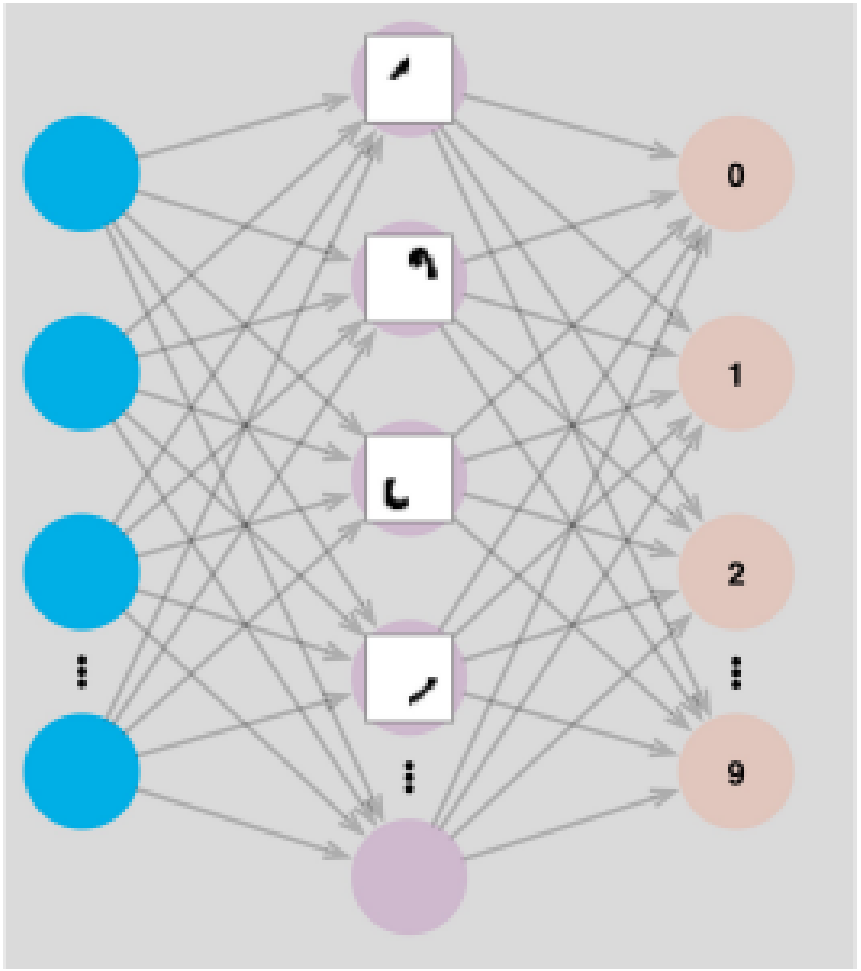
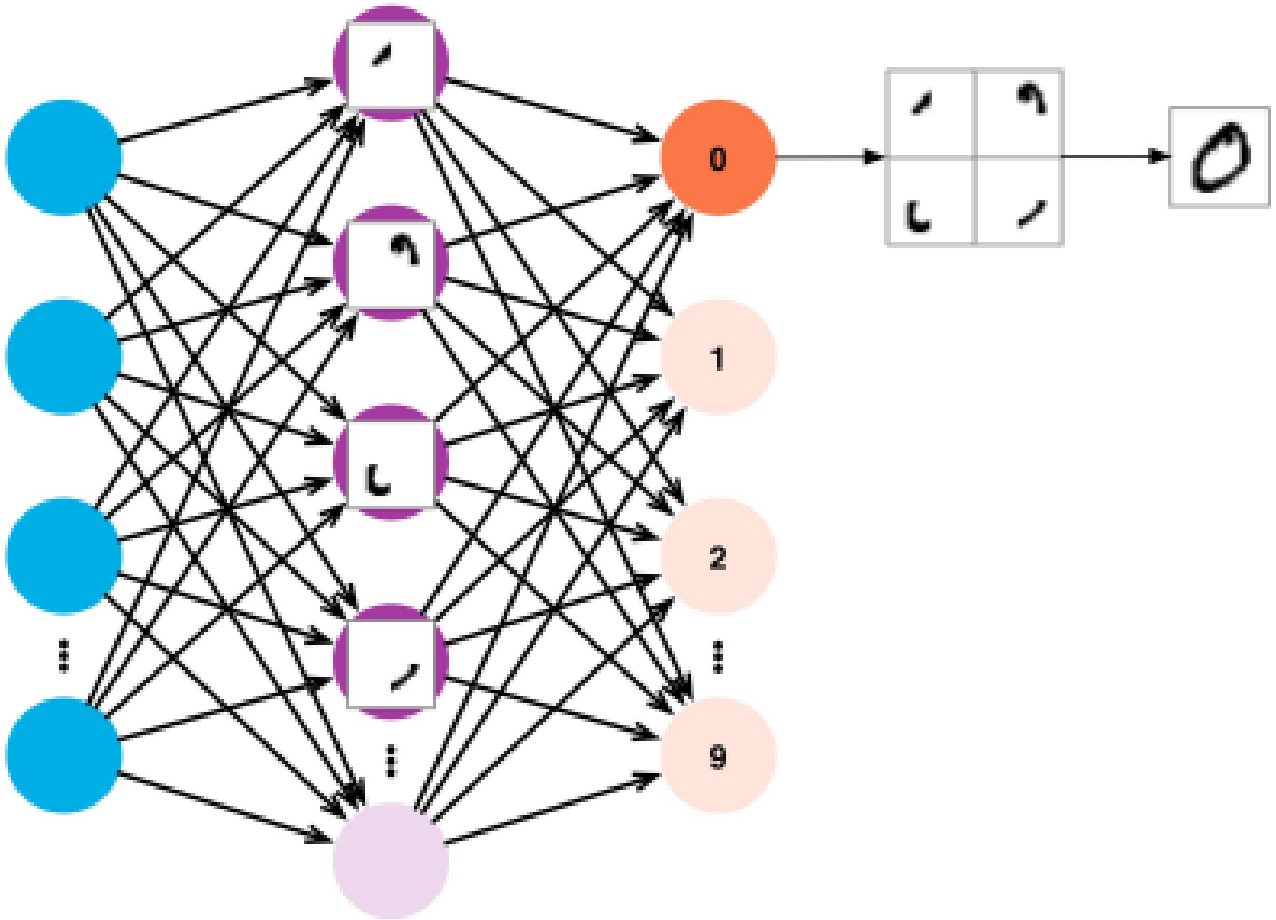
Predictions and Probabilities

