Georgia Institute of Technology

CSIP
Center for Signal & Image Processing

ECE/CS/ISYE 8803

Probabilistic Graphical Models

Module 2 (Part B)
Directed Graphical Models (BNs) and Conditional Independencies

Faramarz Fekri

Center for Signal and Information Processing

Overview

- Independencies Encoded by BN
  - Common parent, Cascade, V Structure
- Minimal map
- Perfect map
- Example (HMM)
- Local Probability Models:
  - Cond. Prob. Tables (CPTs), Deterministic CPDs, Rule/Tree,
  - Generalized Linear Models
  - Hybrid Bayesian Networks

Read Chapter 3 (and a little bit of Ch. 5) of K&F

Knowledge Engineering

Read page 64-67 of book

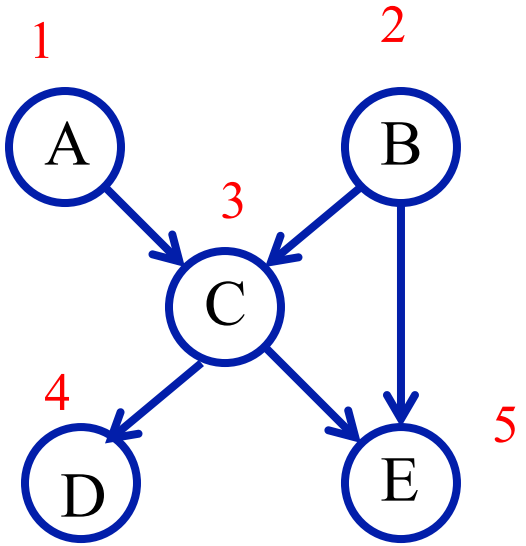
- Variable considerations
  - Observed
  - Hidden variables
  - Irrelevant variables
- Selecting the structure of the graph
  - Causality of variables
  - Generative
  - Coupling
  - Which independencies approximately hold
- Choosing probability distributions
  - Zero probabilities
  - Relative values
  - Order of magnitudes

Generate Samples from Bayesian Networks

- We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.

- First, sort the nodes in topological order.
- Generate a set of samples for (A, B, C, D, E):

$$\text{sample } a_i \sim P(A)$$
$$\text{sample } b_i \sim P(B)$$
$$\text{sample } c_i \sim P(C \mid a_i, b_i)$$
$$\text{sample } d_i \sim P(D \mid c_i)$$
$$\text{sample } e_i \sim P(E \mid c_i, b_i)$$



More on I-MAP

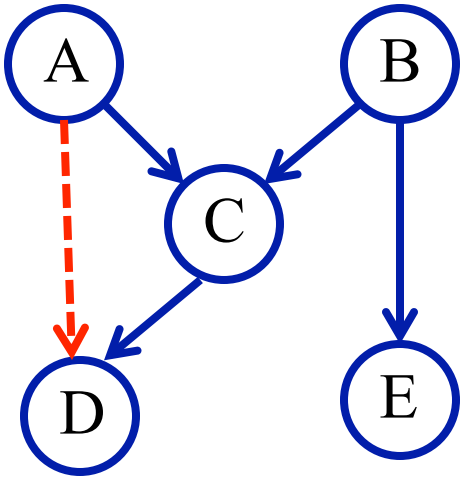
- Adding edges does not violate I-map but it will make graph more expensive.

Theorem: Let **G** be an I-map for ***P***, any DAG **G'** that includes the same directed edges as **G** is also an I-map for ***P***.

Sketch of Proof:

- G is I-map for ***P***, hence ***P*** factorizes according to G.
- A topological ordering for G' is also a topological ordering for G.
- However, some conditional independencies may be lost in G' due to extra edges in G', e.g., $\neg\, A \perp D \mid C$
- Then,

$$I_l(G') \subseteq I_l(G) \subseteq I(P)$$



Active Trail (I)

- **Causal trail** $X \rightarrow Z \rightarrow Y$ **:** active if and only if Z is not observed.
- **Evidential trail** $X \leftarrow Z \leftarrow Y$ **:** active if and only if Z is not observed.
- **Common cause** $X \leftarrow Z \rightarrow Y$ **:** active if and only if Z is not observed.
- **Common effect** $X \rightarrow Z \leftarrow Y$ **:** active if and only if either Z or one of Z's descendants is observed

Active Trail (II)

- Let G be a Bayesian network structure
- Let $X_1 \leftrightarrow \ldots \leftrightarrow X_n$ be a trail in G
- Let **E** be a subset of evidence nodes in G

The trail $X_1 \leftrightarrow \ldots \leftrightarrow X_n$ is active given evidence E if:

- ALL the three-node networks along the trail is active.
  - For every V-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i or one of its descendants is observed
  - No other nodes along the trail is in **E**

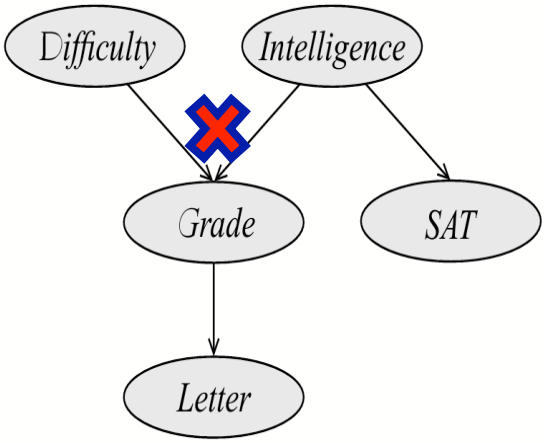
d-separation

Definition : Let X, Y , Z be three sets of nodes in G. We say that X and Y are *d-separated given* Z, denoted *d-sep*$_{\mathcal{G}}(X;Y \mid Z)$, if there is no active trail between any node $X \in X$ and $Y \in Y$ given Z.

