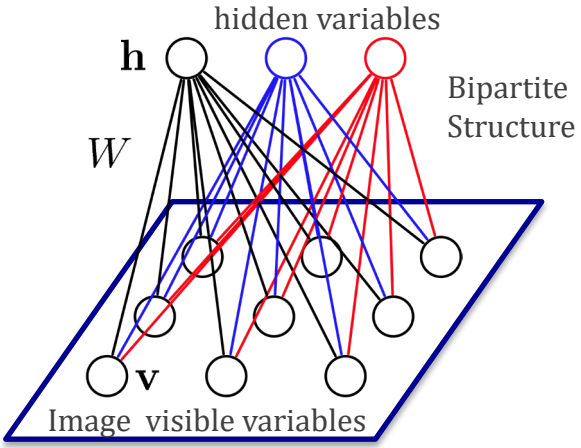
Claim: If A, B are separated by C, then $p(x_A, x_B | x_C) = p(x_A|x_C)\, p(x_B|x_C)$

Example:

Example 6: Restricted Boltzmann Machines

A **latent-variable model**: some of the variables the distribution models are not observed (hidden).

We denote visible and hidden variables with vectors **v, h** respectively:



Visible variables $\mathbf{v} \in \{0,1\}^D$
are connected to hidden variables $\mathbf{h} \in \{0,1\}^F$

.

The energy of the joint configuration:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j$$

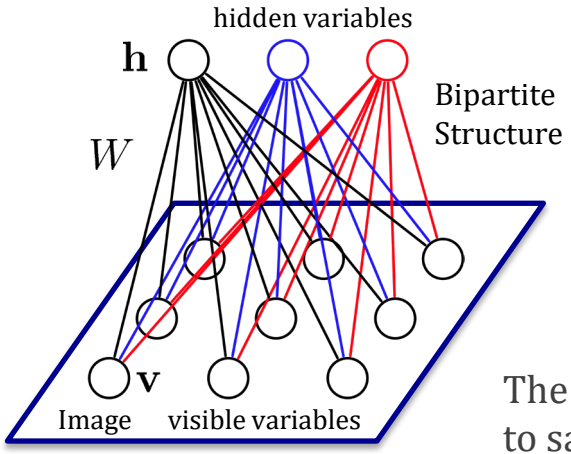
$\theta = \{W, a, b\}$ model parameters.

Probability of the joint configuration is given by the Boltzmann distribution:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big) = \frac{1}{\mathcal{Z}(\theta)} \prod_{ij} e^{W_{ij} v_i h_j} \prod_i e^{b_i v_i} \prod_j e^{a_j h_j}$$

$$\mathcal{Z}(\theta) = \sum_{\mathbf{h},\mathbf{v}} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big)$$

Example 6: Restricted Boltzmann Machines



Restricted: No interaction between hidden variables

The **posterior** over the hidden variables is easy to sample from! (Conditional independence!)

$$P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v}) \qquad P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij} v_i - a_j)}$$

Factorizes

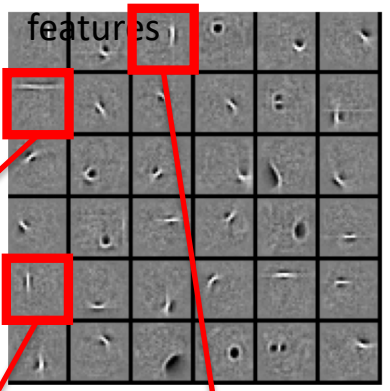
Similarly:

$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) \qquad P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\sum_j W_{ij} h_j - b_i)}$$

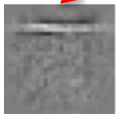
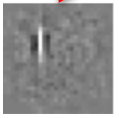
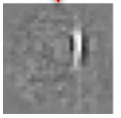
Example: learning alphabets using RBMs

Observed Data
Subset of 25,000 characters

Learned W: "edges"
Subset of 1000 features



New Image:

$= \sigma\Big(0.99 \times$  $+ \; 0.97 \times$  $+ \; 0.82 \times$  $\cdots\Big)$

$p(h_7 = 1|v)$ $p(h_{29} = 1|v)$

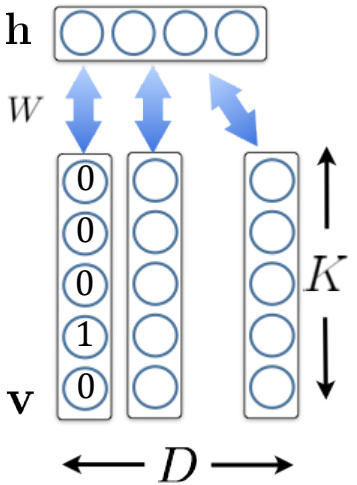
Most hidden variables are off

$\sigma(x) = \frac{1}{1+\exp(-x)}$

Logistic Function: Suitable for modeling binary images

Represent :  as $P(\mathbf{h}|\mathbf{v}) = [0,\ 0,\ 0.82,\ 0,\ 0,\ 0.99,\ 0,\ 0\ \ldots\ ]$