Rahul Shome
IML 2024

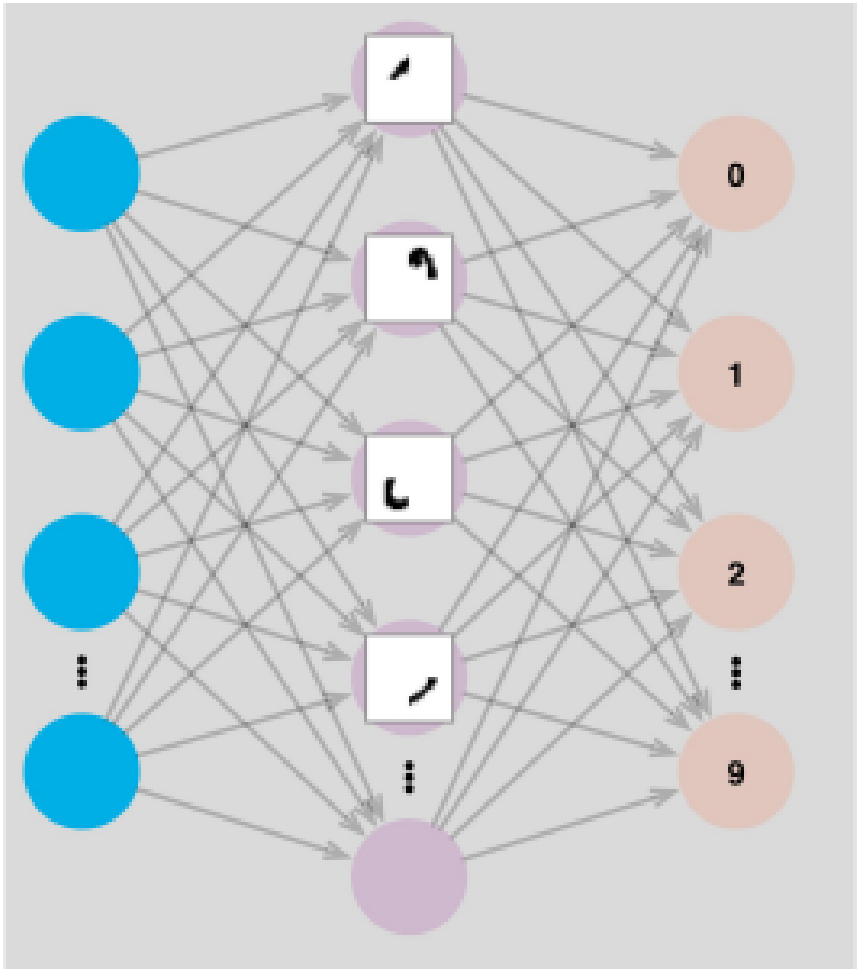
Based on content from Rutgers University