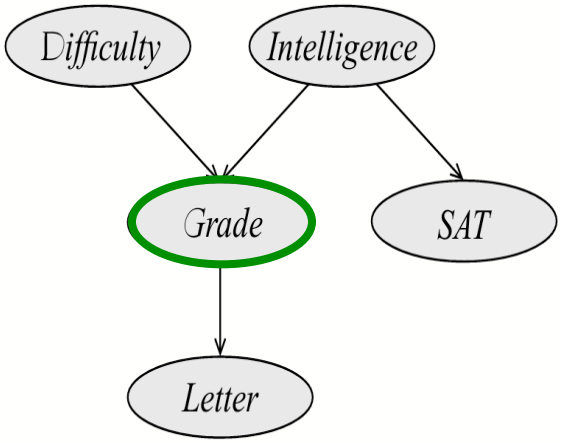
- All independence properties correspond to d-separation:

$$I(G) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : \text{d-sep}_G(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})\}$$

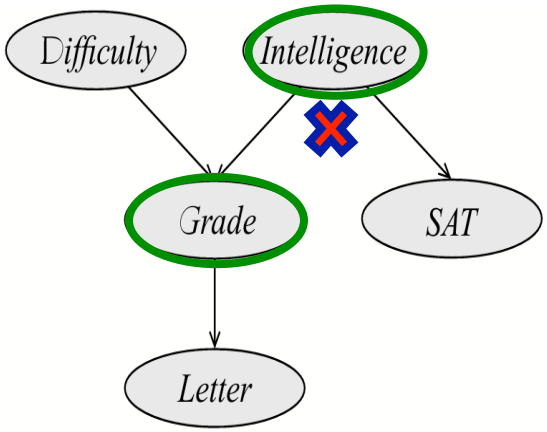
Examples



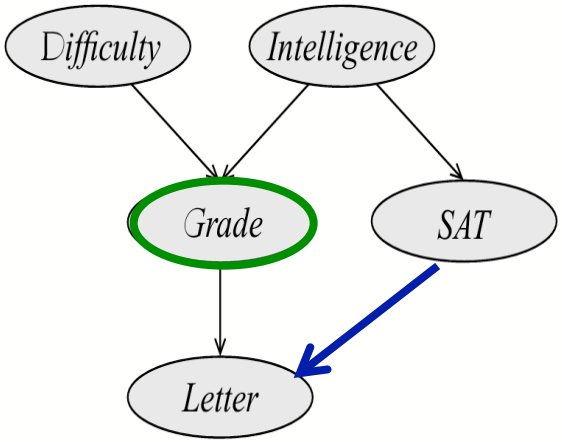
d-sep(S,D): Yes



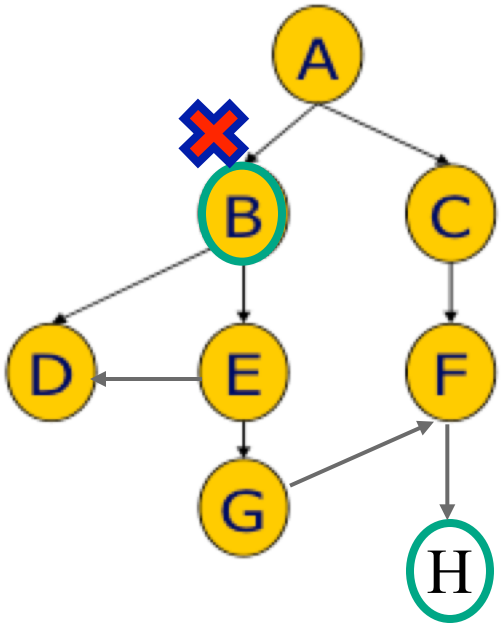
d-sep(S,D | G): No



d-sep(S,D |G, I)
Yes



d-sep(L,I | G): No



d-sep(D,C | B, H): No

Soundness and Completeness of d-separation

- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

$$I(G) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : \text{d-sep}_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

- Soundness of d-separation
  - If P factorizes over G then $I(G) \subseteq I(P)$ (not only $I_{l(G)} \subseteq I(P)$)

- Completeness of d-separation:
  - d-separation captures all possible independencies in P.

- For most P's that factorize over G, $I(G) = I(P)$ (P-map)

- **Theorem :** For **almost all** distributions P that factorize over G, i.e., for all distributions except for a set of "measure zero" in the space of CPD parameterizations, we have that I(P) = I(G)

Example

Consider a distribution P over two variables A and B, where A and B are independent. One possible I-map for P is the network $A \rightarrow B$. For example, we can set the CPD for B to be

