Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 2:
Directed Graphical Models

Some figures courtesy Michael Jordan's draft textbook,
An Introduction to Probabilistic Graphical Models

Discrete Random Variables

$X \longrightarrow$ discrete random variable

$\mathcal{X} \longrightarrow$ sample space of possible outcomes, which may be finite or countably infinite

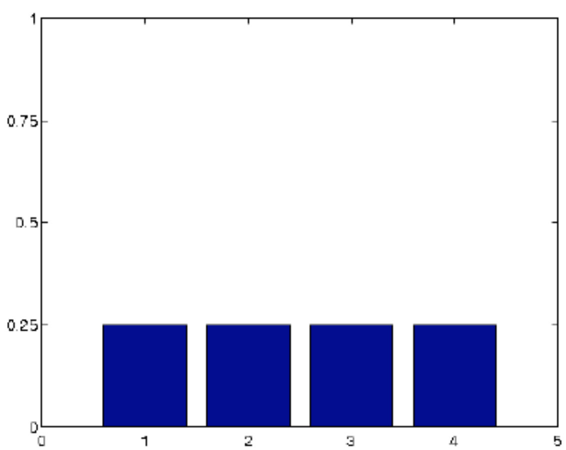
$x \in \mathcal{X} \longrightarrow$ outcome of sample of discrete random variable

$p(X = x) \longrightarrow$ probability distribution (probability mass function)

$p(x) \longrightarrow$ shorthand used when no ambiguity

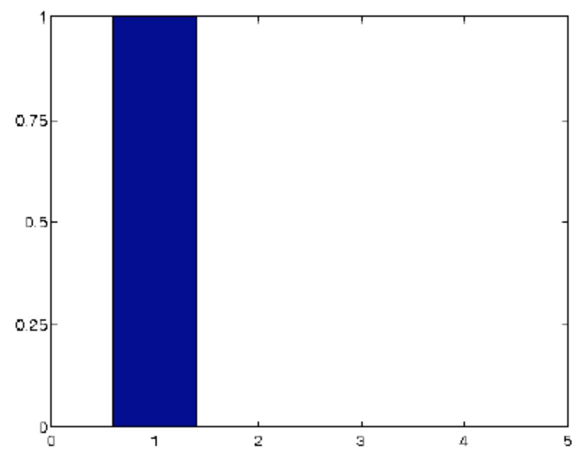
$$0 \leq p(x) \leq 1 \text{ for all } x \in \mathcal{X}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$