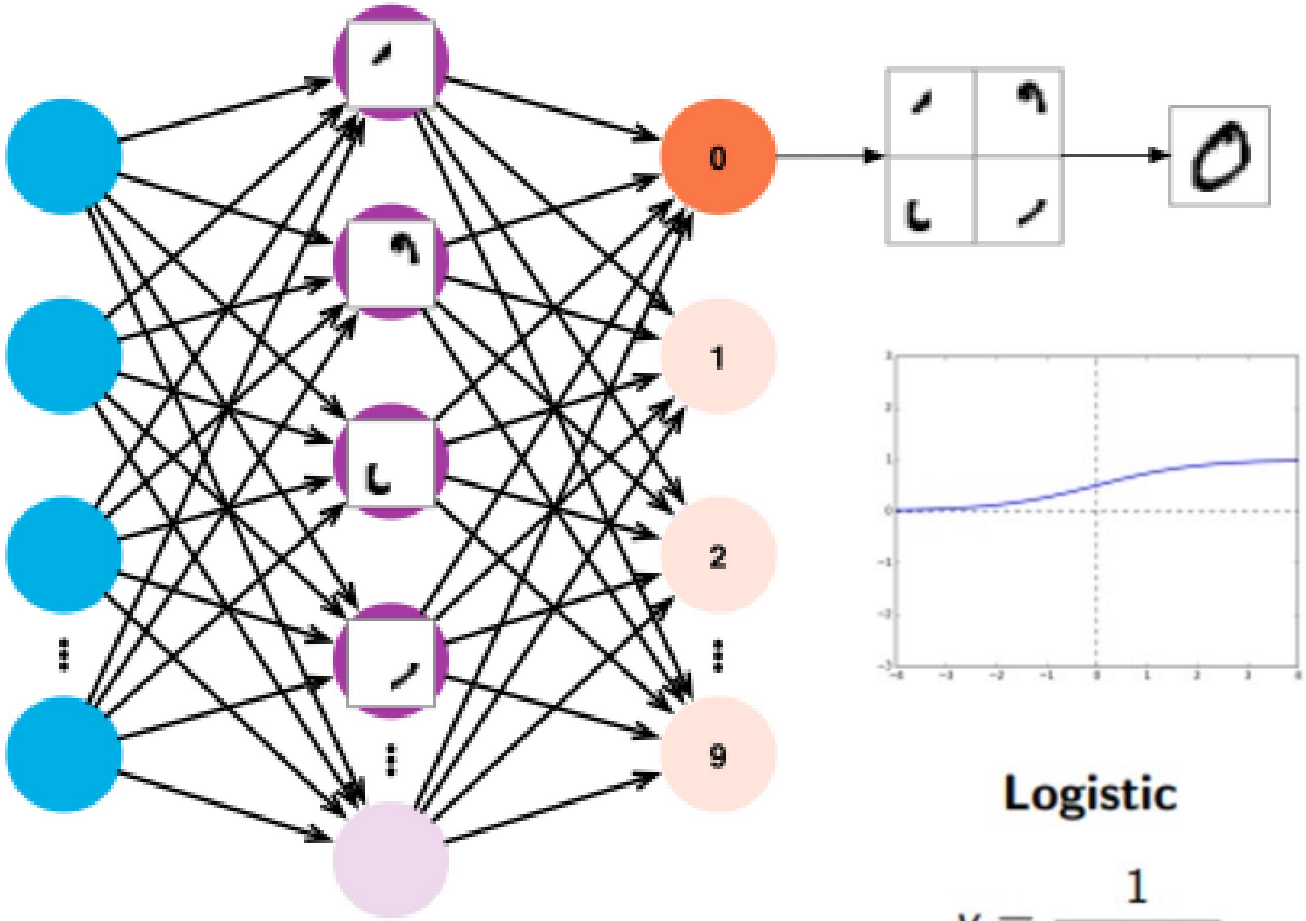
Weights and Hidden Layers



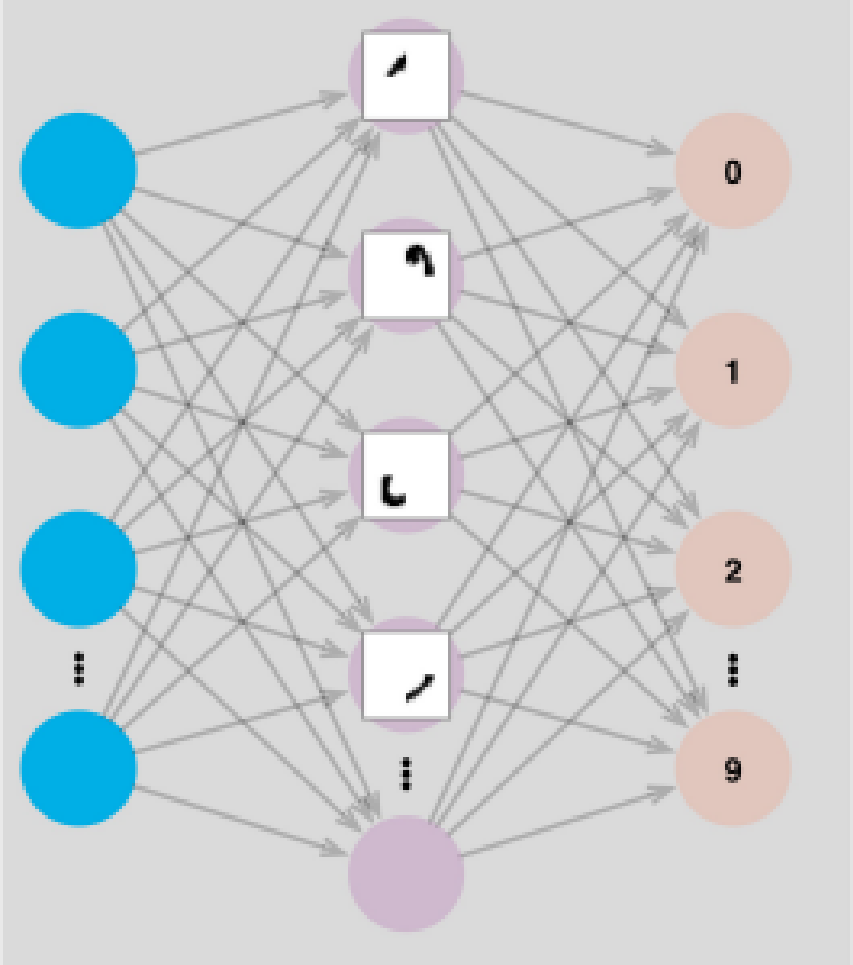


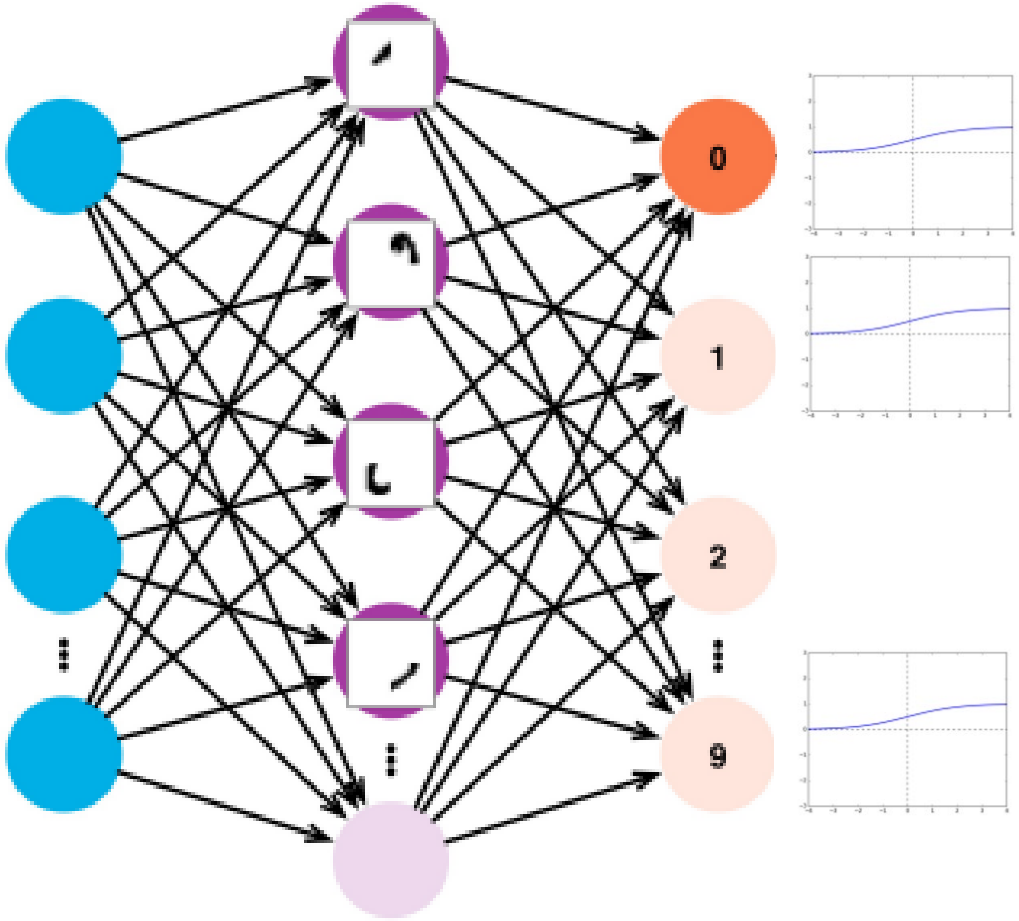
Weights and Hidden Layers



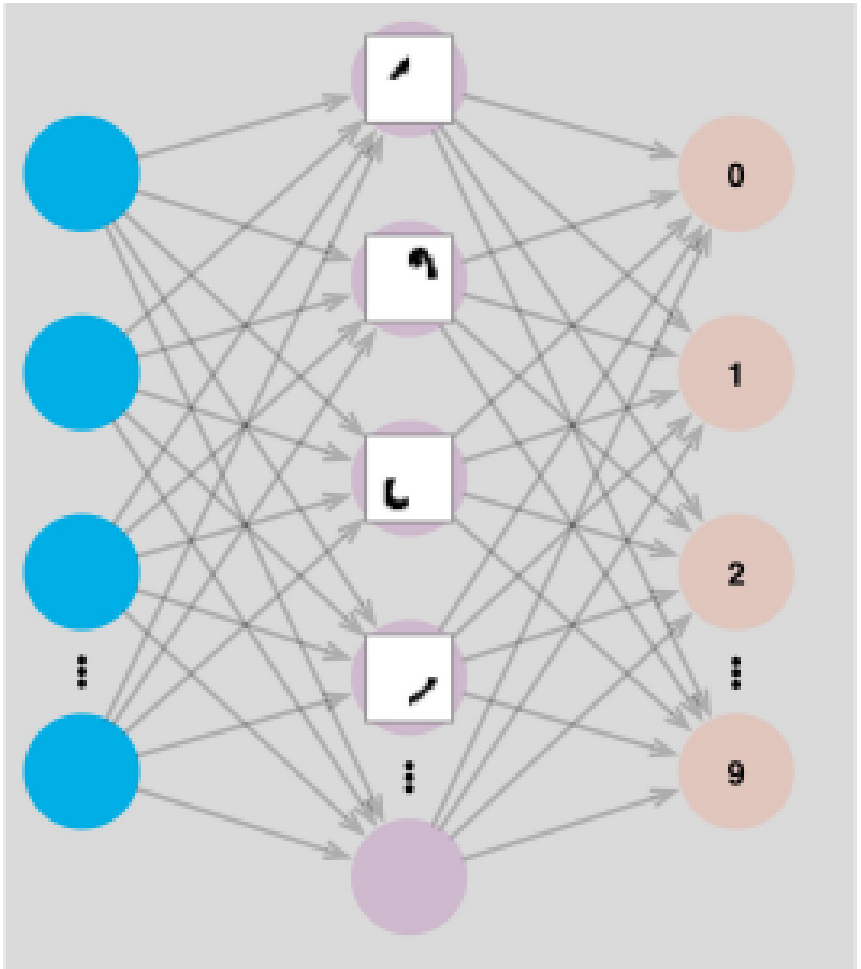


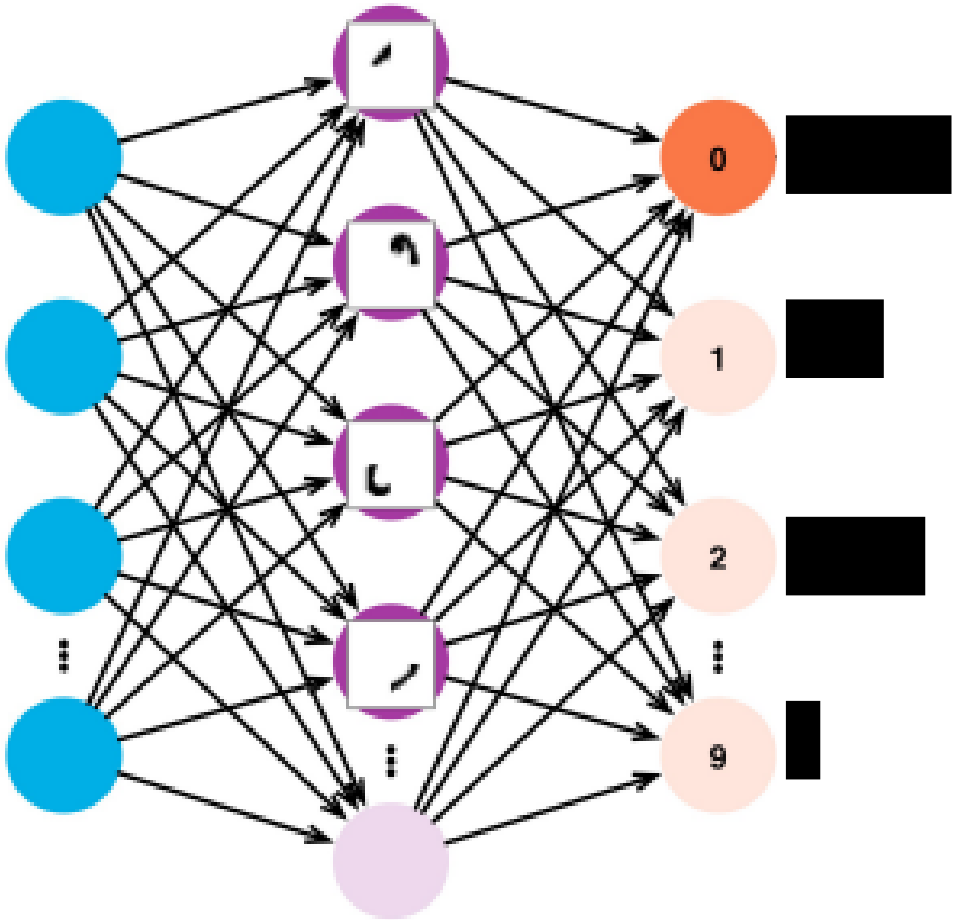
Weights and Hidden Layers





Weights and Hidden Layers





LLM says what?

Linear regression - maximum likelihood

Training data N input, output pairs $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}, \mathbf{x}_n \in \mathbb{R}^D, y_n \in \mathbb{R}$

Assumptions:

- Underlying function is **linear**, $f_\theta(\mathbf{x}) = \sum_{d=1}^{D} \theta_d x_d = \theta^\intercal \mathbf{x}, \theta \in \mathbb{R}^D$
- Due to measurement noise, observed y is a noisy version of $f(\mathbf{x})$

$$\begin{bmatrix} f_\theta(\mathbf{x}_1) \\ f_\theta(\mathbf{x}_2) \\ \vdots \\ f_\theta(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} \theta^\intercal \mathbf{x}_1 \\ \theta^\intercal \mathbf{x}_2 \\ \vdots \\ \theta^\intercal \mathbf{x}_N \end{bmatrix} = X\theta$$

Likelihood: $p(\mathcal{D} \mid \theta) = \prod_n \mathcal{N}(y_n; f(x_n), \sigma^2) = \prod_n \mathcal{N}(y_n; x_n^\intercal \theta, \sigma^2)$

Factorise across data points

Mean = linear mapping

Measurement noise
Constant across data points

Maximum likelihood, $\text{argmax}_\theta\ p(\mathcal{D} \mid \theta)$ **is equiv. to least squares**

Linear regression - Maximum a posteriori