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

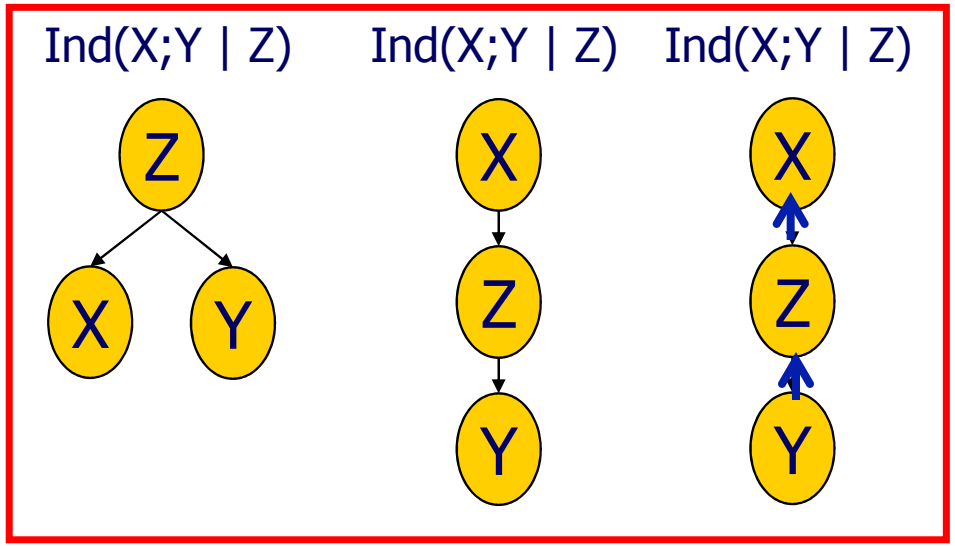
- Graph is an I-map of P
- Yet, independency does not follow from graph d-separation

Algorithm for d-separation

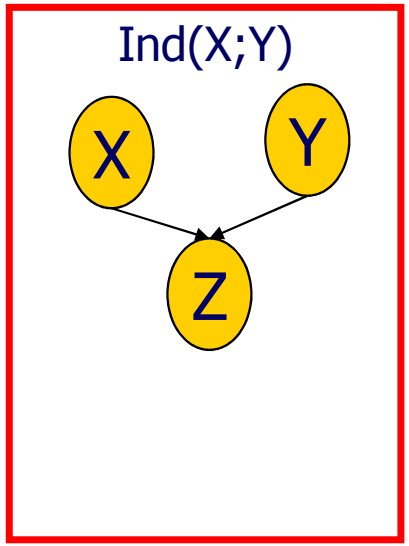
- Goal: answer whether d-sep(**X**;**Y** | **Z**, G)
  - Enumerate all possible trails between X and Y? NO
- Algorithm:
  - Mark all nodes in **Z** or that have descendants in **Z**
  - BFS traverse G from **X**
  - Stop traversal at blocked nodes:
    - Node that is in the middle of a v-structure and not in marked set
    - Not such a node but is in **Z**
  - If we reach any node in **Y** then there is an active path and thus d-sep(**X**;**Y** | **Z**, G) does not hold
- Theorem: algorithm returns all nodes reachable from **X** via trails that are active in G

Uniqueness of BN

- I(G) describe all conditional independencies in G
- Different Bayesian networks can have same Ind.



Equivalence class I

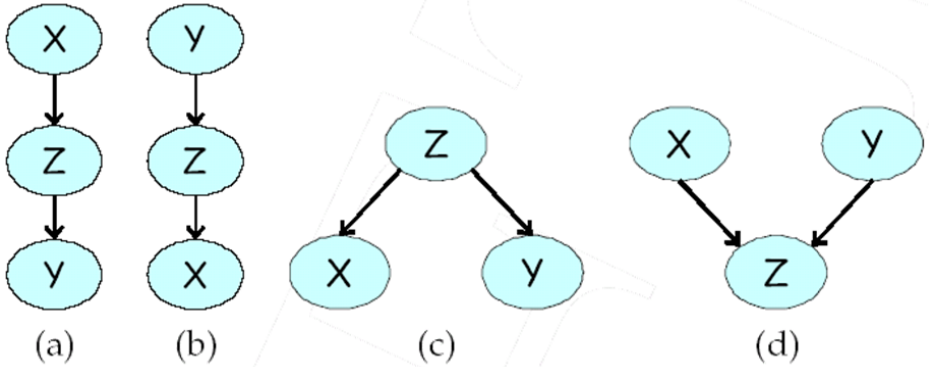


Equivalence class II

Two BN graphs G_1 and G_2 are I-equivalent if $I(G_1) = I(G_2)$

I-equivalence between Graphs

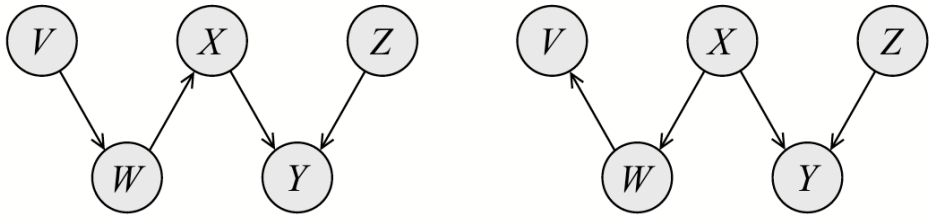
- The set of all graphs over X is partitioned into a set of mutually exclusive and exhaustive *I-equivalence classes*, which are the set of equivalence classes induced by the I-equivalence relation.



- Any distribution P that can be factorized over one of these graphs can be factorized over the other.
- Furthermore, there is no intrinsic property of P that would allow us associate it with one graph rather than an equivalent one.
- This observation has important implications with respect to our ability to determine the directionality of influence.
- Test for I-equivalence: d-separation

Detecting I-equivalence

- Necessary condition: same graph skeleton
  - Otherwise, can find active path in one graph but not other
  - But, not sufficient: v-structures
- Sufficient condition: same skeleton and v-structures
  - But, not necessary: complete graphs (no independence)



- Define X→Z←Y as immoral if X, Y are not directly connected
  - Necessary and Sufficient: same skeleton and immoral set of v-structures