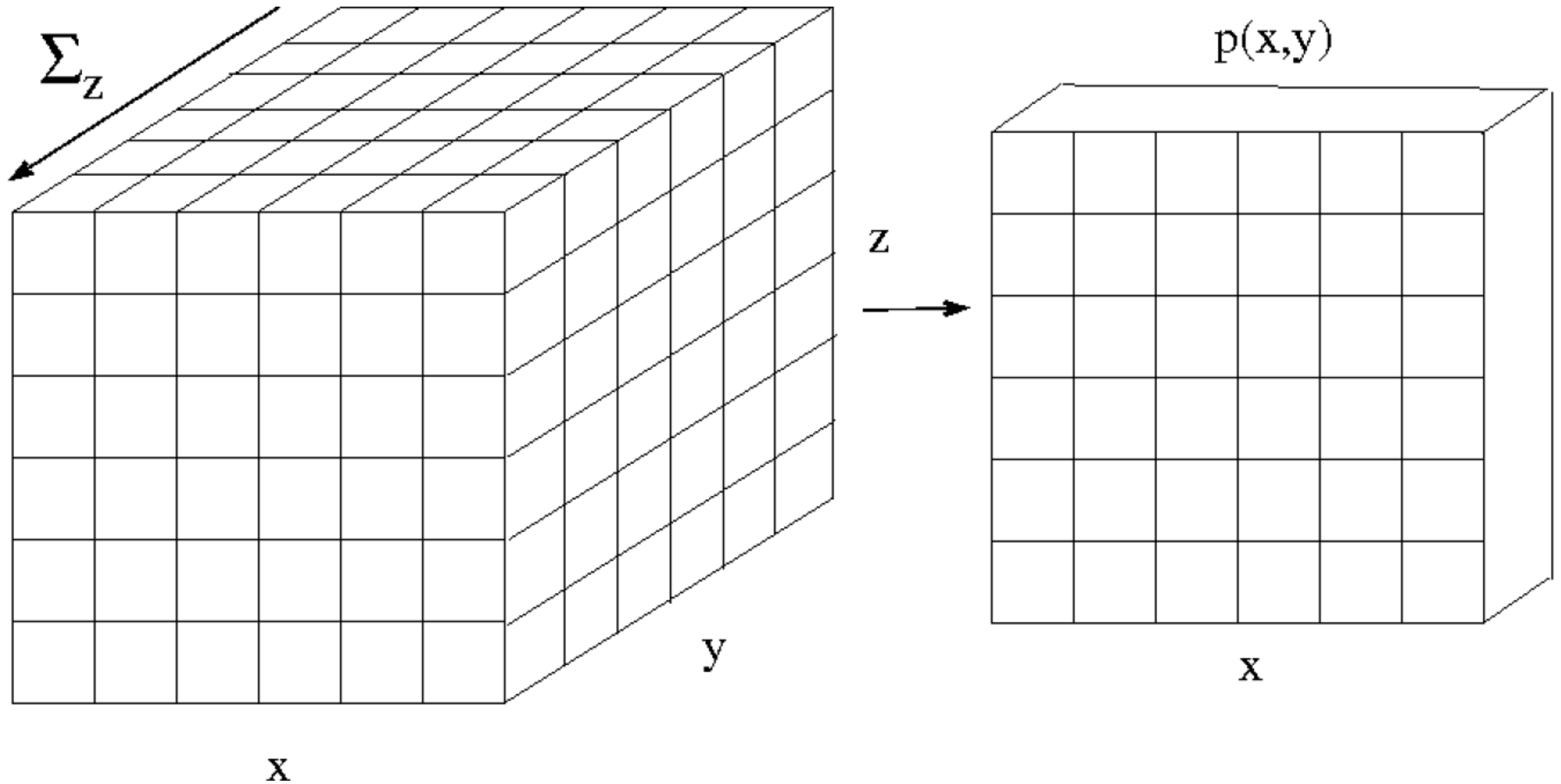
uniform distribution

$$\mathcal{X} = \{1, 2, 3, 4\}$$



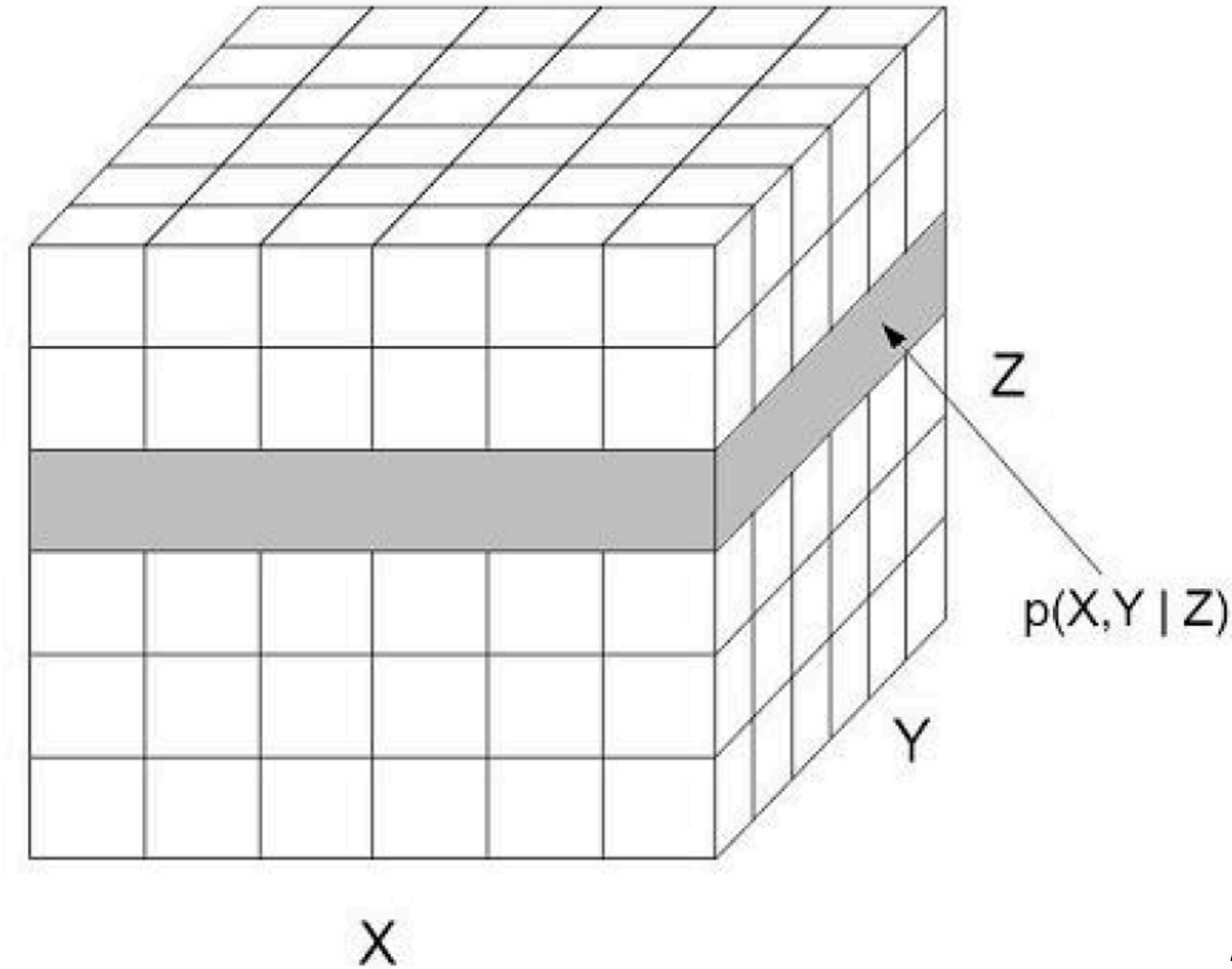
degenerate distribution

Marginal Distributions



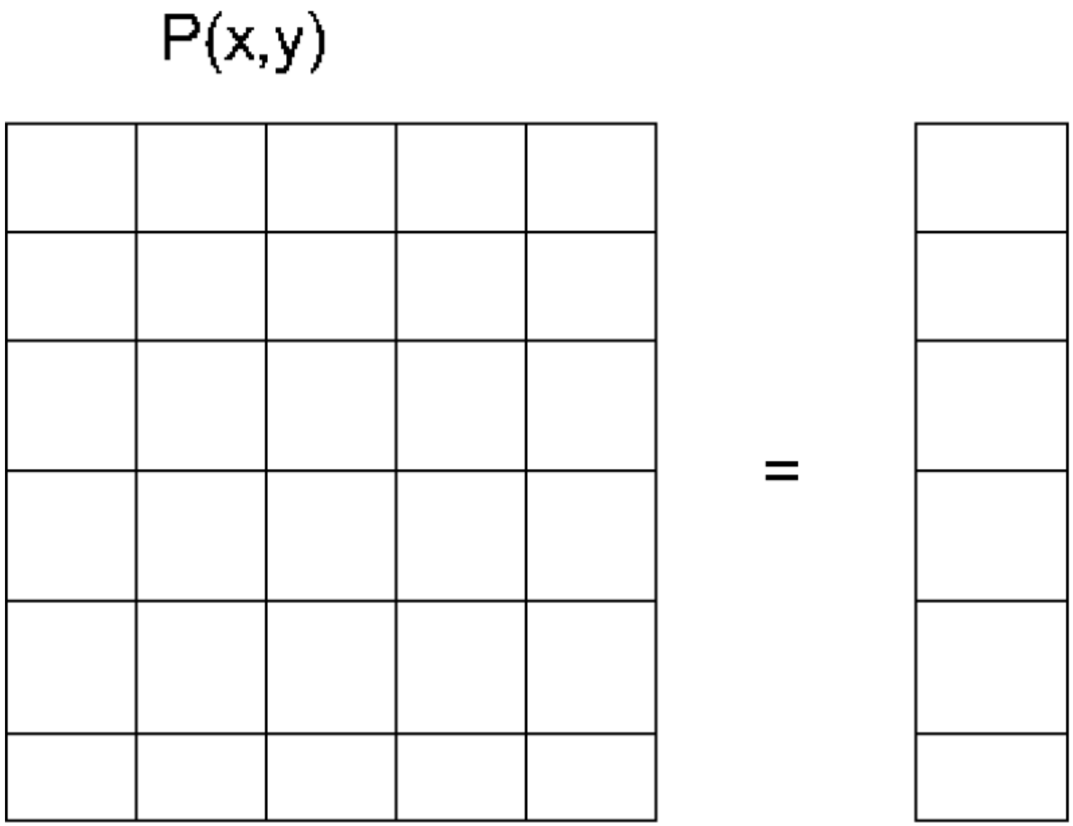
$$p(x, y) = \sum_{z \in \mathcal{Z}} p(x, y, z) \qquad p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

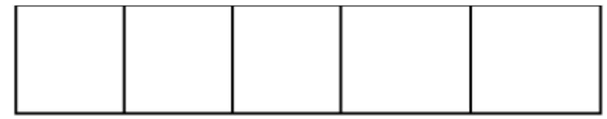
Conditional Distributions



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

Independent Random Variables





$$X \perp Y$$

$$\Updownarrow$$

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

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

Equivalent conditions on conditional probabilities:

$$p(x \mid Y = y) = p(x) \text{ and } p(y) > 0 \text{ for all } y \in \mathcal{Y}$$

$$p(y \mid X = x) = p(y) \text{ and } p(x) > 0 \text{ for all } x \in \mathcal{X}$$

Bayes Rule (Bayes Theorem)

$$p(x, y) = p(x)p(y \mid x) = p(y)p(x \mid y)$$

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(y \mid x)p(x)}{\sum_{x' \in \mathcal{X}} p(x')p(y \mid x')}$$

$$\propto p(y \mid x)p(x)$$

- A basic identity from the definition of conditional probability
- Used in ways that have nothing to do with Bayesian statistics!
- Typical application to learning and data analysis:

$X \longrightarrow$ unknown parameters we would like to infer

$Y = y \longrightarrow$ observed data available for learning

$p(x) \longrightarrow$ prior distribution (domain knowledge)

$p(y \mid x) \longrightarrow$ likelihood function (measurement model)

$p(x \mid y) \longrightarrow$ posterior distribution (learned information)

Binary Random Variables

Bernoulli Distribution: Single toss of a (possibly biased) coin

$$\mathcal{X} = \{0, 1\}$$

$$0 \leq \theta \leq 1$$

$$\mathrm{Ber}(x \mid \theta) = \theta^{\delta(x,1)}(1-\theta)^{\delta(x,0)}$$

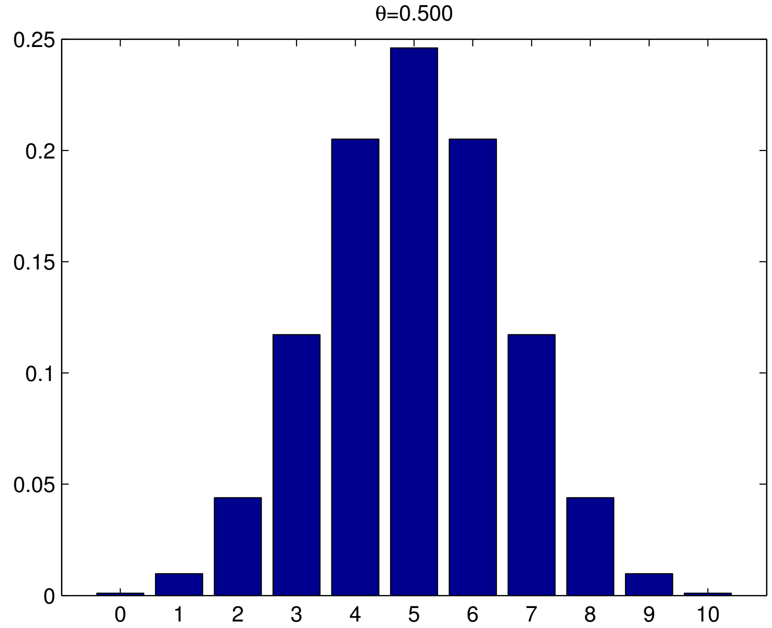
Binomial Distribution: Toss a single (possibly biased) coin *n* times, and record the number *k* of times it comes up heads

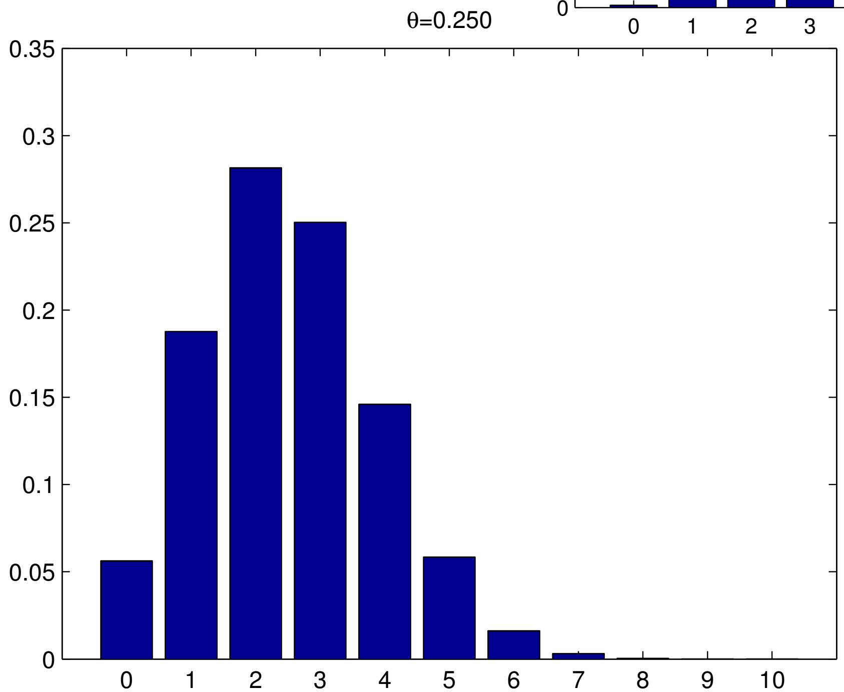
$$\mathcal{K} = \{0, 1, 2, \ldots, n\}$$

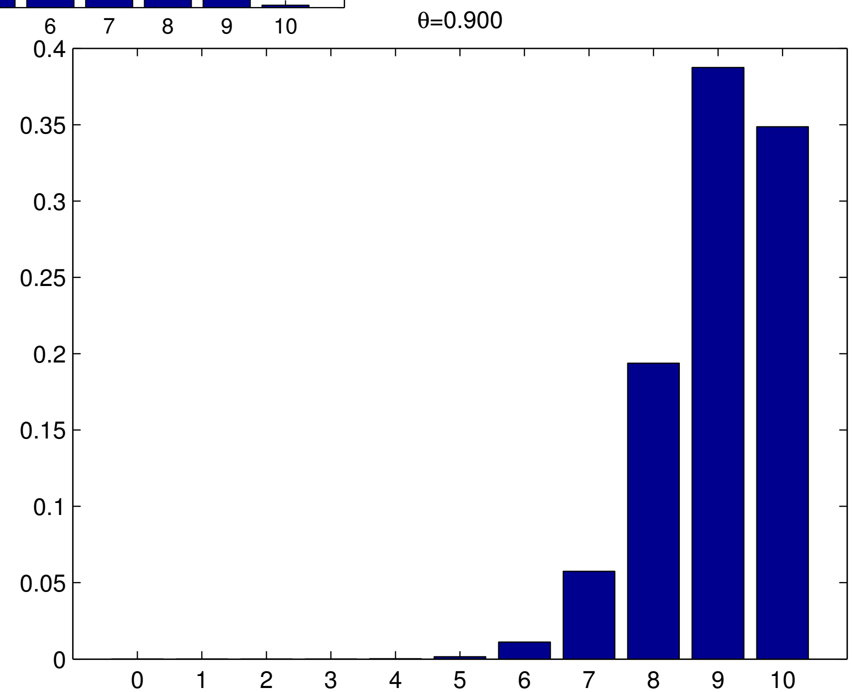
$$0 \leq \theta \leq 1$$

$$\mathrm{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \qquad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Binomial Distributions







Categorical Random Variables

Multinoulli Distribution: Single roll of a (possibly biased) die

$$\mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$

binary vector encoding

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K), \theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1$$

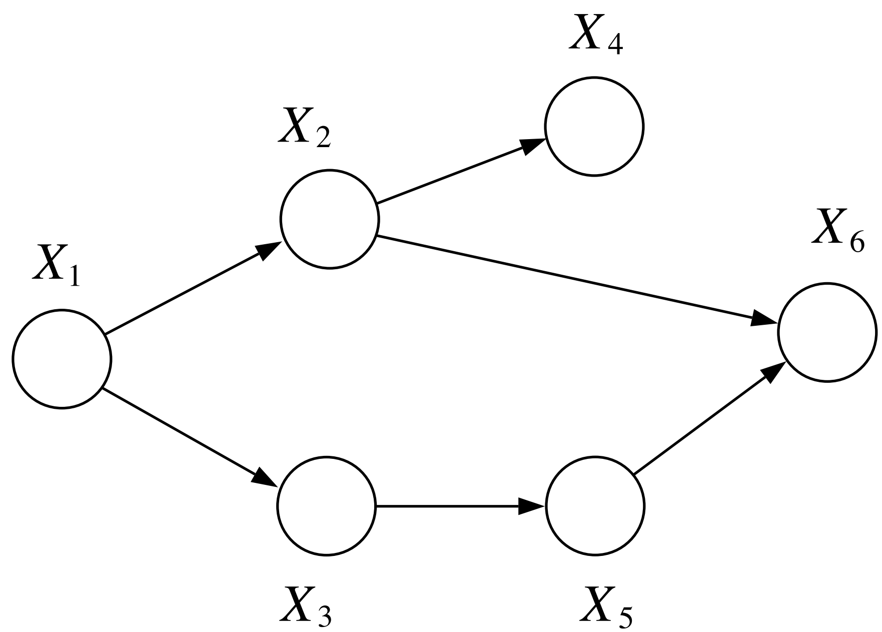
$$\mathrm{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k}$$