Training data N input, output pairs $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}, \mathbf{x}_n \in \mathbb{R}^D, y_n \in \mathbb{R}$

Assumptions:

- Underlying function is **linear**, $f_\theta(\mathbf{x}) = \sum_{d=1}^{D} \theta_d x_d = \theta^\intercal \mathbf{x}, \theta \in \mathbb{R}^D$
- Due to measurement noise, observed y is a noisy version of $f(\mathbf{x})$

$$\begin{bmatrix} f_\theta(\mathbf{x}_1) \\ f_\theta(\mathbf{x}_2) \\ \vdots \\ f_\theta(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} \theta^\intercal \mathbf{x}_1 \\ \theta^\intercal \mathbf{x}_2 \\ \vdots \\ \theta^\intercal \mathbf{x}_N \end{bmatrix} = X\theta$$

Likelihood: $p(\mathcal{D} \mid \theta) = \prod_n \mathcal{N}(y_n; f(x_n), \sigma^2) = \prod_n \mathcal{N}(y_n; x_n^\intercal \theta, \sigma^2)$

Prior: $p(\theta) = \mathcal{N}(\theta; \mathbf{0}, \sigma_o^2 \mathrm{I}) = \prod_d \mathcal{N}(\theta_d; 0, \sigma_o^2)$

Factorise across dimensions — Zero mean — Same variance for all dimensions

Maximum a posteriori (MAP), $\operatorname{argmax}_\theta p(\mathcal{D} \mid \theta) p(\theta)$ **is equiv. to regularised least squares**

Linear regression - Prediction

Point estimate

$$\theta_{\mathrm{ML}} = (X^\intercal X)^{-1} X^\intercal \mathbf{y} \text{ or } \theta_{\mathrm{MAP}} = (X^\intercal X + N\lambda \mathrm{I})^{-1} X^\intercal \mathbf{y}$$

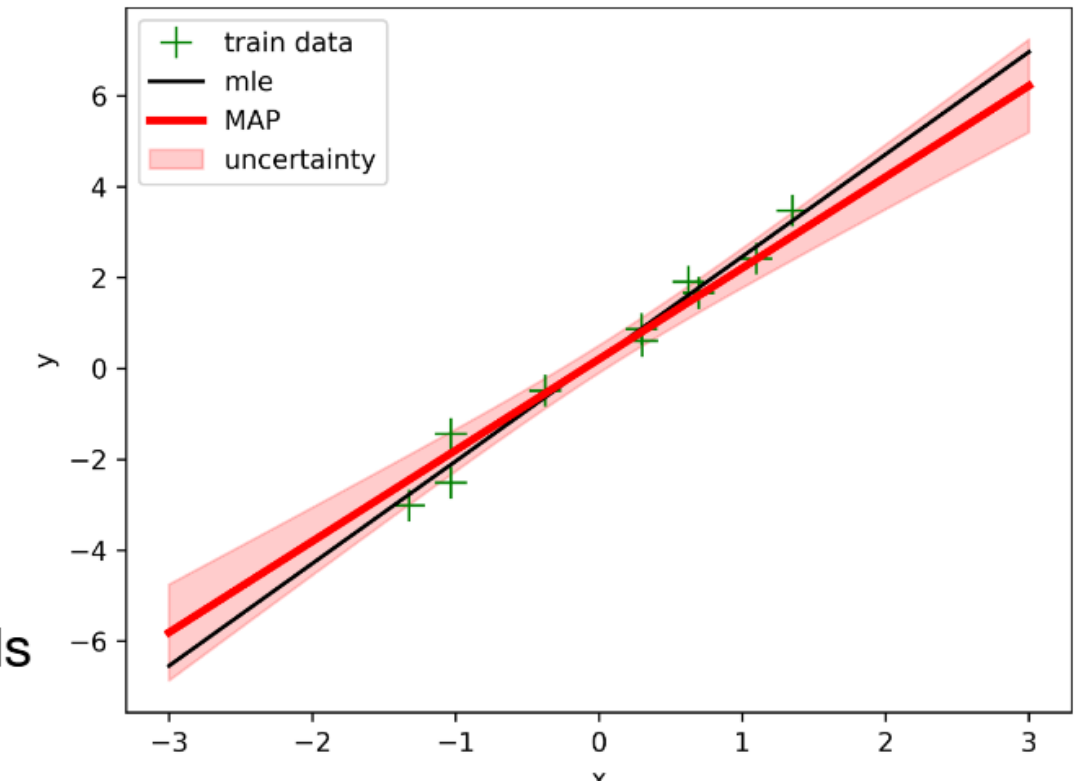
$$f(\mathbf{x}^*) = \theta_{\mathrm{ML}}^\intercal \mathbf{x}^* \quad \text{or} \quad \theta_{\mathrm{MAP}}^\intercal \mathbf{x}^*$$