Example: inferring topic structure using RBMs



Pair-wise Unary

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K}\sum_{j=1}^{F} W_{ij}^k v_i^k h_j + \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k b_i^k + \sum_{j=1}^{F} h_j a_j\right)$$

$$\theta = \{W, a, b\}$$

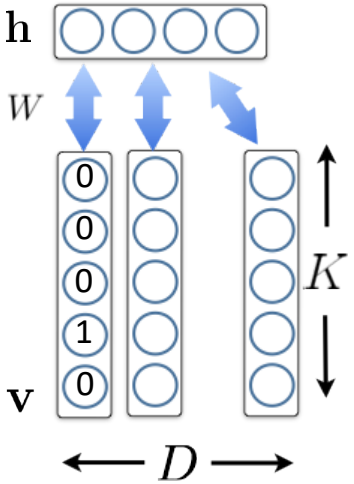
$$P_\theta(v_i^k = 1|\mathbf{h}) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}$$

Replicated Softmax Model: *undirected* topic model:

- Stochastic 1-of-K visible variables.
- Stochastic binary hidden variables **h**
- Bipartite connections.

(Salakhutdinov & Hinton, NIPS 2010, Srivastava & Salakhutdinov, NIPS 2012)

Example: inferring topic structure using RBMs



Pair-wise Unary

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K}\sum_{j=1}^{F} W_{ij}^k v_i^k h_j + \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k b_i^k + \sum_{j=1}^{F} h_j a_j\right)$$

$$\theta = \{W, a, b\}$$

$$P_\theta(v_i^k = 1|\mathbf{h}) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}$$

Reuters dataset:
804,414 **unlabeled** newswire stories
Bag-of-Words

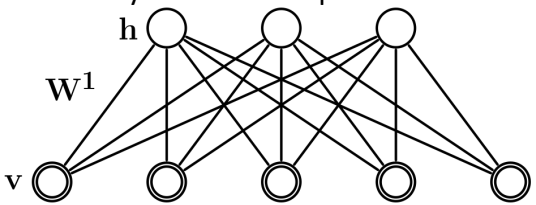
Learned features: ``topics''

russian	clinton	compute	trade	stock
russia	house	r	country	wall
mosco	presiden	system	import	street
w	t	product	world	point
yeltsin	bill	software	econom	dow
soviet	congress	develop	y	

Example: movie predictions using RBMs

$$P_\theta(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{ijk} W_{ij}^k v_i^k h_j + \sum_{ik} b_i^k v_i^k + \sum_j a_j h_j \right)$$

Binary hidden: user preferences



Multinomial visible: user ratings

Netflix dataset:
480,189 users
17,770 movies
Over 100 million ratings

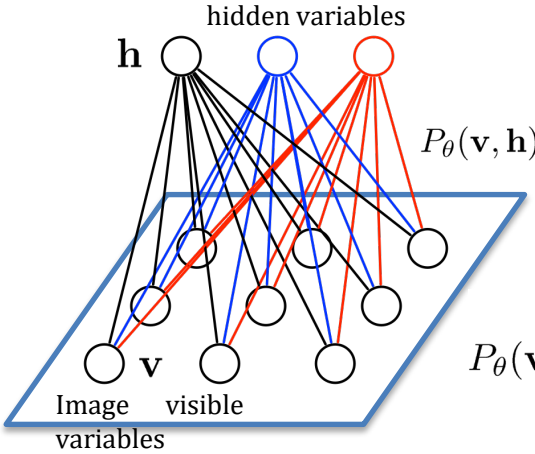
Learned features: ``genre''

Fahrenheit 9/11	Independence Day
Bowling for Columbine	The Day After Tomorrow
The People vs. Larry Flynt	Con Air
Canadian Bacon	Men in Black II
La Dolce Vita	Men in Black
Friday the 13th	Scary Movie
The Texas Chainsaw Massacre	Naked Gun
Children of the Corn	Hot Shots!
Child's Play	American Pie
The Return of Michael Myers	Police Academy

State-of-the-art performance on the Netflix dataset.