Multinomial Distribution: Roll a single (possibly biased) die *n* times, and record the number n_k of each possible outcome

$$\mathrm{Mu}(x \mid n, \theta) = \binom{n}{n_1 \ldots n_K} \prod_{k=1}^{K} \theta_k^{n_k}$$

$$n_k = \sum_{i=1}^{n} x_{ik}$$

Directed Acyclic Graphs (DAGs)



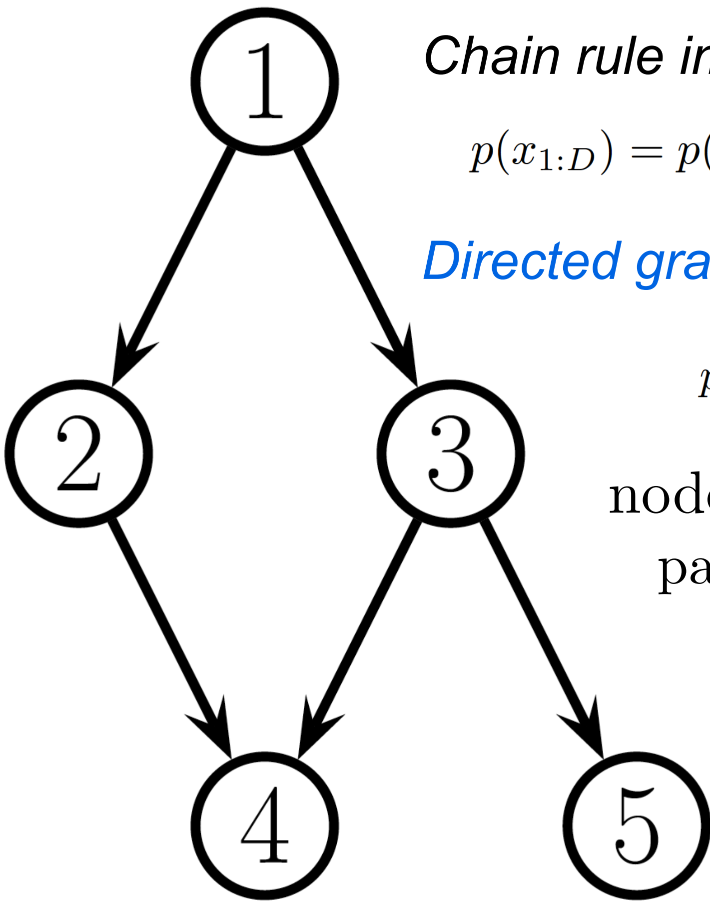
$\mathcal{V} \longrightarrow$ set of N nodes or vertices, $\{1, 2, \ldots, N\}$

$\mathcal{E} \longrightarrow$ set of oriented edges (s,t) linking parents s to children t, so that the set of parents of a node is

$$\mathrm{pa}(t) = \Gamma(t) = \{s \in \mathcal{V} \mid (s,t) \in \mathcal{E}\}$$

$X_s = x_s \longrightarrow$ random variable associated with node s

Directed Graphical Models



Chain rule implies that any joint distribution equals:

$$p(x_{1:D}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_1,x_2,x_3)\dots p(x_D|x_{1:D-1})$$

Directed graphical model implies a restricted factorization:

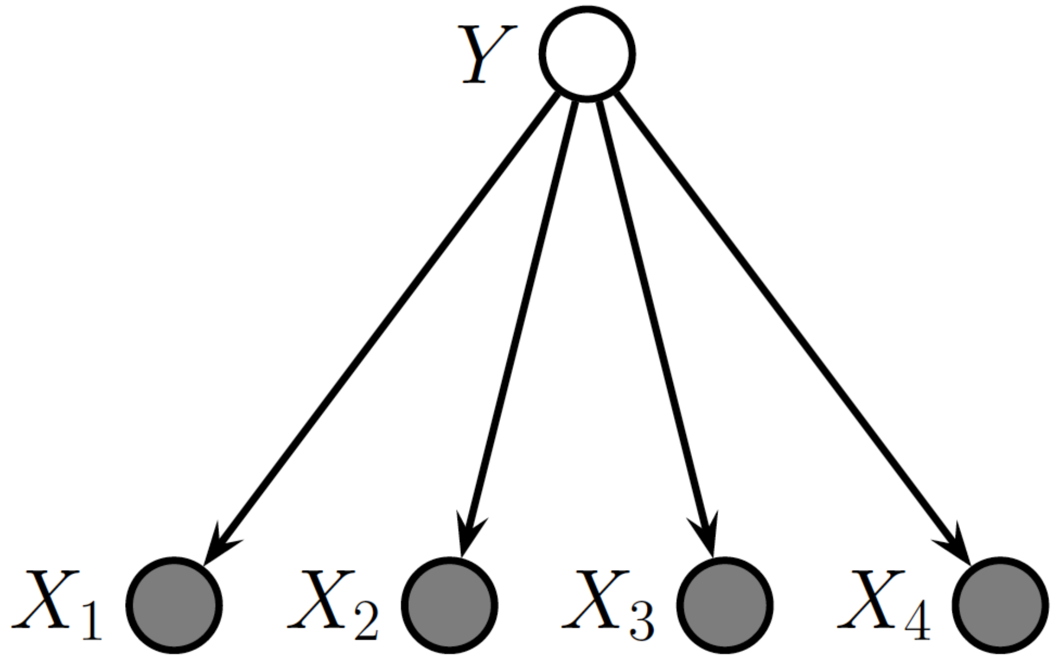
$$p(\mathbf{x}_{1:D}|G) = \prod_{t=1}^{D} p(x_t|\mathbf{x}_{\mathrm{pa}(t)})$$

$$\begin{aligned} \text{nodes} &\rightarrow \text{random variables} \\ \mathrm{pa}(t) &\rightarrow \text{parents with edges pointing to node } t \end{aligned}$$

Valid for any directed acyclic graph (DAG): equivalent to dropping conditional dependencies in standard chain rule

$$\begin{aligned} p(\mathbf{x}_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1,\cancel{x_2})p(x_4|\cancel{x_1},x_2,x_3)p(x_5|\cancel{x_1},\cancel{x_2},x_3,\cancel{x_4}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2,x_3)p(x_5|x_3) \end{aligned}$$

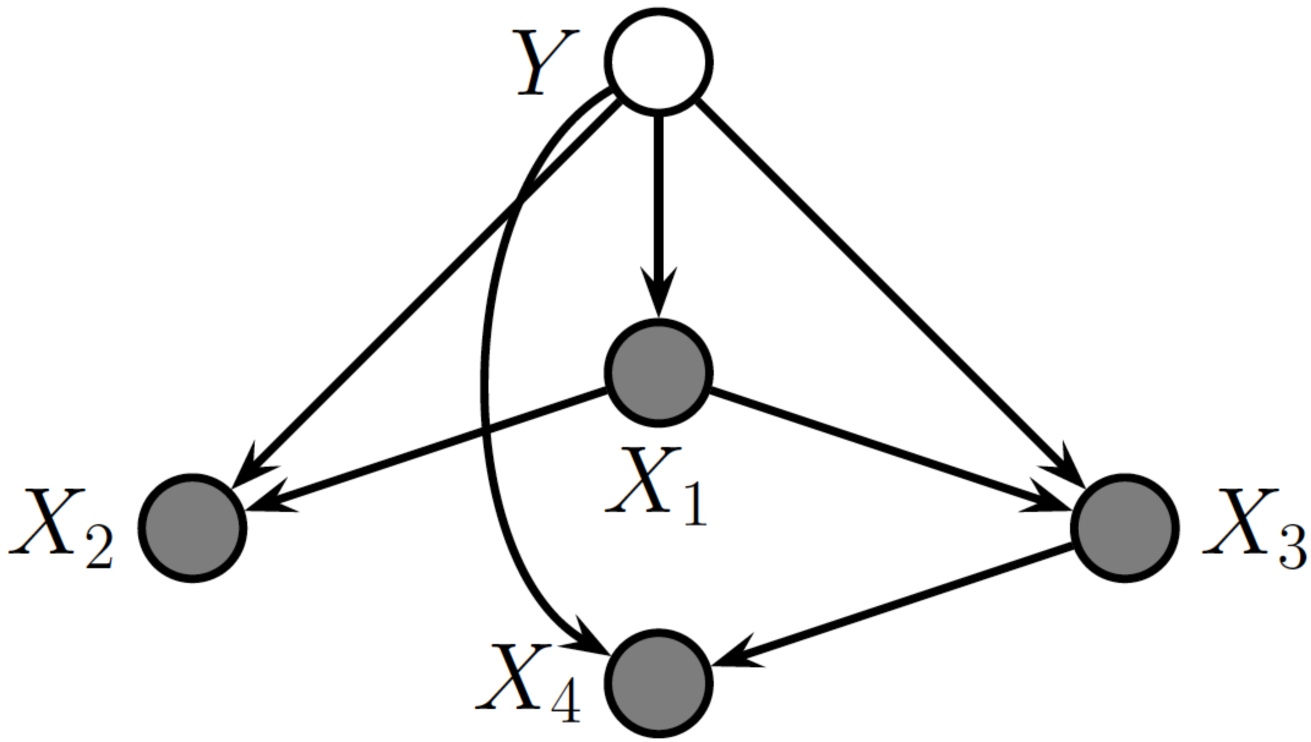
Name That Model



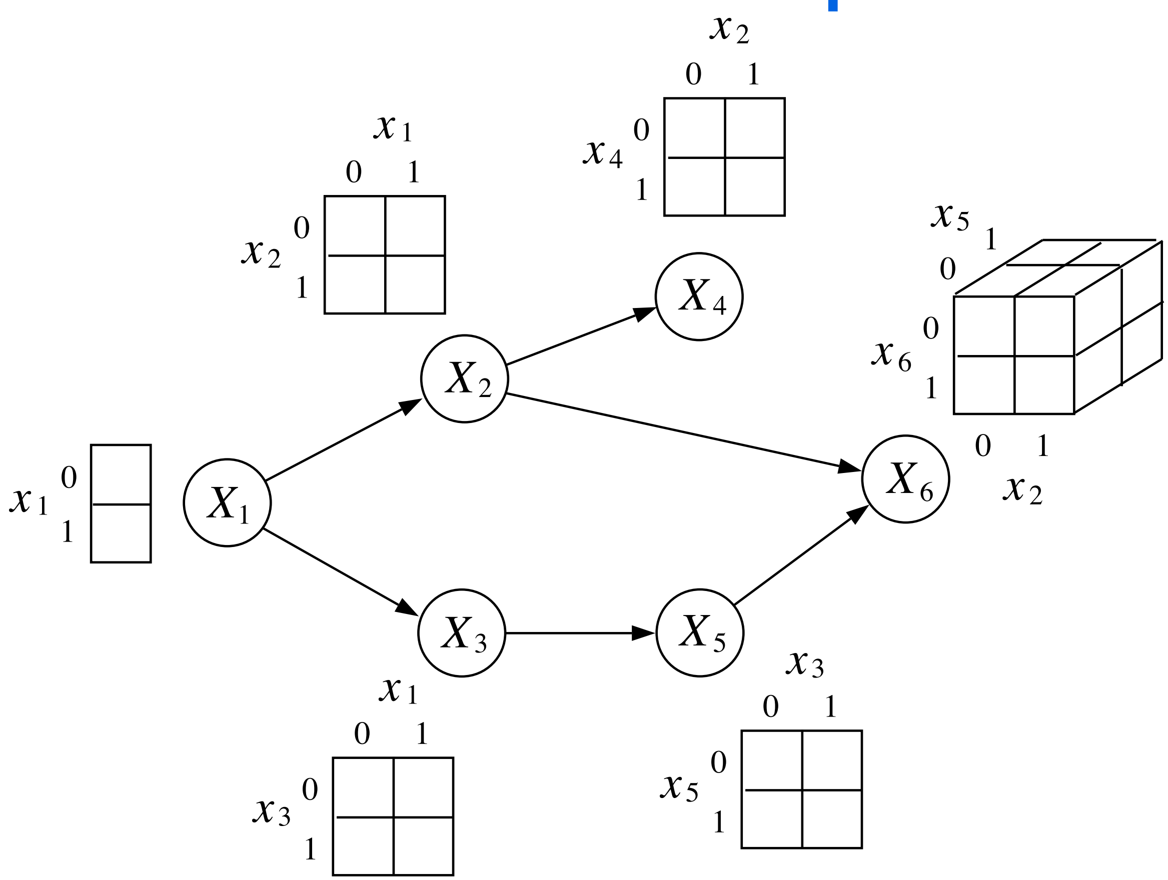
Naïve Bayes:

$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$

Tree-Augmented Naïve Bayes



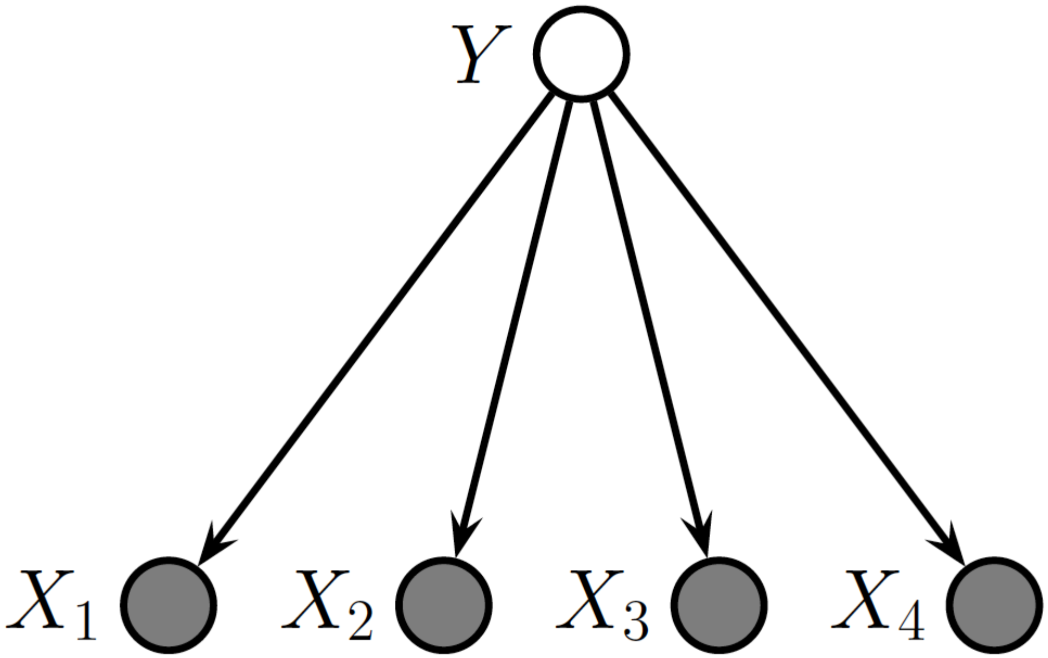
Parameterization & Representation

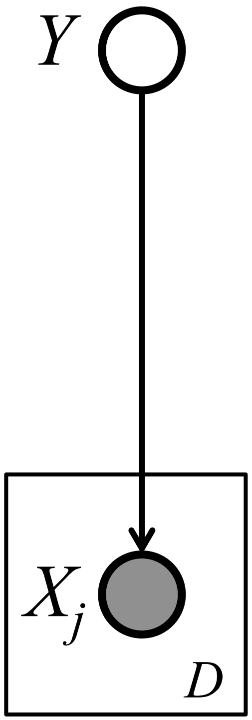


Representational (storage, learning, computation) Complexity

- *Joint distribution:* Exponential in number of variables
- *Directed graphical model:* Exponential in number of parents ("fan-in") of each node, linear in number of nodes

Shading & Plate Notation



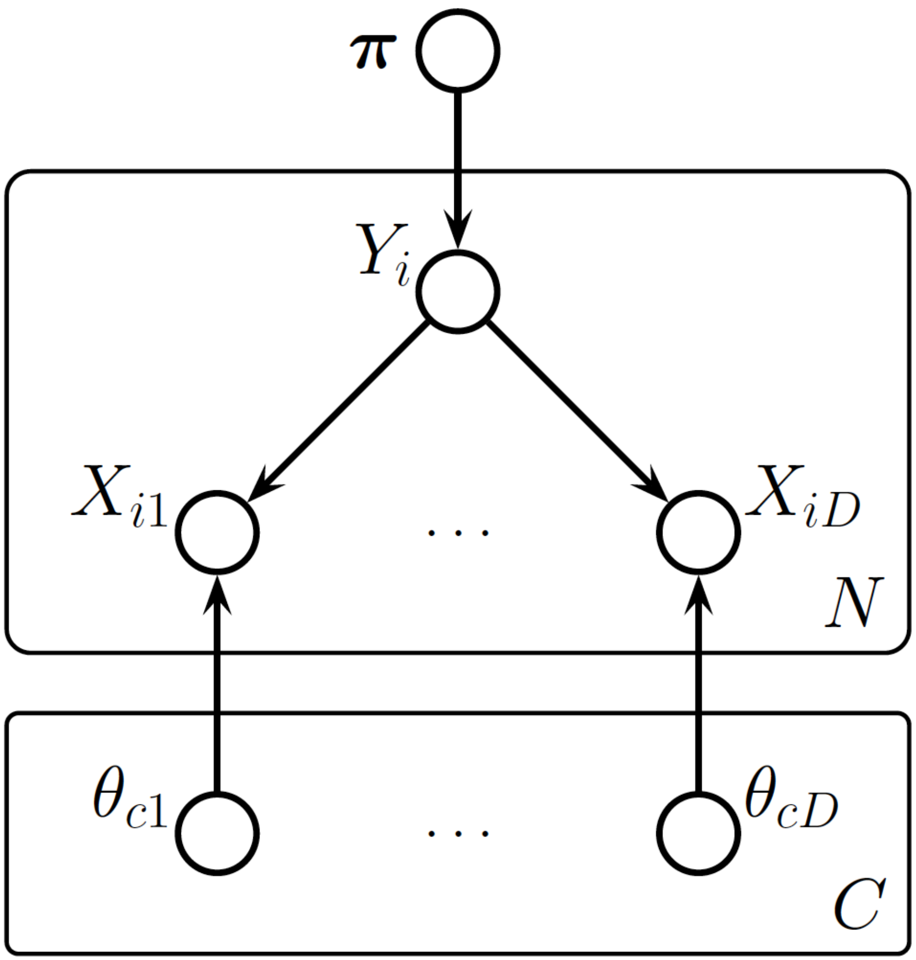


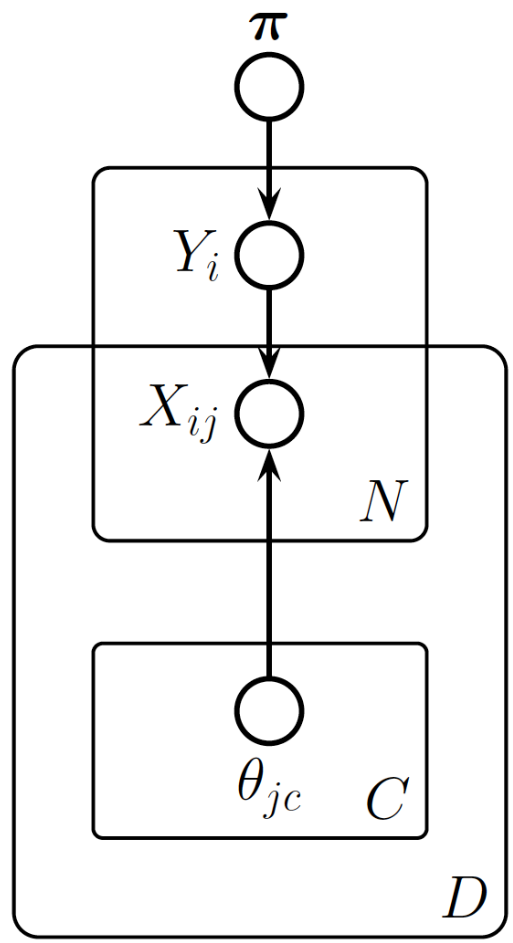
Plates denote replication of random variables

Naïve Bayes Inference: $$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$

Convention: Shaded nodes are observed, open nodes are latent/hidden

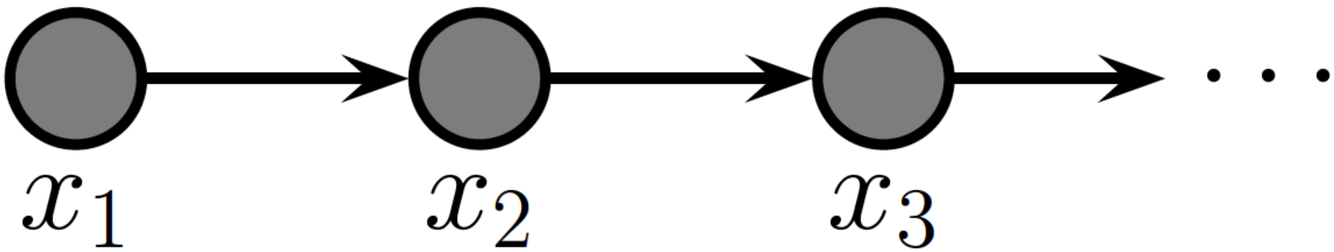
Learning and Unknown Parameters





$$p(\pi)\left[\prod_{c=1}^{C}\prod_{j=1}^{D} p(\theta_{cj})\right]\prod_{i=1}^{N}\left[p(y_i \mid \pi)\prod_{j=1}^{D} p(x_{ij} \mid y_i, \theta_{j1}, \ldots, \theta_{jC})\right]$$