Exact posterior $p(\theta \mid \mathcal{D}) = \mathcal{N}(\theta; \mu, \Sigma)$

$$p(f(\mathbf{x}^*) \mid \mathbf{x}^*, \mathcal{D}) = \mathcal{N}(f(\mathbf{x}^*); \mu^\intercal \mathbf{x}^*, \mathbf{x}^{*\intercal} \Sigma \mathbf{x}^*)$$

Nice things about the Bayesian perspective:

- MLE and MAP as special cases
- Capture all plausible solutions*
- Be explicit about the assumptions:
  - Gaussian independent measurement noise
  - Gaussian prior over parameters
- Can be adapted to handle other priors/likelihoods



*Model mis-specification is an issue

Probabilities

An example

Outcome of a coin flip, O, is random

There are two possible outcomes: head (H) or tail (T)

Tail Head

Questions we may want to ask:

- What is the probability of getting a head?
- What is the probability of getting a tail?
- Is it a fair coin?
- If we flip the coin many many times and we get one dollar for a head and zero for a tail, how much money will we make?
- What happens if instead we get 2 dollars per head and lose 25 dollars per tail?

Random variables - Continuous [1]

Many physical measurements or parameters in ML can take any value in a continuous range.

For D-dimensional continuous random variables X [each dimension is a random variable], we assoc

$$f: \mathbb{R}^D \to \mathbb{R} \forall \mathbf{x} \in \mathbb{R}^D: f(\mathbf{x}) \geq 0 \, and \int_{\mathbb{R}^D} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$$

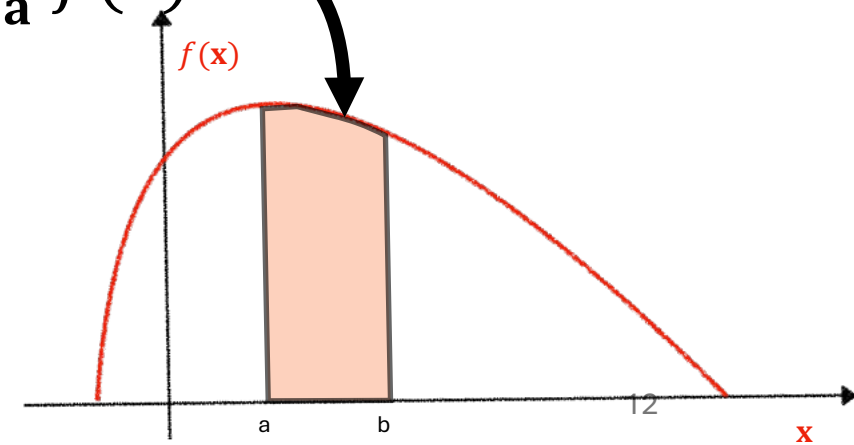
Reminder: in the discrete case [slide 11], the pmf satisfies: $0 \leq P(X = a) \leq 1$ and $\sum_a P(X = a) = 1$.

We say X is distributed according to $f(\mathbf{x})$ if $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x}) \mathrm{d}\mathbf{x}$

Note: $P(\mathbf{X} = \mathbf{a}) = \int_{\mathbf{a}}^{\mathbf{a}} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$

Hand-wavy: we can measure X between a and a + δ, but can never say X = a exactly

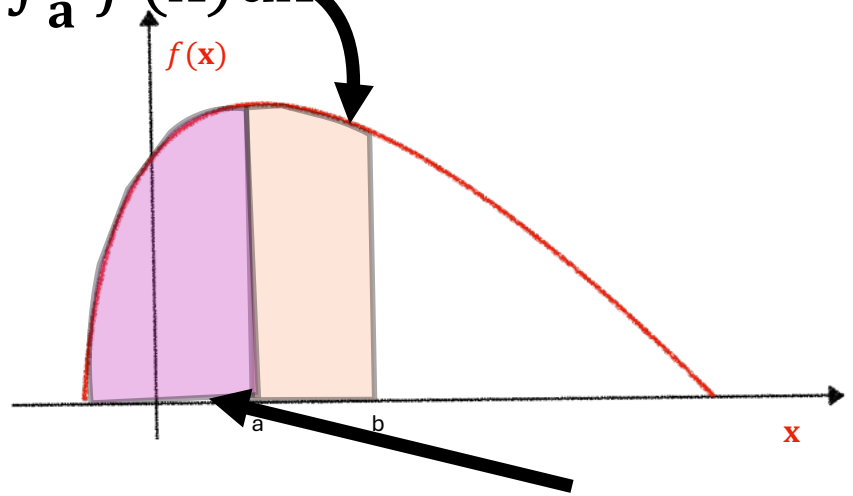
We often write p(x) instead of f(x). For continuous r.v., p(x) is not probability!



Random variables - Continuous [2]

We say X is distributed according to $f(\mathbf{x})$ if $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x})\mathrm{d}\mathbf{x}$

Area under f(x) between a a



Area under f(z) on the left of x

We can also find $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{z})\mathrm{d}\mathbf{z}$

This is called the cumulative distribution function (cdf).

Note that: $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{-\infty}^{\mathbf{b}} f(\mathbf{z})\mathrm{d}\mathbf{z} - \int_{-\infty}^{\mathbf{a}} f(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})$

and $f(\mathbf{x}) = \frac{d}{d\mathbf{x}} F(\mathbf{x})$

Example: Find $f(x)$, given $F(x) = \begin{cases} 0 & if x < 1 \\ 2(x-1) & if 1 \leq x \leq 1.5 \\ 1 & if x > 1.5 \end{cases}$

Distributions: discrete vs continuous

Discrete:

Probability mass function (pmf): $\forall a: 0 \leq P(X = a) \leq 1 \;\; and \sum_a P(X = a) = 1$

Sum rule $P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$ Product rule $P(X, Y) = P(X|Y)P(Y)$

$= P(Y|X)P(X)$

Continuous:

Probability density function (pdf): $\forall \mathbf{x} \in \mathbb{R}^D: f(\mathbf{x}) \geq 0 and \int_{\mathbb{R}^D} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$

Cumulative density function (cdf): $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{z}) \mathrm{d}\mathbf{z}$

Sum rule $f(\mathbf{x}) = \int_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$ Product rule $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}|\mathbf{y}) f(\mathbf{y}) = f(\mathbf{y}|\mathbf{x}) f(\mathbf{x})$

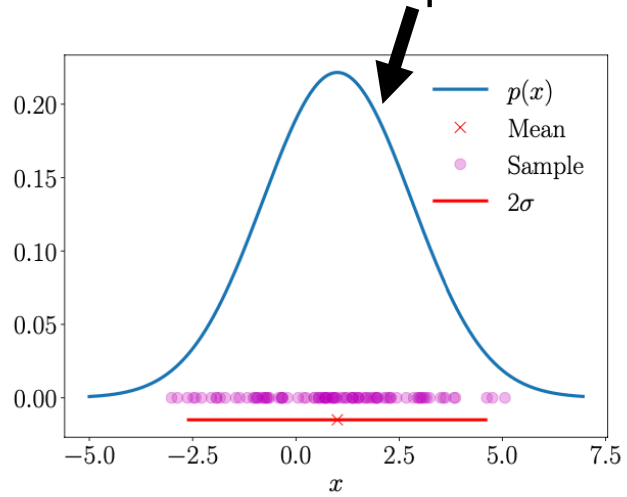
Note: f(x) can be larger than 1. Example: the uniform example in the last slide

Gaussian distributions - univariate

A continuous, real-valued, univariate Gaussian or normal random variable has the following probability density function

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

bell shaped curve



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.