(Salakhutdinov, Mnih, Hinton, ICML 2007)

Example: RBMs for image data



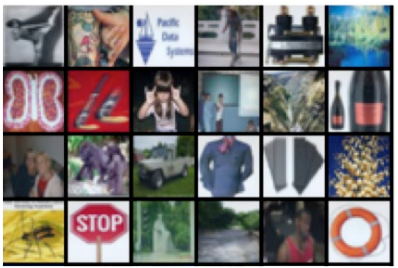
$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} h_j \frac{v_i}{\sigma_i} + \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{F} a_j h_j \right)$$

Pair-wise **Unary**

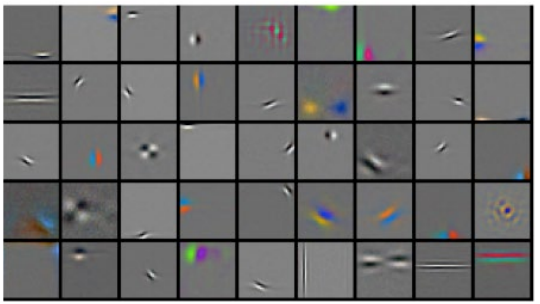
$$\theta = \{W, a, b\}$$

$$P_\theta(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{D} P_\theta(v_i|\mathbf{h}) = \prod_{i=1}^{D} \mathcal{N}\left(b_i + \sum_{j=1}^{F} W_{ij} h_j, \sigma_i^2 \right)$$

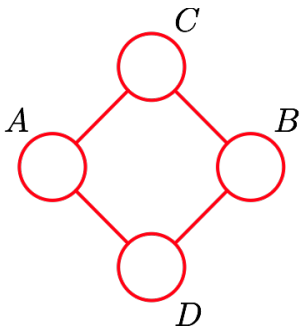
4 million **unlabelled** images



Learned features (out of 10,000)



Conditional independence: a more general principle



More generally, we can describe the family of distributions whose conditional independence structure "tracks" a given (undirected) graph.

The way to formalize this is in terms of **maximal cliques** C (clique = fully connected subset of nodes) of the graph:

$$p(x) \propto \Pi_C \phi_C(x_C)$$

For example, the joint distribution above factorizes as:

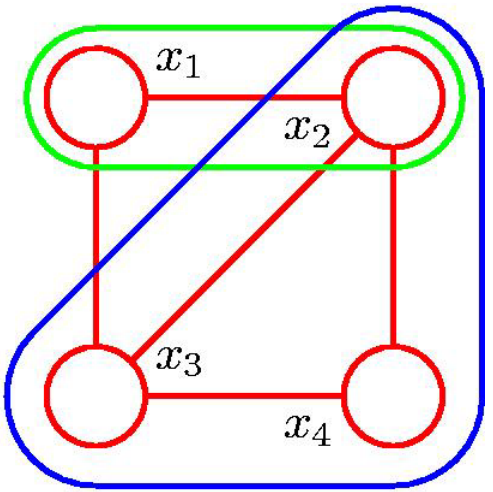
$$p(A,B,C,D) \propto \phi_{AC}(A,C)\phi_{BC}(B,C)\phi_{AC}(B,D)\phi_{AD}(A,D)$$

Maximal cliques

The subsets that are used to define the potential functions are represented by maximal cliques in the undirected graph.

Clique: a subset of nodes such that there exists an edge between all pairs of nodes in a subset.

Maximal Clique: a clique such that it is not possible to include any other nodes in the set without it ceasing to be a clique.



This graph has 5 cliques:

$$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\},$$
$$\{x_4, x_2\}, \{x_1, x_3\}.$$

Two maximal cliques:

$$\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}.$$

How are pairwise graphical models a special case of this?

$$p(x) \propto \exp\left(\sum_{ij} \phi_{ij}(x_i, x_j)\right)$$

Hammersley-Clifford theorem

Consider the following two sets of distributions:

- The set of distributions consistent with the conditional independence relationships defined by separations in an undirected graph G.

- The set of distributions consistent with the factorization defined by potential functions on maximal cliques of the graph G.

Hammersley-Clifford theorem: these two sets of distributions are the same.

Hammersley-Clifford theorem

Part 1 (Easier): Distribution consistent w/ factorization defined by potential fns on maximal cliques of a graph => distribution consistent w/ conditional independence relationships given by separations in an undirected graph.

Hammersley-Clifford theorem

Part 2 (harder): A distribution consistent w/ conditional independence relationships given by separations in an undirected graph => Distribution consistent w/ factorization defined by potential fns on maximal cliques of a graph

(See link in schedule, not covered in class.)