From Distributions to BNs

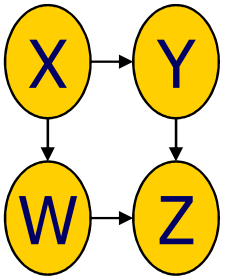
- From d-separation we learned:
  - Start from local Markov assumptions, obtain all independence assumptions encoded by graph
  - For most *P*'s that factorize over *G*, I(*G*) = I(*P*)
  - All of this discussion was for a given *G* that is an I-map for *P*
- Now, give me a *P*, how can I get a *G*?
  - i.e., give me the independence assumptions entailed by *P*
  - Many *G* are "equivalent", how do I represent this?
  - Most of this discussion is not about practical algorithms, but useful concepts that will be used by practical algorithms

Minimal I-map

- Complete graph is a (trivial) I-map for any distribution, yet it does not reveal any of the independence structure in the distribution.
  - Meaning that the graph dependence is arbitrary, thus by careful parameterization an dependencies can be captured
  - We want a graph that has the maximum possible I(G), yet still $\subseteq$ I(P)

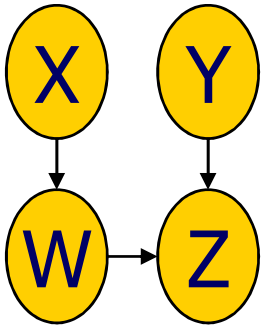
- **Defn :** A graph object G is a *minimal I-map* for a set of independencies I if it is an I-map for I, and if the removal of even a single edge from G renders it not an I-map.

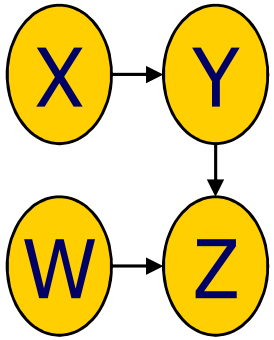
Example: minimal I-map

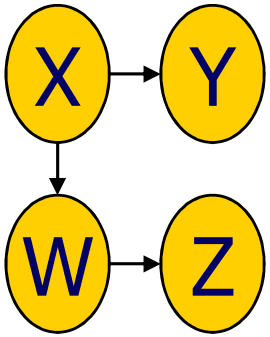
- Example: if 

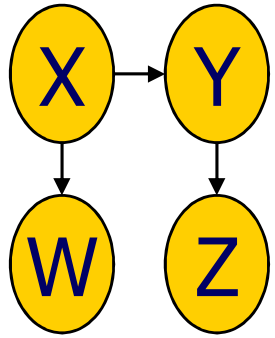
is a minimal I-map for P,

Then the followings are not I-map for P.









Constructing Minimal I-map

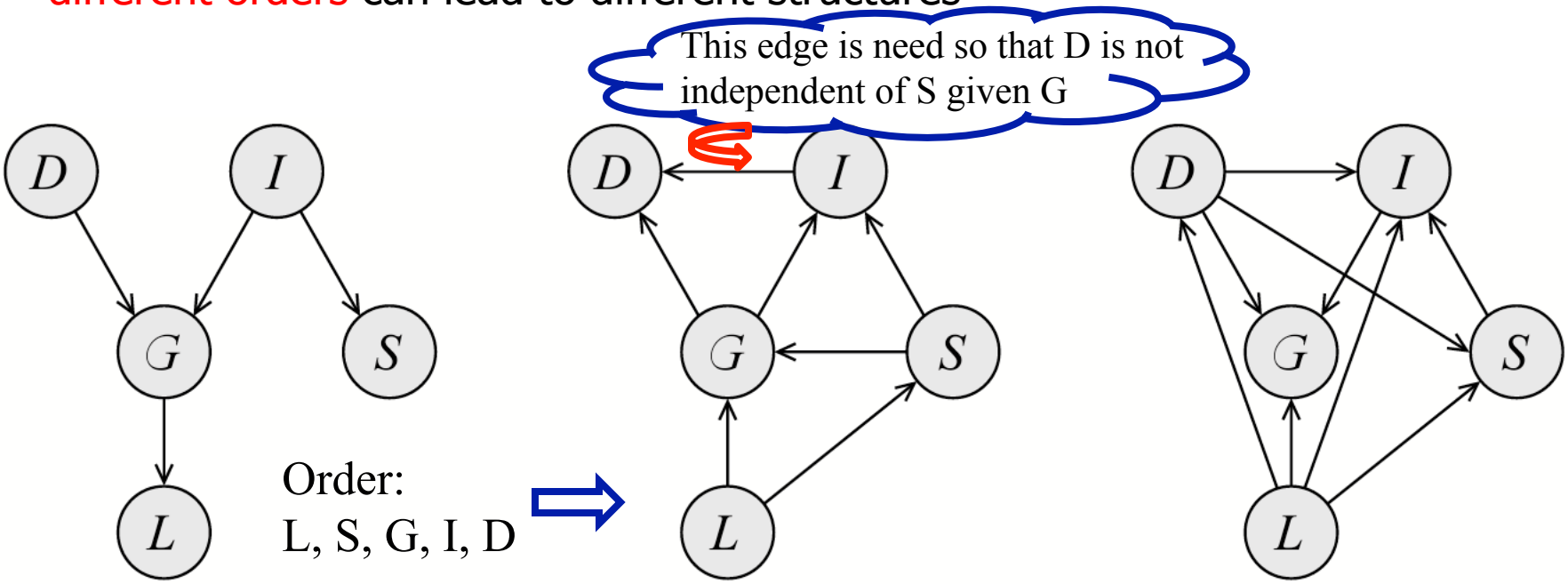
- A Bayesian network is a pair (G,P)
  - P factorizes over G
  - P is specified as set of CPDs associated with G's nodes
  - Additional requirement: G is a minimal I-map for P

- Given a set of variables and conditional independence assertions of P
- Choose an ordering on variables, e.g., X_1, , X_n
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, $\mathbf{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1, \ldots, X_{i-1}\}$ such that local Markov assumption holds – X_i independent of rest of $\{X_1, \ldots, X_{i-1}\}$, given parents $\mathbf{Pa}_{X_i}$
  - Define/learn CPT – $P(X_i | \mathbf{Pa}_{X_i})$

Minimal I-map is Not Unique

- A distribution may have several minimal I-maps
  - Each corresponding to a specific node-ordering
- Example:

  Applying the same I-Map construction process with different orders can lead to different structures



- Note that all three graphs are minimal I-map for P but they fail to capture some or all inependencies that hold in P.