$mean = \mu$, $variance = \sigma^2$, $standard\ deviation = \sigma$

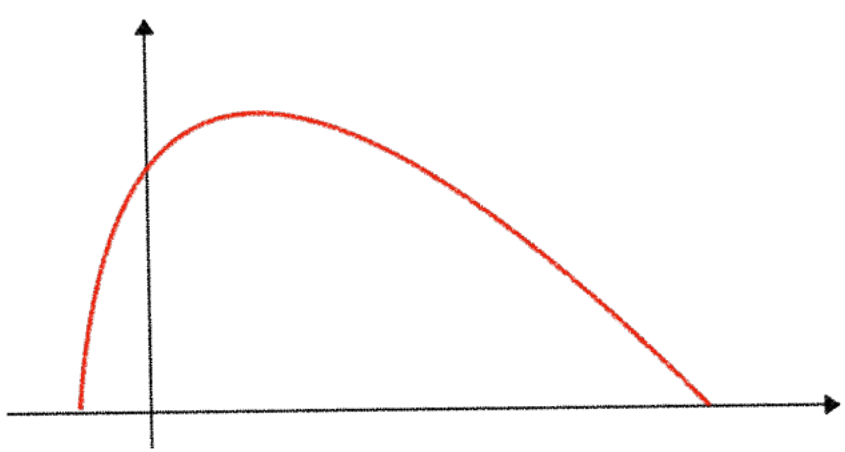
We often write: $p(x|\mu,\sigma^2) = \mathcal{N}(x;\mu,\sigma^2) or x \sim \mathcal{N}(\mu,\sigma^2)$

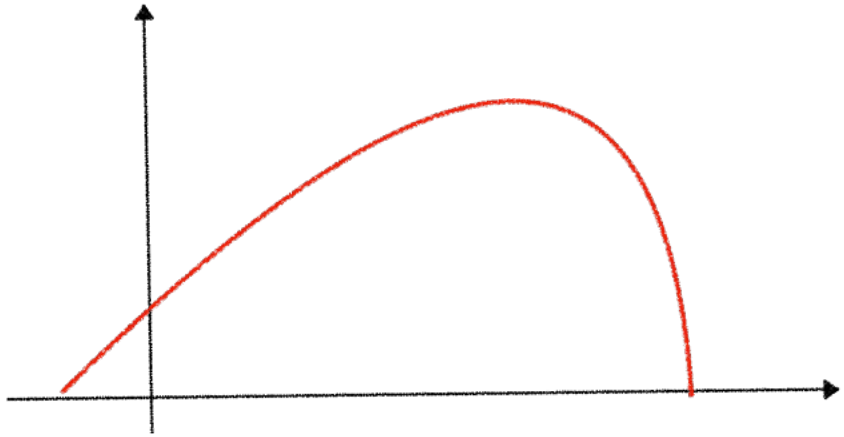
Standard normal distribution: $\mu = 0\ and\ \sigma = 1$

Cumulative probability distribution = probability to the left of x $\mathrm{cdf}(x)$

$= \int_{-\infty}^{x} p(z)\mathrm{d}z = \Phi\left(\frac{x-\mu}{\sigma}\right)$ where Φ is the Gauss error function (erf)

Comparing Distributions





- Entropy and cross entropy

Two fundamental rules - Sum and Product

Terminology:

- P(X, Y) is the joint distribution of two random variables X and Y
- P(X), P(Y) are the marginal distributions
- P(Y | X) is the conditional distribution of Y given X
- P(X | Y) is the conditional distribution of X given Y

Sum rule

$$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$$

where $\mathcal{Y}$ is the outcome space of Y. Also called marginalisation property.

Product rule

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Check worked example in previous slides

Bayes' rule - inverse probability

Product rule: $$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

We can rewrite: $P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y|X)}{P(Y)}$ or $P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(Y)P(X|Y)}{P(X)}$

This is called Bayes' rule or Bayes' theorem.

Posterior
belief about X after knowing Y $=$ $\frac{\text{Prior: prior belief about X} \times \text{Likelihood: likelihood of X given Y}}{\text{Marginal likelihood or evidence}}$

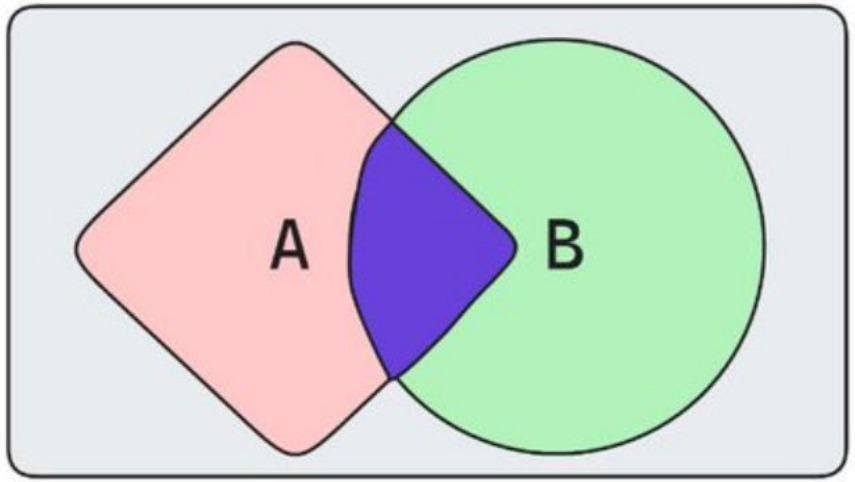
likelihood averaged over all potential X

Machine learning model + Bayes = probabilistic/Bayesian machine learning

X is the model or model parameters [unknown], Y = data [observed]

Learning = computing the posterior; model/hyperparameter selection using evidence;

Prediction = summing over all plausible models/model parameters



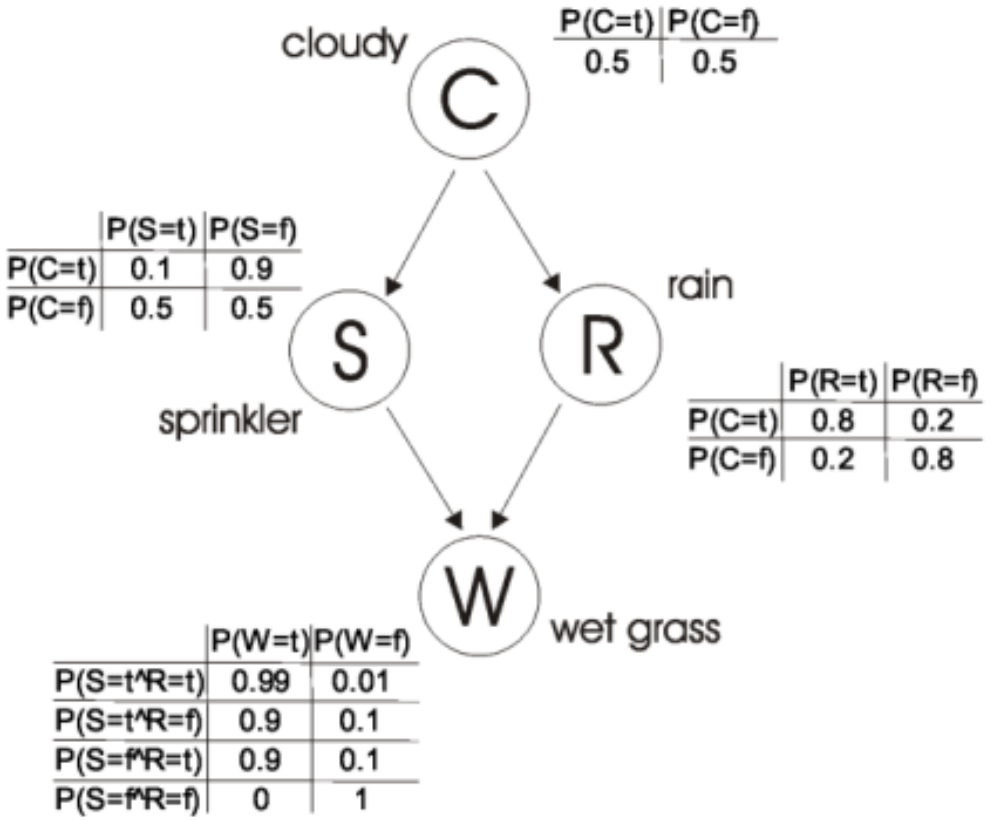
$$P(A) = \frac{\diamond}{\square} \qquad P(B) = \frac{\bigcirc}{\square} \qquad P(A|B) = \frac{\text{◗}}{\bigcirc} \qquad P(B|A) = \frac{\text{◗}}{\diamond}$$