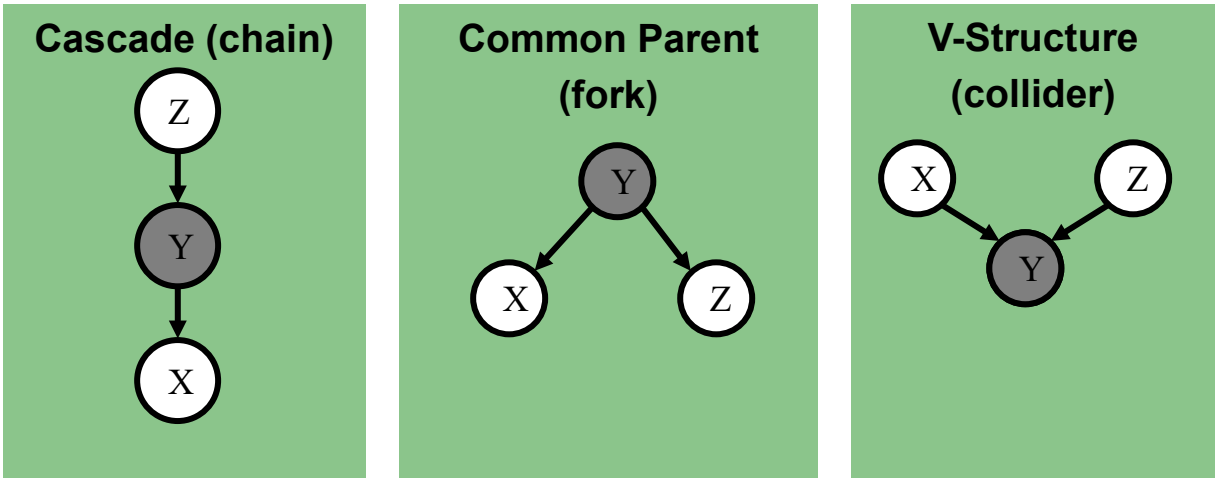
What independencies does a DGM represent?

- By **definition**: Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

- But what other dependencies are encoded?

Some special cases

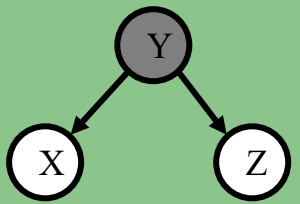


Some special cases

Cascade (chain)

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

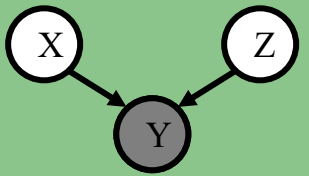
Common Parent (fork)



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

Knowing Y
decouples X and Z

V-Structure (collider)

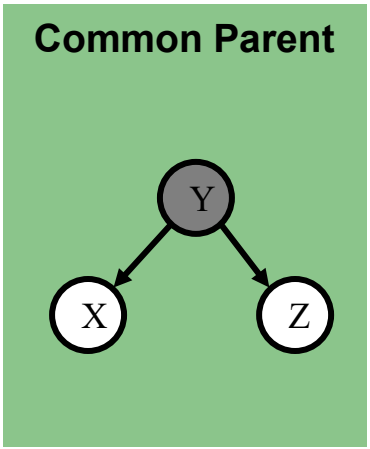


$X \not\perp\!\!\!\perp Z \mid Y$

Knowing Y
couples X and Z

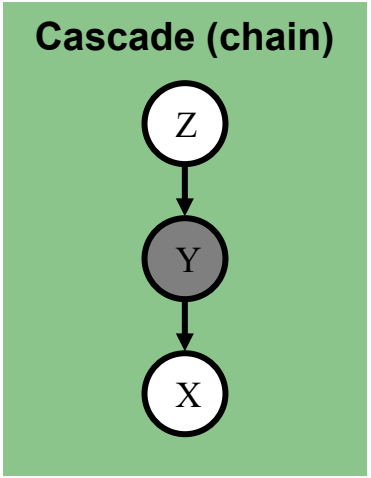
Proof of conditional independence in fork

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



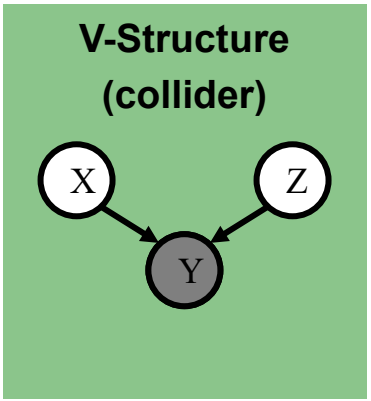
Proof of conditional independence in chain

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



Conditional dependence in collider ("explaining away")

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



General principle: moralization

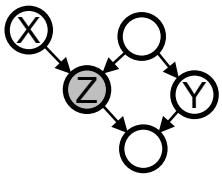
X and Z are **conditionally independent** given the **set** E, if X and Z are separated in the undirected moralized graph.