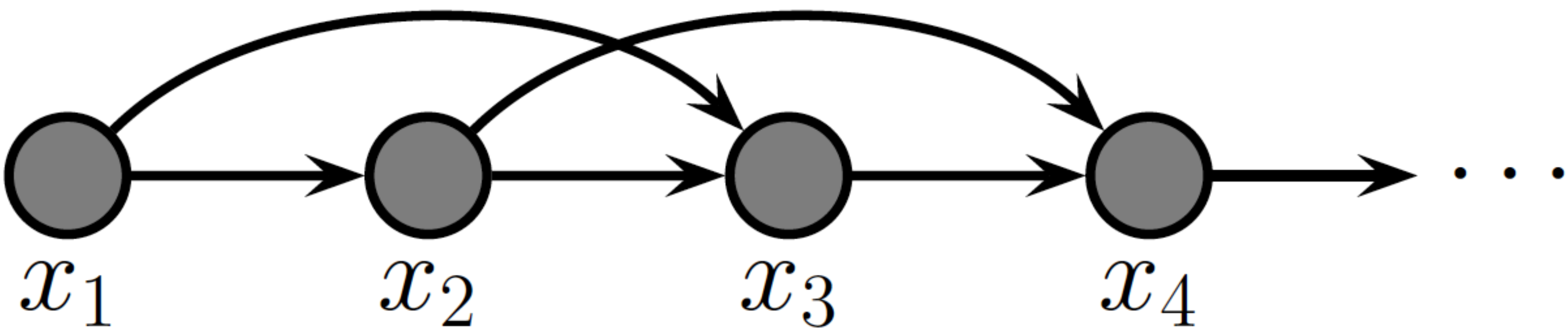
Example: Markov Chains



$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3) \cdots$$

Markov Property

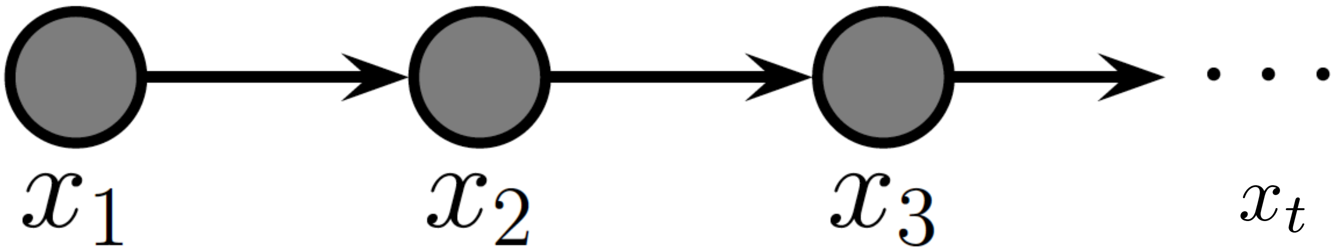
Conditioned on the present, the past and future are independent



$$p(\mathbf{x}_{1:T}) = p(x_1, x_2)p(x_3|x_1, x_2)p(x_4|x_2, x_3) \ldots = p(x_1, x_2) \prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$$

Graphical Models vs. State Diagrams

Graphical Model: *One node per time point*



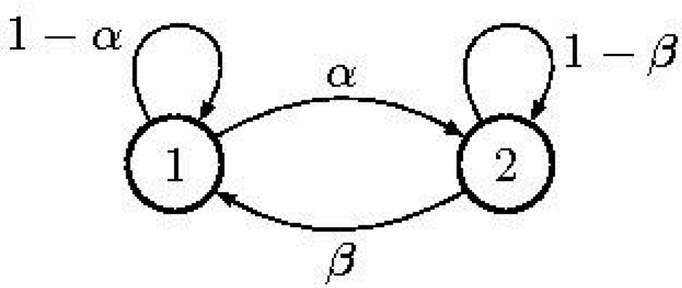
$x_t \in \{1, 2, \ldots, K\}$

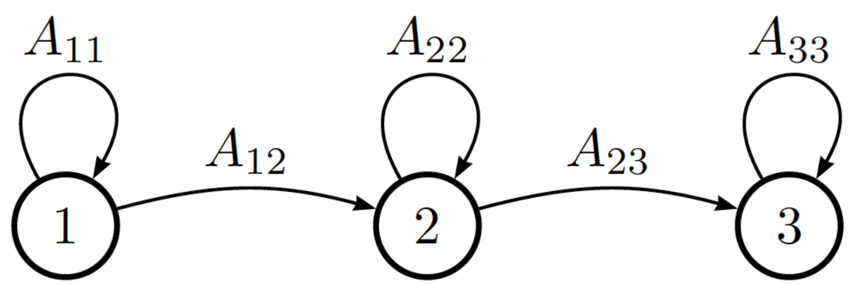
$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)\cdots$$

Interesting when Markov chain is part of a more complex model.

State Transition Matrix: $A \in \mathbb{R}^{K \times K}, A_{ij} = p(x_t = j \mid x_{t-1} = i)$

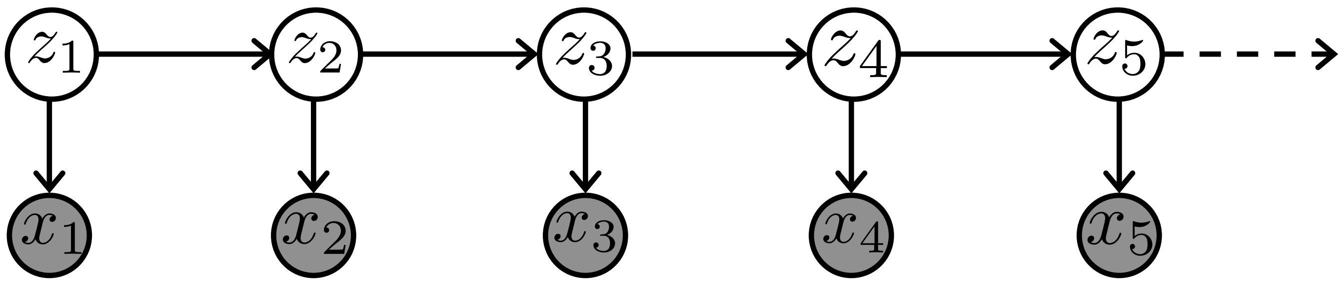
State Transition Diagram: *One node per discrete state*





Not a graphical model! Interesting when state transition matrix is sparse.

Hidden Markov Models (HMMs)



$$p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) \quad = \quad p(\mathbf{z}_{1:T})p(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) = \left[p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \right] \left[\prod_{t=1}^{T} p(\mathbf{x}_t|z_t) \right]$$

$z_t \longrightarrow$ Hidden states taking one of *K* discrete values

$x_t \longrightarrow$ Observations taking values in any space

Discrete: M observation symbols $\rightarrow B \in \mathbb{R}^{K \times M}$

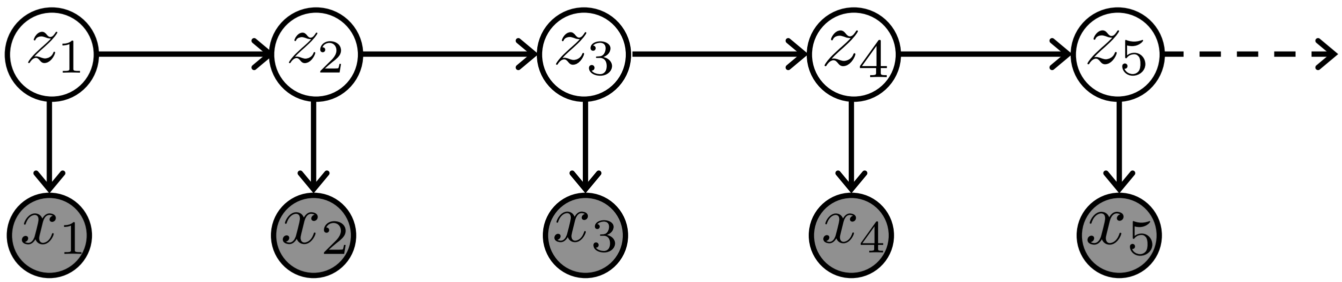
$$p(x_t = \ell \mid z_t = k) = B_{k\ell}$$

Continuous Gaussian:

$$p(x_t \mid z_t = k) = \mathcal{N}(x_t \mid \mu_k, \Sigma_k)$$

Or any convenient family, e.g. an exponential family…

Examples: Sequence Labeling in NLP



Part of speech (POS) tagging:

$\mathbf{z}$:	DT	JJ	NN	VBD	NNP	.
$\boldsymbol{x}$:	the	big	cat	bit	Sam	.

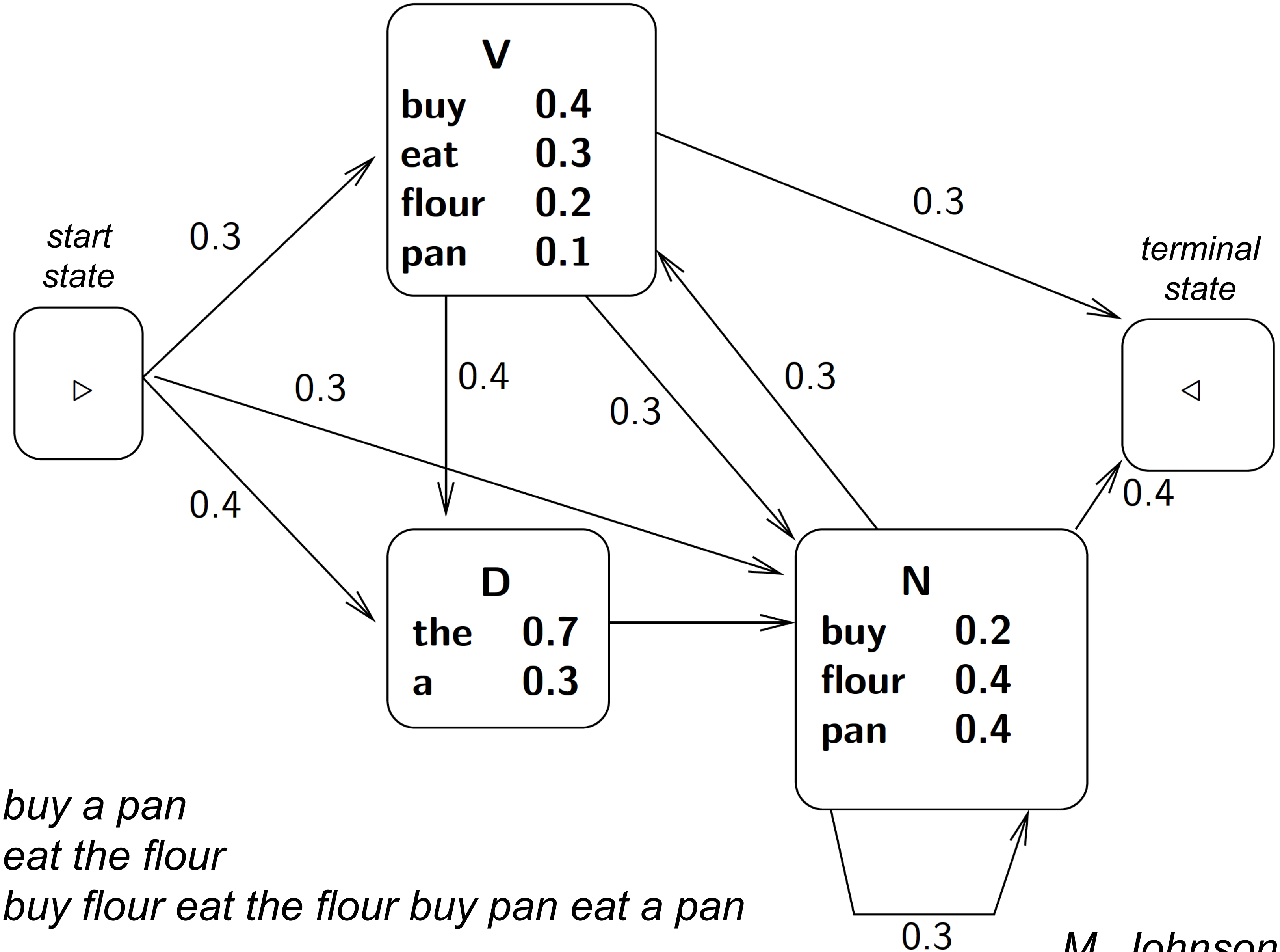
Named entity detection:

$\mathbf{z}$:	[CO	CO]	_	[LOC]	_	[PER]	_
$\boldsymbol{x}$:	XYZ	Corp.	of	Boston	announced	Spade's	resignation

Speech recognition: The $\boldsymbol{x}$ are 100 msec. time slices of acoustic input, and the $\mathbf{z}$ are the corresponding phonemes (i.e., $\mathbf{z}_i$ is the phoneme being uttered in time slice x_i)

M. Johnson, 2009

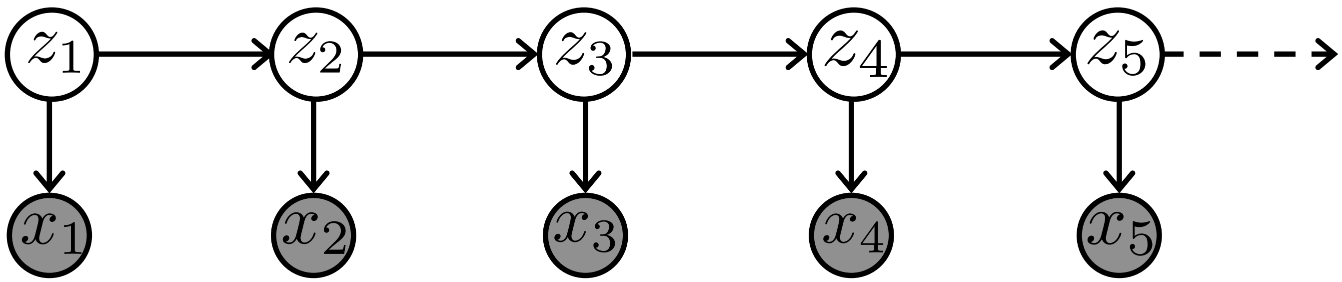
Example: Discrete Language HMM

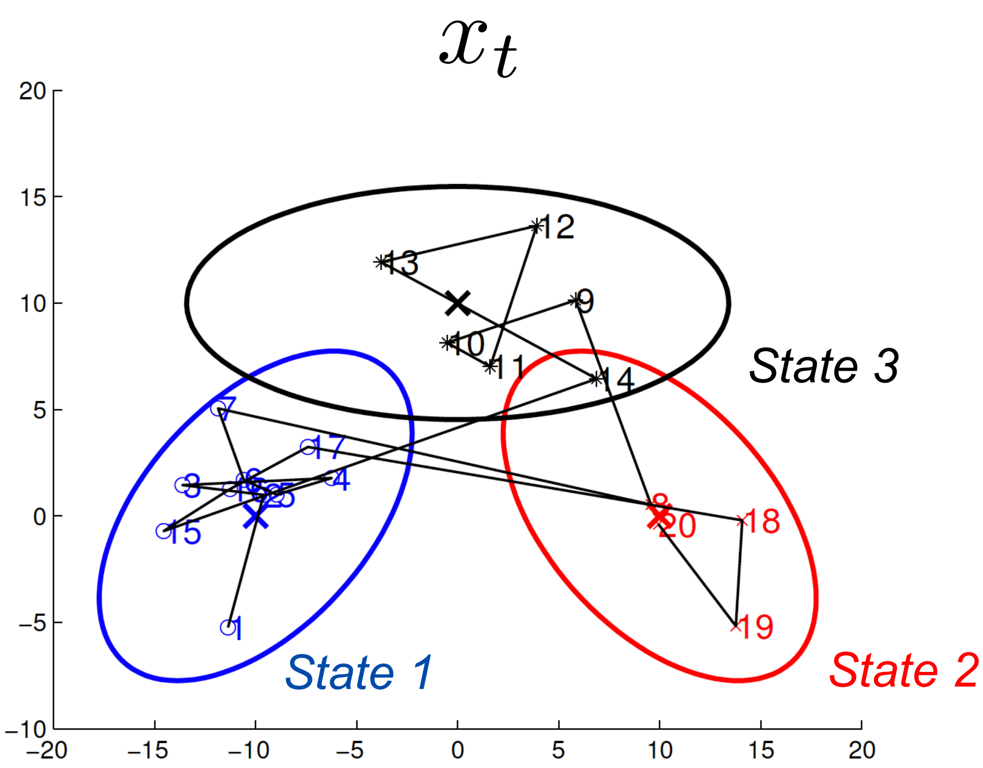


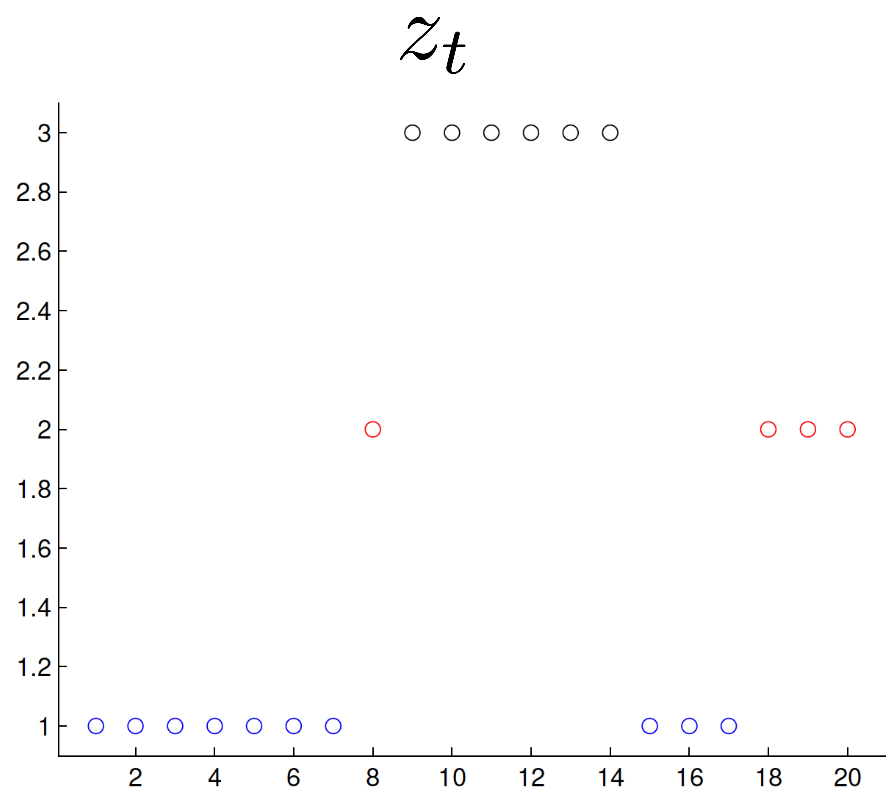
buy a pan
eat the flour
buy flour eat the flour buy pan eat a pan

M. Johnson, 2009

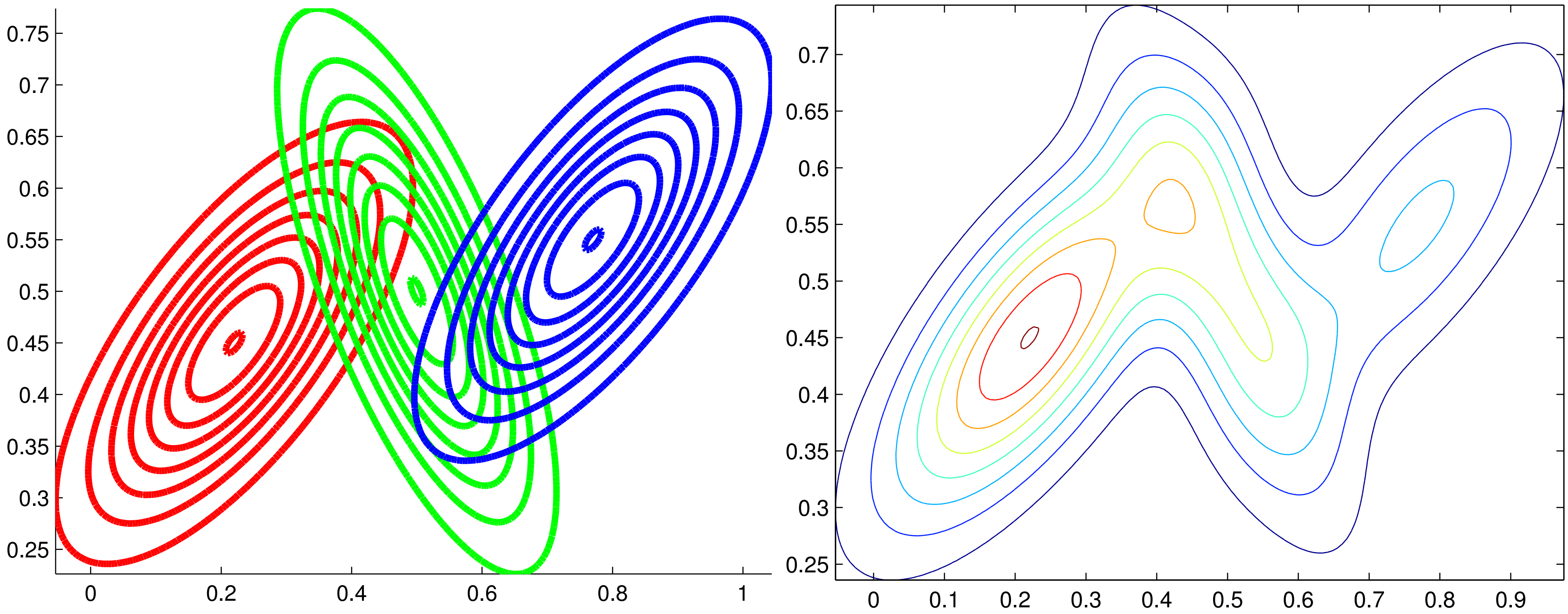
Example: 3-State Gaussian HMM







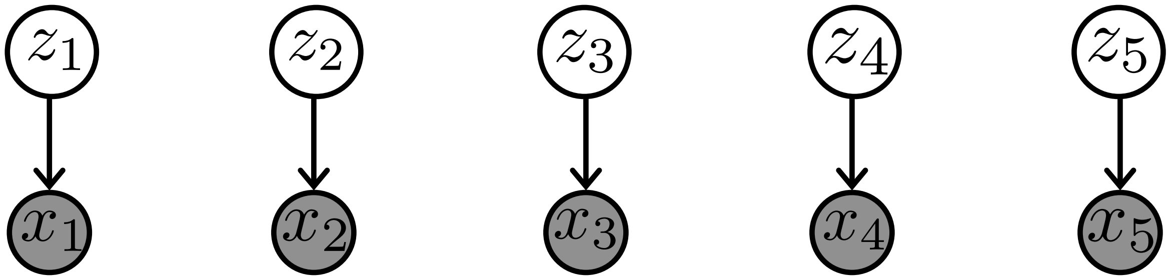
Gaussian Mixture Models



Mixture models are a special case of HMMs, in which the state transition distribution happens to not depend on the previous state, and becomes the mixture prior probability.