$$\frac{\text{◗}}{\bigcirc} = P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{\frac{\text{◗}}{\diamond} * \frac{\diamond}{\square}}{\frac{\bigcirc}{\square}} = \frac{\text{◗}}{\bigcirc}$$

@akshay_pachaar, X

Bayesian Network

- A graph that accurately represents the joint probability distributions of a large number of variables.



Bayesian Network: Application

A patient has been suffering from shortness of breath (called **dyspnoea**) and visits the doctor, worried that he has lung **cancer**. The doctor knows that other diseases, such as tuberculosis and bronchitis, are possible causes, as well as lung cancer. She also knows that other relevant information includes whether or not the patient is a **smoker** (increasing the chances of cancer and bronchitis) and what sort of **air pollution** he has been exposed to. A **positive X-ray** would indicate either TB or lung cancer.[a]

Problems it solves

Assume the patient has dyspnoea, is not a smoker, and has not been exposed to high pollution. If the x-ray comes positive, what is the probability that the patient has cancer?

Applications

This problem is know as medical diagnostic: we have several possible causes of observed effects (symptoms).

Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}

Applications

This problem is know as medical diagnostic: we have several possible causes of observed effects (symptoms).

Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}

$$P(cancer = true \mid dyspnoea = true, \\ X\text{-}ray = positive, \\ smoker = false, \\ pollution = low).$$

Applications

$P(cancer = t \mid dyspnoea = t, X\text{-}ray = pos, smoker = f, pollution = low)$ can be estimated from data as:

$$\frac{\text{number of cases } (cancer = t, dyspnoea = t, X\text{-}ray = pos, smoker = f, pollution = low)}{\text{number of cases } (dyspnoea = t, X\text{-}ray = pos, smoker = f, pollution = low)}$$

Applications

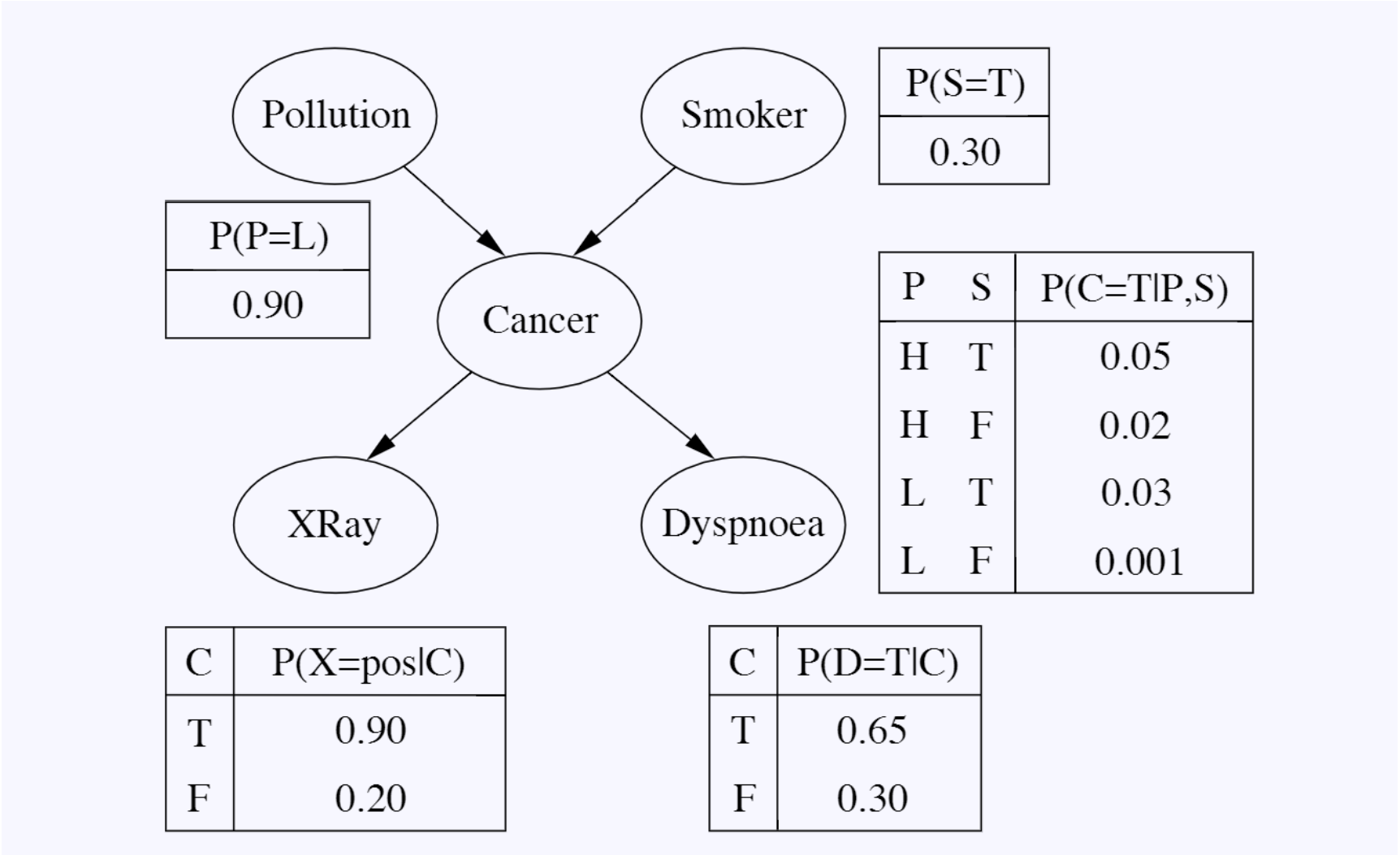
However, such an estimation will be hard to obtain accurately given the large number of parameters defining each instance.

- For example, assume there are no cases with non-smokers and low pollution, but a lot of cases with non-smokers and high pollution and a lot of cases with smokers and and low pollution.
- This should not be a problem if we know in advance that there is no strong causal relation between smoking and air pollution.

Applications

- Instead of estimating $P(cancer \mid \text{everything else})$ (which is difficult), we model simpler relations such as $P(cancer \mid smoker, pollution)$ and $P(X\text{-}ray \mid cancer)$.
- These simpler relations are derived from the causal relations between different variables: smoking causes cancer, cancer causes positive X-rays.
- These simpler relations are easier to estimate from data.
- From the marginal probabilities of these simpler relations, one can compute the joint probability distribution $P(cancer = t \mid dyspnoea = t, X\text{-}ray = pos, smoker = f, pollution = low)$.
- The joint probability distribution can be used to compute other interesting probabilities, such as $P(Dyspnoea = pos \mid cancer)$.