Definition: $X \perp\!\!\!\perp Z|E$ iff X and Z are separated by E in the **undirected ancestral moral** graph.

1. **Ancestral graph**: keep only X, Z, E and their ancestors
2. **Moral graph**: add undirected edge between all pairs of each node's parents
3. **Undirected graph**: convert all directed edges to undirected
4. Givens Removed: delete any nodes in E

Moralization: example

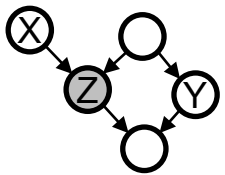
Is X ⫫ Y | Z ?



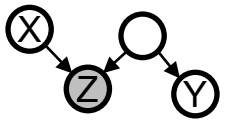
Original graph

Moralization: example

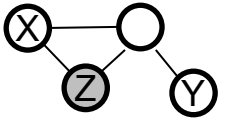
Is X ⫫ Y | Z ?



Original graph



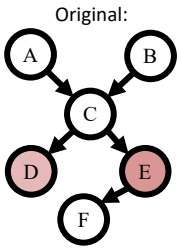
ancestral



Moral ancestral

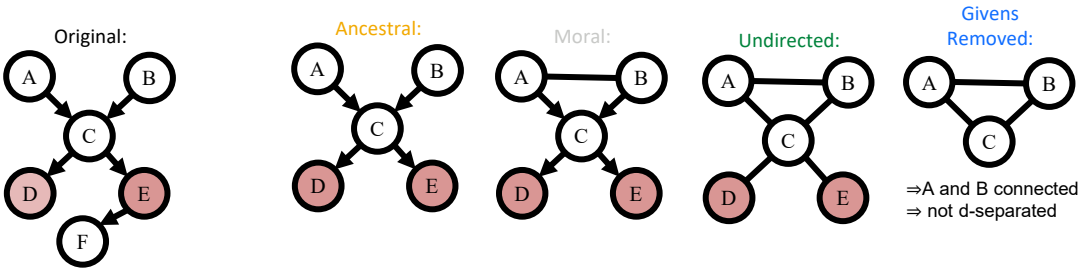
Moralization: example

Is A ⫫ B | {D, E}?



Moralization: example

Is A ⫫ B | {D, E}?



Why does moralization work?

1. **Ancestral graph** suffices : $P(X,Z|E) = \sum_y P(X,Z,Y=y|E)$

   Y which are not parents of X,Z or E can be ignored (X,Z,E are conditionally independent given value of parents).

2. Moral **undirected** graph suffices:

$$P(X,Z,Y|E) = \frac{P(X,Z,Y,E)}{P(E)}$$

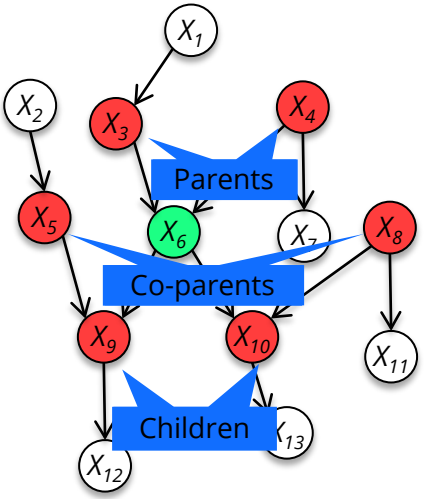
$$\propto \Pi_i p(X_i = x_i | pa(X_i))$$

Can be viewed as factors of an undirected graphical model, which obey maximal clique factorization of moralized graph.

Conditional independence follows by Hammersley-Clifford

Markov blankets in DGMs

The **Markov Blanket** of a node in a directed graphical model is the set containing the node's parents, children, and co-parents.



Another principle: D-separation

If variables X and Z are **d-separated** given a **set** of variables E

Then X and Z are **conditionally independent** given the **set** E

Definition: Variables X and Z are **d-separated** given a **set** of evidence vars E iff every (undirected) path from X to Z is blocked.

A path is blocked whenever:

1. ∃Y on path s.t. Y ∈ E and Y is a "common parent"

2. ∃Y on path s.t. Y ∈ E and Y is in a "cascade"

3. ∃Y on path s.t. {Y, descendants(Y)} ∉ E and Y is in a "v-structure"