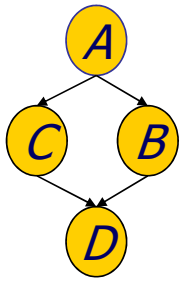
Perfect Maps

We aim to find a graph $\mathcal{G}$ that precisely captures the independencies in a given distribution P.

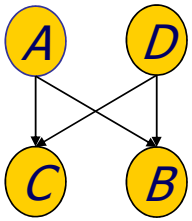
We say that a graph $\mathcal{K}$ is a perfect map (P-map) *for a set of independencies $\mathcal{I}$ if we have that $\mathcal{I}(\mathcal{K}) = \mathcal{I}$. We say that $\mathcal{K}$ is a perfect map for P if $\mathcal{I}(\mathcal{K}) = \mathcal{I}(P)$.*

- Thm: not every distribution has a perfect map as DAG.
  - some structures cannot be represented in a BN
- Proof by counter example:

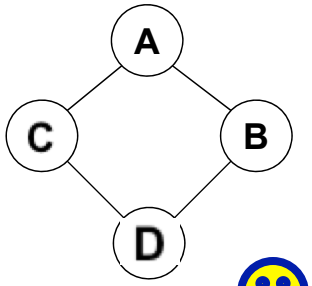
  Independencies in P: Ind(A;D | B,C), and Ind(B;C | A,D)



Ind(B;C | A,D) does not hold



Ind(A,D) also holds



MRF

Finding a Perfect Map

- If P has a P-Map, can we find it?
  - Not uniquely, since I-equivalent graphs are indistinguishable
  - Thus, represent I-equivalent graphs and return it
- Recall I-Equivalence
  - Necessary and Sufficient: same skeleton and immoral set of v-structures
- Finding P-Maps
  - Step I: Find skeleton
  - Step II: Find immoral set of v-structures
  - Step III: Direct constrained edges

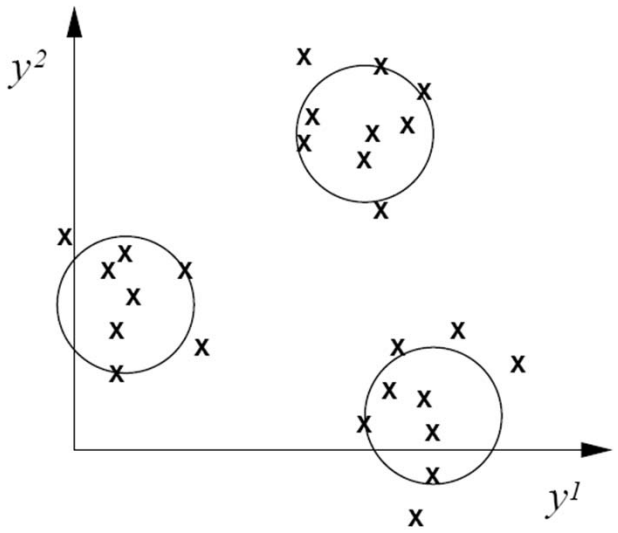
Summary

- **Definition:** A *Bayesian network* is a pair (G, P) where P factorizes over G, and where P is specified as set of local conditional probability dist. CPDs associated with G's nodes.
- A BN capture "causality", "generative schemes", "asymmetric influences", etc., between entities

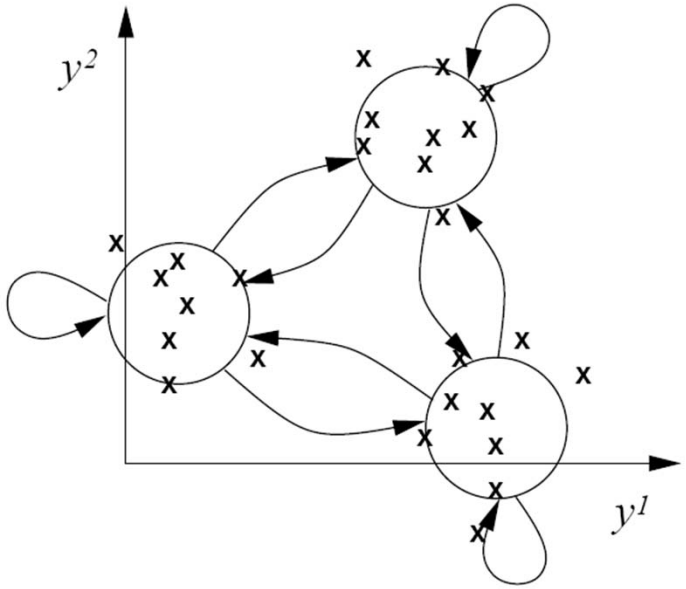
- Local independencies $I_L(G)$– basic BN independencies
- d-separation – all independencies via graph structure
- G is an I-Map of P if and only if P factorizes over G
- I-equivalence – graphs with identical independencies
- Minimal I-Map
  - All distributions have I-Maps (sometimes more than one)
  - Minimal I-Map does not capture all independencies in P
- Perfect map – not every distribution P has one