Gaussian Mixture Models vs. HMMs

Mixture Model

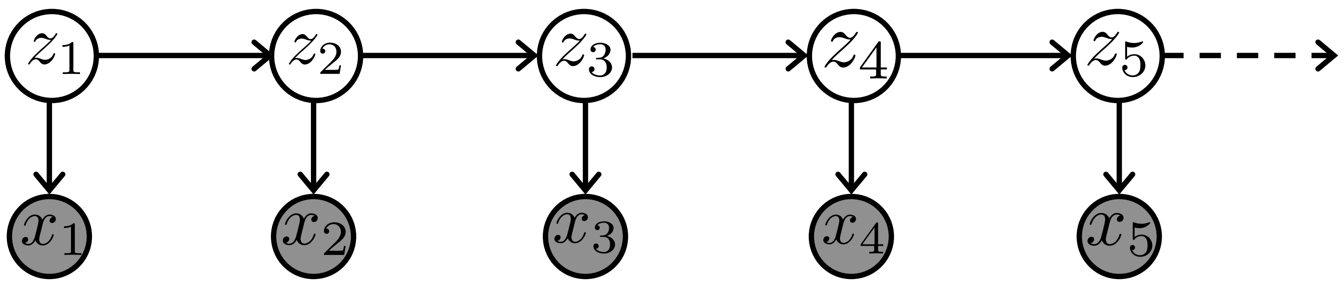


$z_i \in \{1, \ldots, K\}$

$$p(z_i \mid \pi, \mu, \Sigma) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathrm{Norm}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Hidden Markov Model

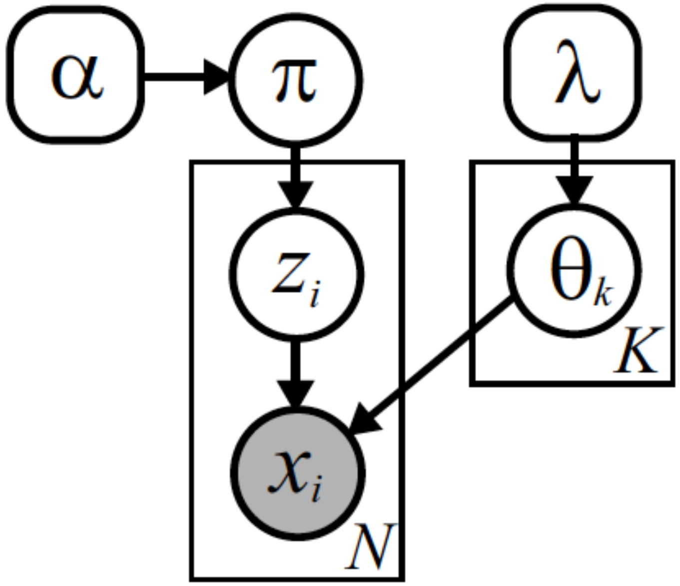


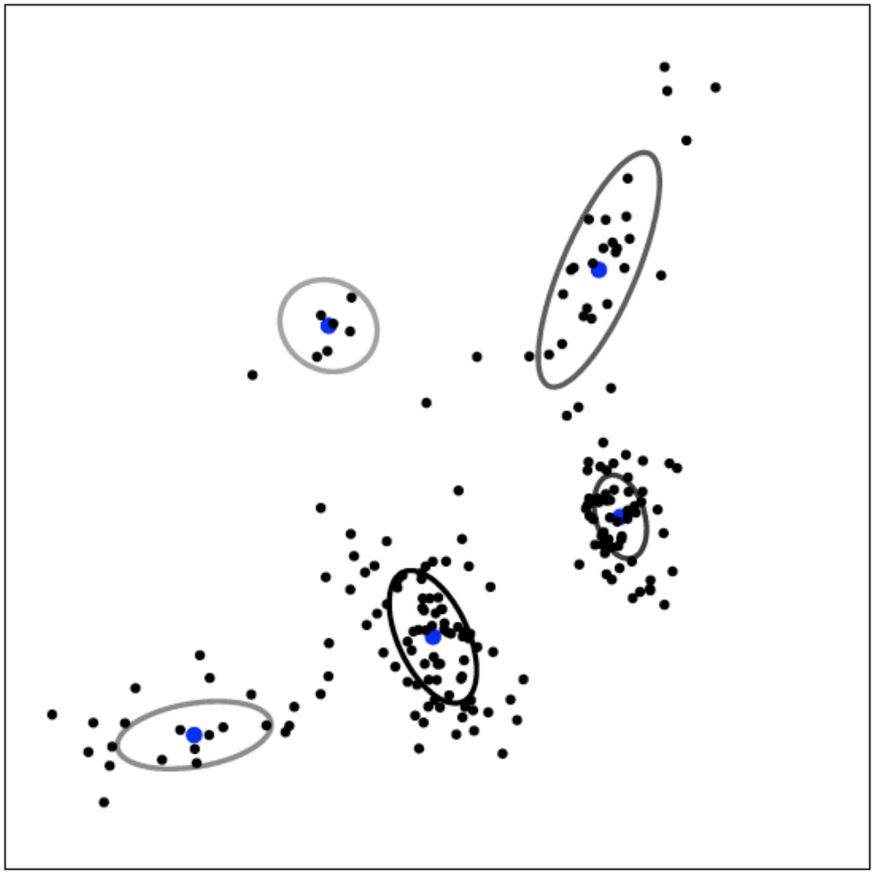
$$p(z_t \mid \pi, \mu, \Sigma, z_{t-1}, z_{t-2}, \ldots) = \mathrm{Cat}(z_t \mid \pi_{z_{t-1}})$$

$$p(x_t \mid z_t, \pi, \mu, \Sigma) = \mathrm{Norm}(x_t \mid \mu_{z_t}, \Sigma_{z_t})$$

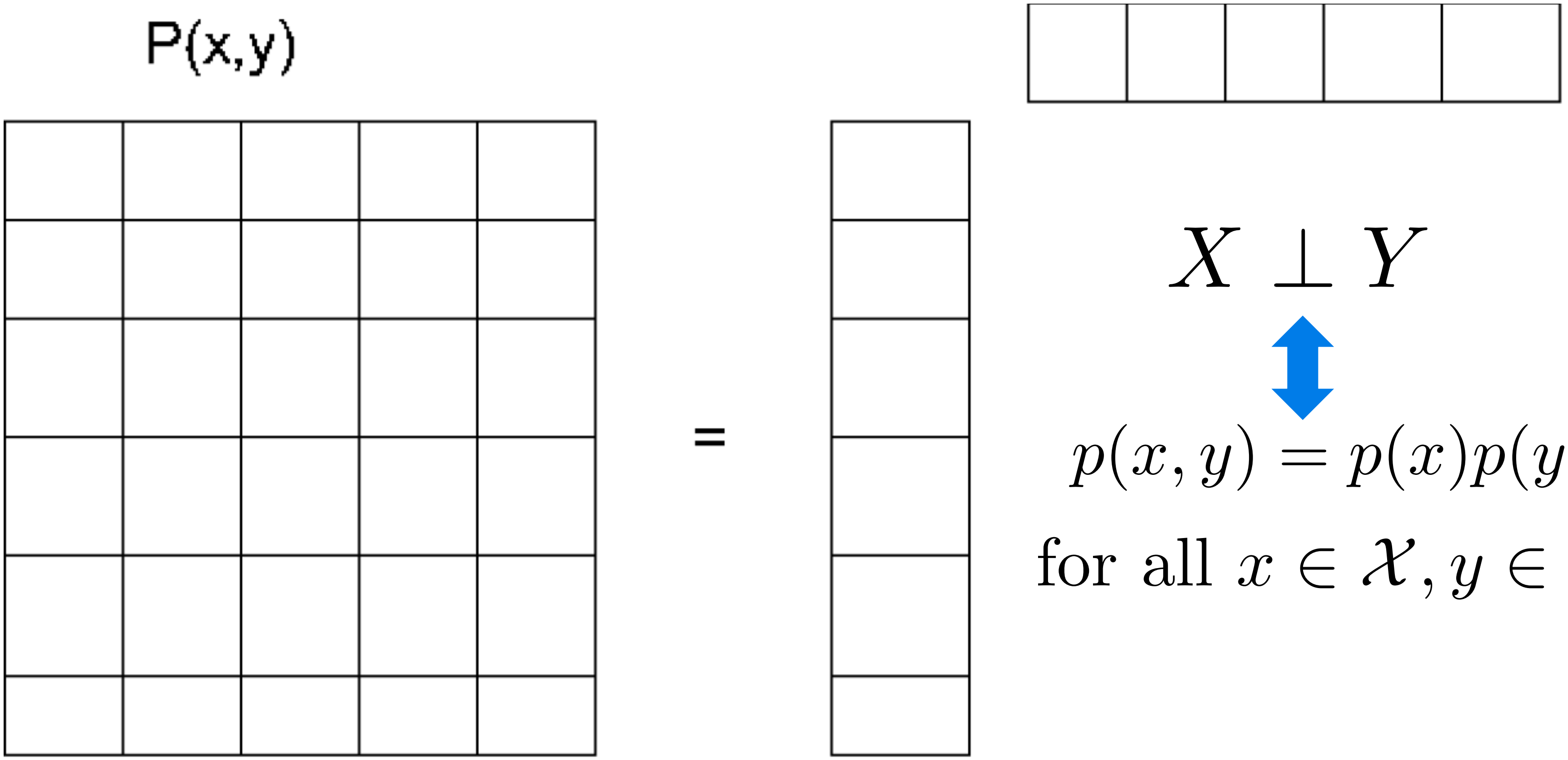
Recover mixture model when all rows of state transition matrix are equal.

Learning Mixture Models





Graphs and Independence



$$X \perp Y$$

$$\Updownarrow$$

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

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

$$x \rightarrow y \qquad\qquad x \leftarrow y \qquad\qquad x \quad y$$

$$p(x, y) = p(x)p(y \mid x) \qquad p(x, y) = p(y)p(x \mid y) \qquad p(x, y) = p(x)p(y)$$

Conditional Independence

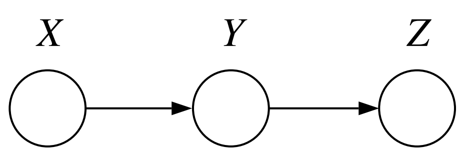
$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B)$$

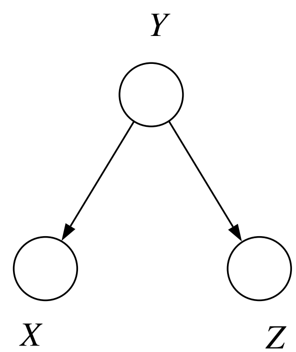
$$p(x_A \mid x_B, x_C) = p(x_A \mid x_B)$$

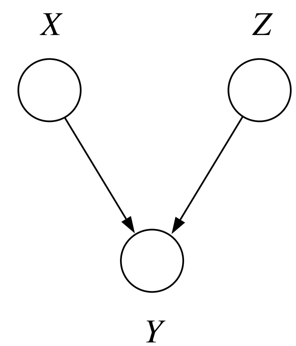
$\Leftrightarrow$ A, C are independent given B

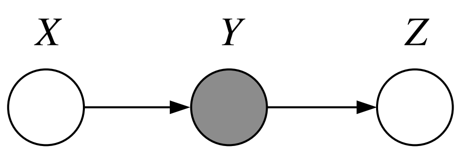
$$A, B, C \subseteq \mathcal{V}$$

GOAL: Characterize conditional independencies which hold for *all* joint distributions which factorize as in a directed graph

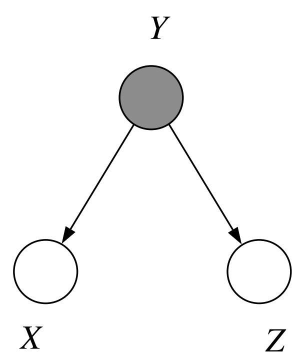








$$X \perp\!\!\!\perp Z \mid Y$$



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

$$X \perp\!\!\!\perp Z$$

Marginally independent but conditionally dependent!