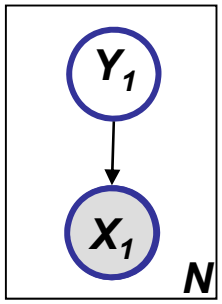
BN Example: Hidden Markov Models

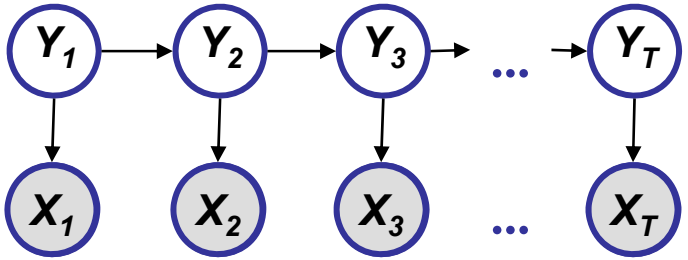
Static mixture



Dynamic mixture







Notation: Plate model

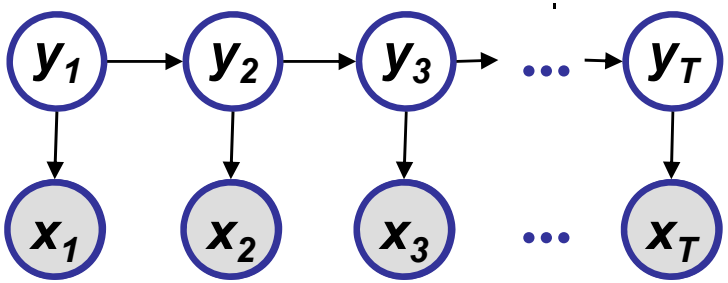
E. Xing

HMM setup

- Observation space

  Alphabetic set: $\mathbb{C} = \{c_1, c_2, \cdots, c_K\}$

  Euclidean space: $\mathbb{R}^d$



- Index set of hidden states

$$\mathbb{I} = \{1, 2, \cdots, M\}$$

- Transition probabilities between any two states

$$p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j},$$

or $$p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,1}, \ldots, a_{i,M}), \forall i \in \mathbb{I}.$$

- Start probabilities

$$p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$$

- Emission probabilities associated with each state

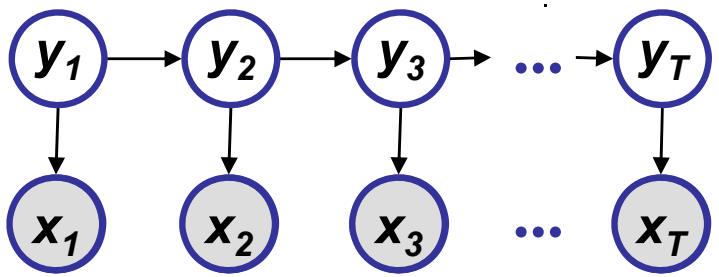
$$p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,1}, \ldots, b_{i,K}), \forall i \in \mathbb{I}.$$

or in general:

$$p(x_t \mid y_t^i = 1) \sim \text{f}(\cdot \mid \theta_i), \forall i \in \mathbb{I}.$$

Probability of a Parse in HMM

- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$ and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T$,
- To find how likely is the parse: (given our HMM and the sequence)



$$p(\mathbf{x}, \mathbf{y}) = p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \quad \text{(Joint probability)}$$

$$= p(y_1)\, p(x_1 \mid y_1)\, p(y_2 \mid y_1)\, p(x_2 \mid y_2) \ldots p(y_T \mid y_{T-1})\, p(x_T \mid y_T)$$

$$= p(y_1)\, \mathrm{P}(y_2 \mid y_1) \ldots p(y_T \mid y_{T-1}) \times p(x_1 \mid y_1)\, p(x_2 \mid y_2) \ldots p(x_T \mid y_T)$$

$$= p(y_1, \ldots\ldots, y_T)\, p(x_1 \ldots\ldots x_T \mid y_1, \ldots\ldots, y_T)$$

- Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \pi_{y_1} \prod_{t=2}^{T} a_{y_{t-1}, y_t} \prod_{t=1}^{T} p(x_t \mid y_t)$
- Posterior probability: $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$

- We will learn how to do this explicitly (polynomial time)

Example of HMM: Casino

A casino has two dice:

- Fair die
  P(1) = P(2) = P(3) = P(5) = P(6) = 1/6
- Loaded die
  P(1) = P(2) = P(3) = P(5) = 1/10
  P(6) = 1/2

Casino player switches back-&-forth between fair and loaded die once every 20 turns

Game:

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

GIVEN: A sequence of rolls by the casino player

12455264621461461361366616646616366163661636165156151151461235623 44

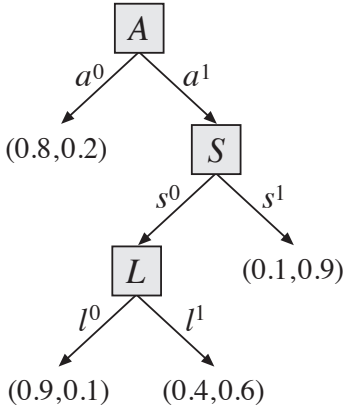
E. Xing

Local Probability Models (Chapter 5)

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Pa_{X_i})$$

- For solely discrete-valued random variables, we can always resort to a tabular representation of CPDs.
  - However, tabular representation can rapidly become large and unwieldy as the number of parents grows.
  - Deterministic CPDs (e.g., the child is OR operation on parents)
  - Tree/rule CPDs

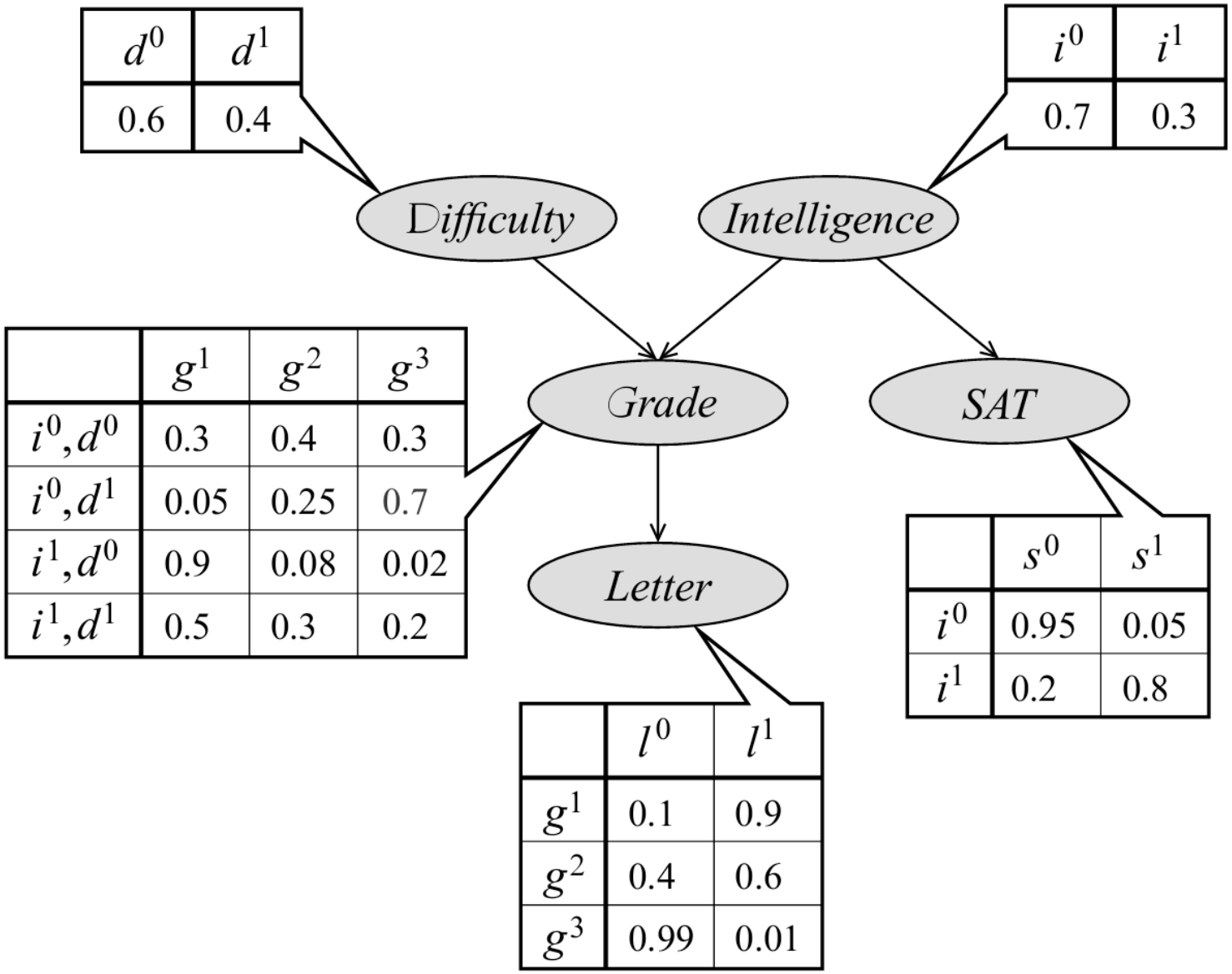
$$P(J \mid A, S, L)$$



  - Independence of Causal Influence
    - Noisy-OR-Model
    - Generalized Linear Models

logistic Sigmoid (binary Y): $P(y^1 \mid X_1, \ldots, X_k) = \text{sigmoid}(w_0 + \sum_{i=1}^{k} w_i X_i)$

Conditional Probability Tables (Tabular Model)



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Local Prob. Density Functions (Continuous case)

- How about random variables with continuous values?

- One solution: discretize
  - Often requires too many value states
  - Loses domain structure

- Can combine continuous and discrete variables, resulting in hybrid networks
- Inference and learning may become more difficult

Review: Multivariate Gaussian Density functions

- A multivariate Gaussian distribution over $X_1,...X_n$ has
  - n_x1 mean vector μ
  - n_xn positive definite covariance matrix Σ
    (positive definite: $\forall x \in \Re^n : x^T \sum x > 0$)
  - Joint density function:

$$p(x) = \frac{1}{(2\pi)^{n/2} |\sum|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right]$$

    - $\mu_i = E[X_i]$
    - $\Sigma_{ii} = Var[X_i]$
    - $\Sigma_{ij} = Cov[X_i,X_j] = E[X_iX_j] - E[X_i]E[X_j]$ $(i \neq j)$

- Independencies can be determined from parameters
  - If $\mathbf{X}=X_1,...X_n$ have a joint normal distribution $N(\mu;\Sigma)$ then $(X_i \perp X_j)$ iff $\Sigma_{ij}=0$ (for $i \neq j$)
  - Does not hold in general for non-Gaussian distributions

Linear Gaussian Model- special case

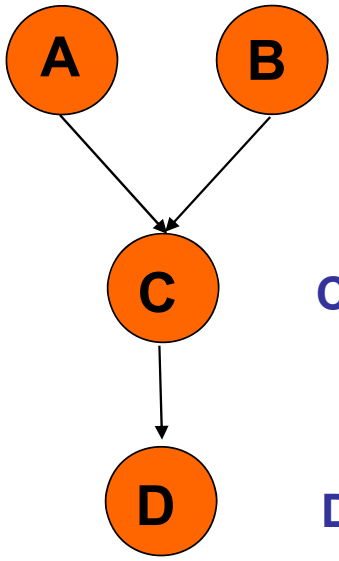
Gaussian density functions

Single variable case:

$$P(X) \sim N(\mu, \sigma^2) \quad \text{if} \quad p(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

- Linear Model: Mean of child is a linear function of parents, and that the variance of child does not depend on parents.