Reachability: Bayes Ball Algorithm

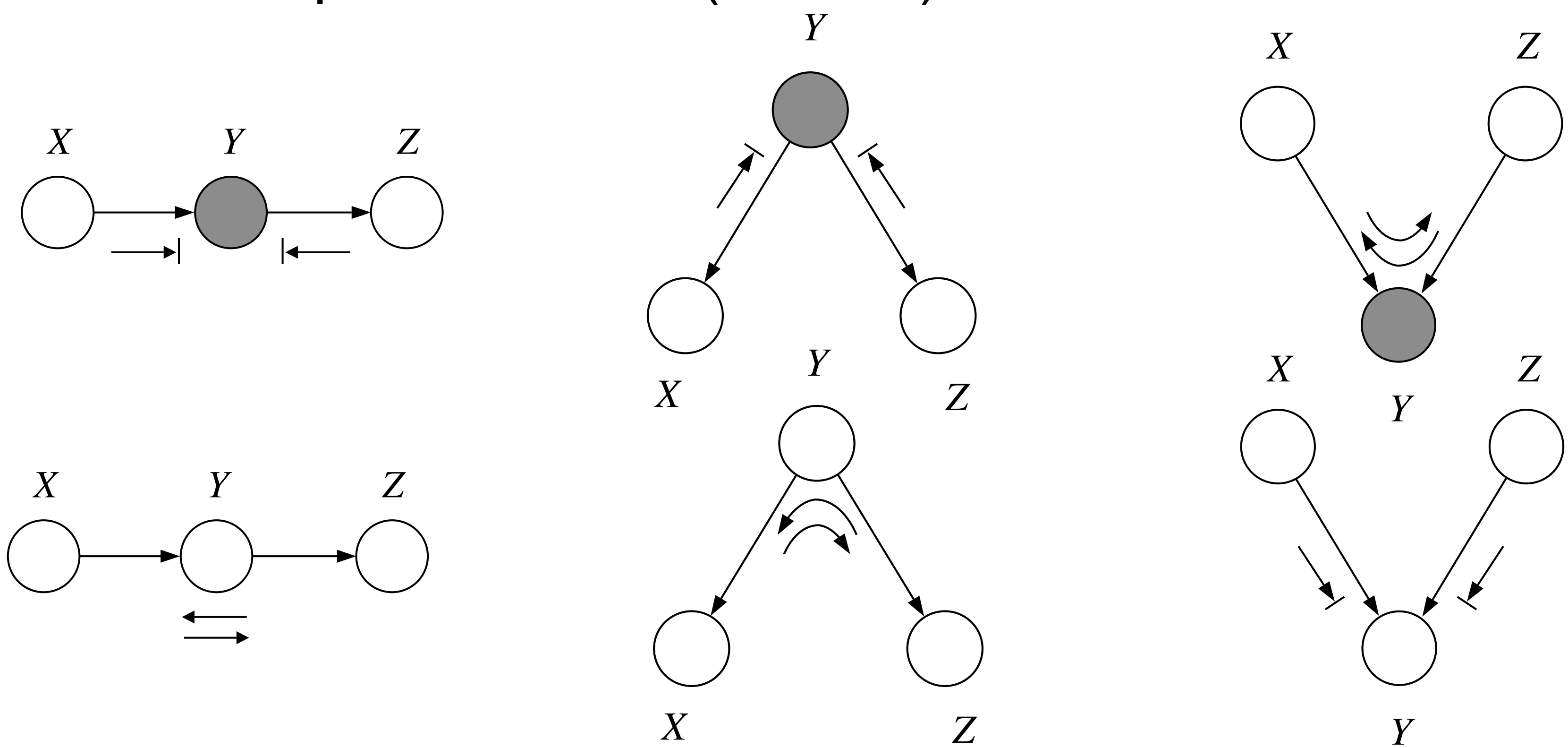
$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B)$$

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

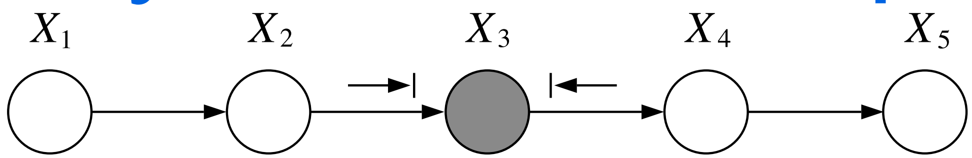
$\Leftrightarrow$ A, C are independent given B

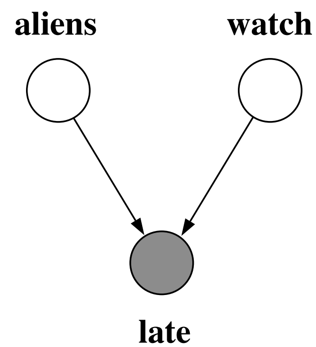
$$A, B, C \subseteq \mathcal{V}$$

Place a ball at each node A, allow to bounce around graph according to rules below, check whether any balls reach nodes C. We interpret observed (shaded) nodes B as follows:

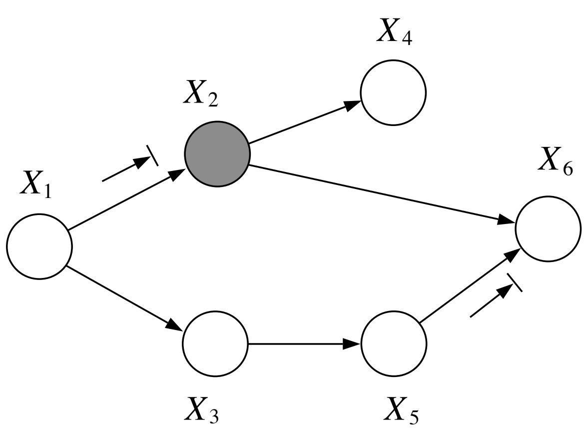


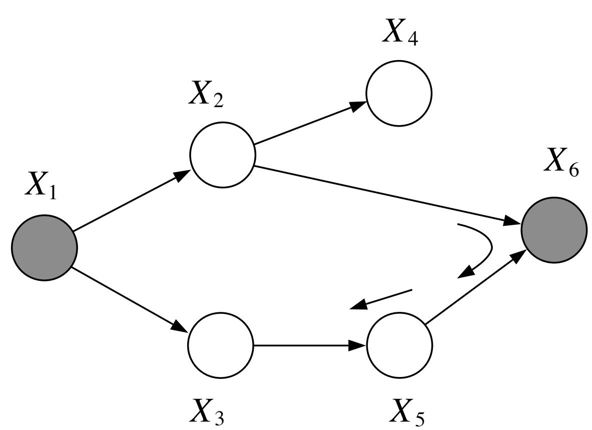
Bayes Ball Examples

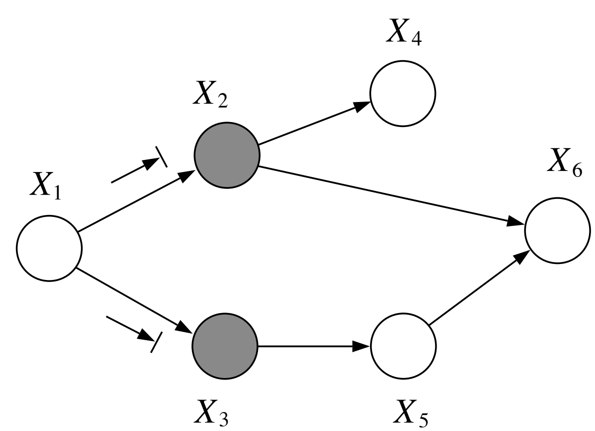




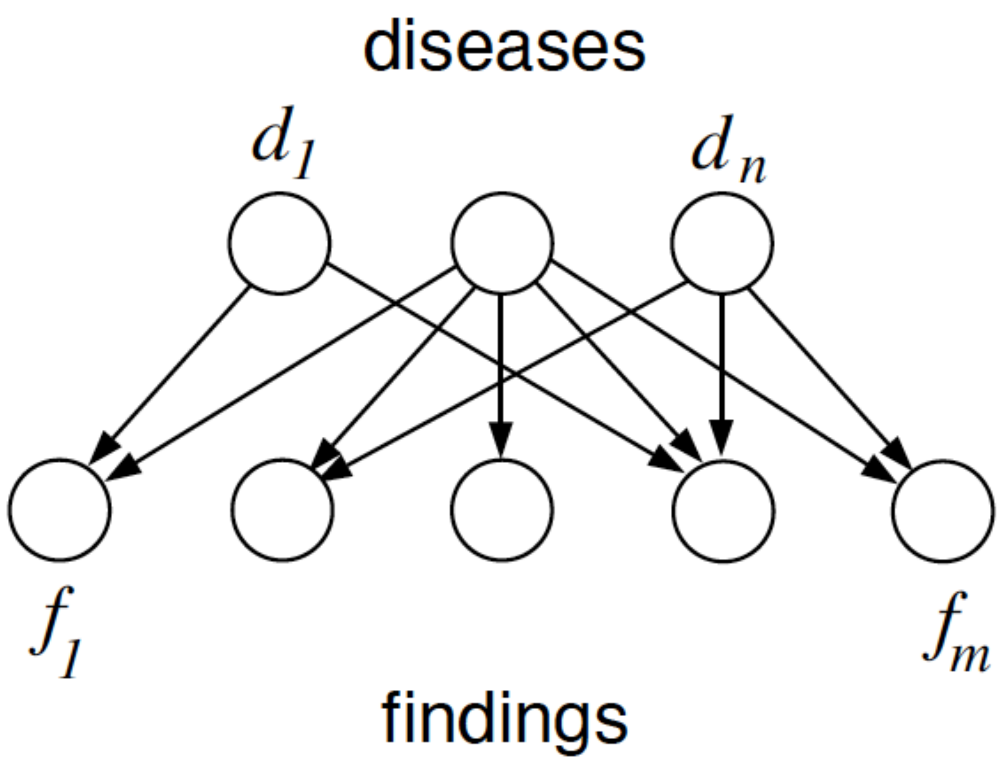
Explaining Away







Example: Medical Diagnosis



Parameterization: Noisy-OR, logistic regression, generalized linear models…