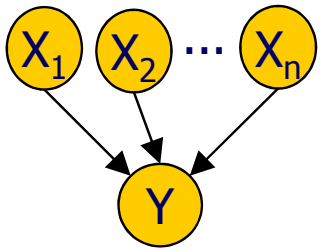
$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d+C, \Sigma_d)$

More in Chapter 5
(not covered by lectures)

Linear Gaussian CPDs (General Case)

- Y is a continuous variable with parents $X_1,...X_n$
- Y has a **linear Gaussian model** if it can be described using parameters $\beta_0,...,\beta_n$ and σ^2 such that
  - $P(Y \mid x_1,...,x_n) = N(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n; \sigma^2)$
  - Vector notation: $P(Y \mid \boldsymbol{x}) = N(\beta_0 + \beta^T \boldsymbol{x}; \sigma^2)$



Can be viewed as (where x_i are constants):

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

where ϵ is a Gaussian random variable with mean 0 and variance σ^2, representing the noise in the system.

- Cons
  - Fixed variance (variance cannot depend on parents values)

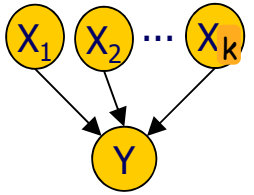
More in Chapter 5
(not covered by lectures)

Gaussian Bayesian Networks

- All variables are continuous
- All of the CPDs are linear Gaussians

Let Y be a linear Gaussian of its parents $X_1, \ldots, X_k$:

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$



Assume that $X_1, \ldots, X_k$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$. Then:

- *The distribution of Y is a normal distribution $p(Y) = \mathcal{N}\left(\mu_Y; \sigma_Y^2\right)$ where:*

$$\begin{aligned} \mu_Y &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu} \\ \sigma_Y^2 &= \sigma^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}. \end{aligned}$$

- *The joint distribution over $\{\boldsymbol{X}, Y\}$ is a normal distribution where:*

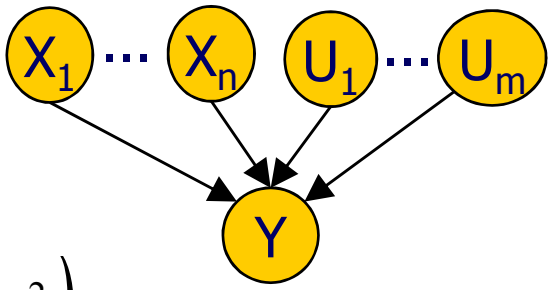
$$\mathbf{Cov}[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}.$$

Revisit this in couple of lectures

Hybrid Bayesian Networks

- Models of continuous and discrete variables
  - Continuous variables with discrete parents
  - Discrete variables with continuous parents

- Conditional Linear Gaussians
  - Y continuous variable
  - **X** = $\{X_1,\ldots,X_n\}$ continuous parents
  - **U** = $\{U_1,\ldots,U_m\}$ discrete parents
  - $\forall \boldsymbol{u} \in \boldsymbol{U}: \quad P(Y \mid \boldsymbol{u}, \boldsymbol{x}) = N\left(a_{\boldsymbol{u},0} + \sum_{i=1}^{n} a_{\boldsymbol{u},i}\boldsymbol{x}_i; \sigma_{\boldsymbol{u}}^2\right)$



- A conditional Linear Bayesian network is one where
  - Discrete variables have only discrete parents
  - Continuous variables have only CLG CPDs

More in Chapter 5
(not covered by lectures)

Hybrid BN: Cont. Parent for a discrete Child

Continuous parents for discrete children

- Threshold models

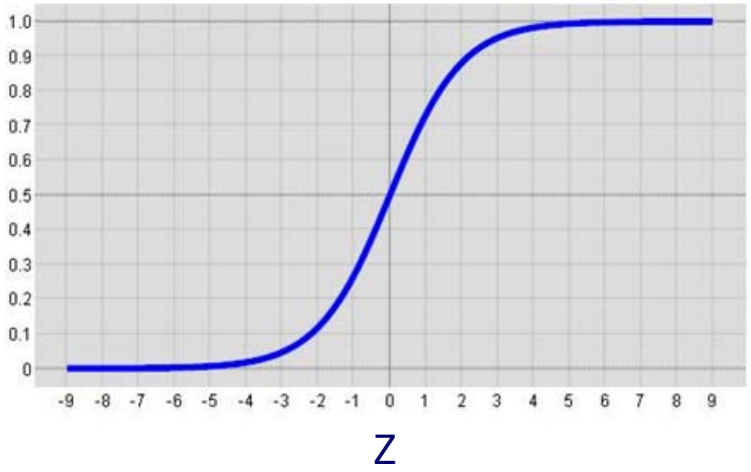
$$P(Y = y^1 \mid x) = \begin{cases} 0.9 & x < 10 \\ 0.05 & \text{otherwise} \end{cases}$$

- Linear sigmoid (logit function)

$$P(Y = y^1 \mid x_1, \ldots, x_k) = \text{logit}(w_o + \sum_{i=1}^{n} w_i x_i)$$

Logit function (smooth step function)

$$\text{logit}(z) = \frac{e^z}{1 + e^z}$$



More in Chapter 5
(not covered by lectures)