Joint Probability Distributions



P(S=T)
0.30

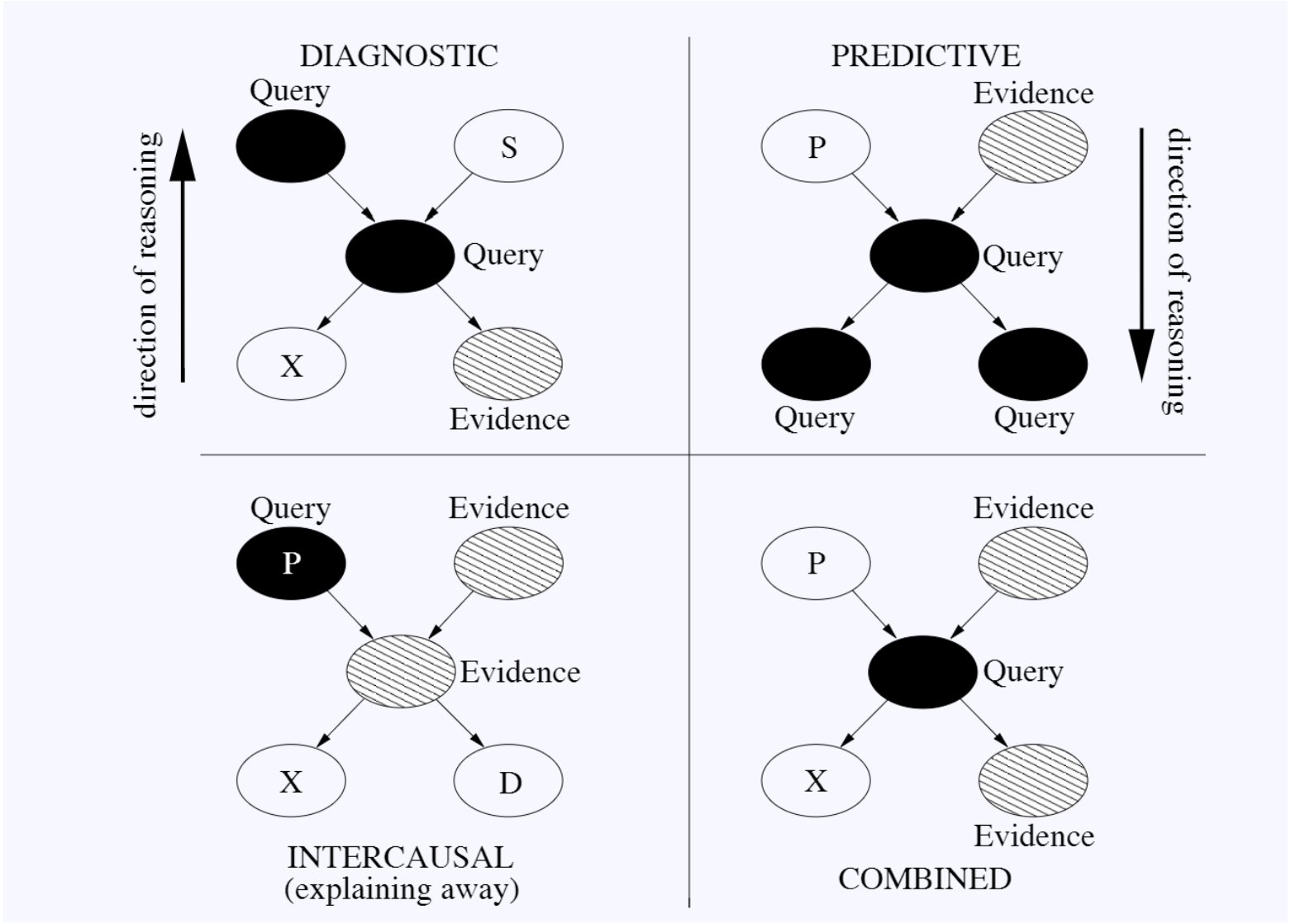
P(P=L)
0.90

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

Types of Reasoning



Bayesian Networks

Definition

A Bayesian network is a Directed Acyclic Graph (DAG) where

- Each node corresponds to a random variable.
- If there is an arrow from node X to node Y, X is said to be a parent of Y.
- Each node X_i has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node.

What is the meaning of the links?

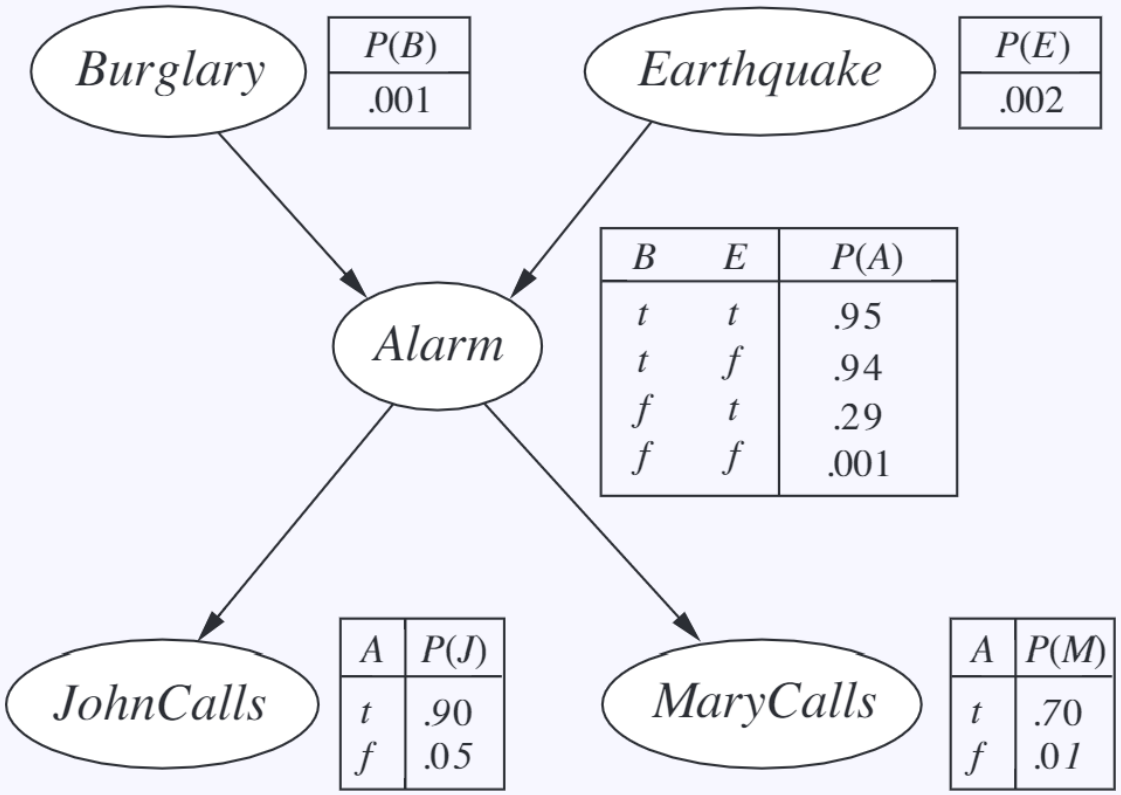
The links in a Bayesian network reflect causal relations between variables.

Bayesian Networks

Why do we use Bayesian networks?

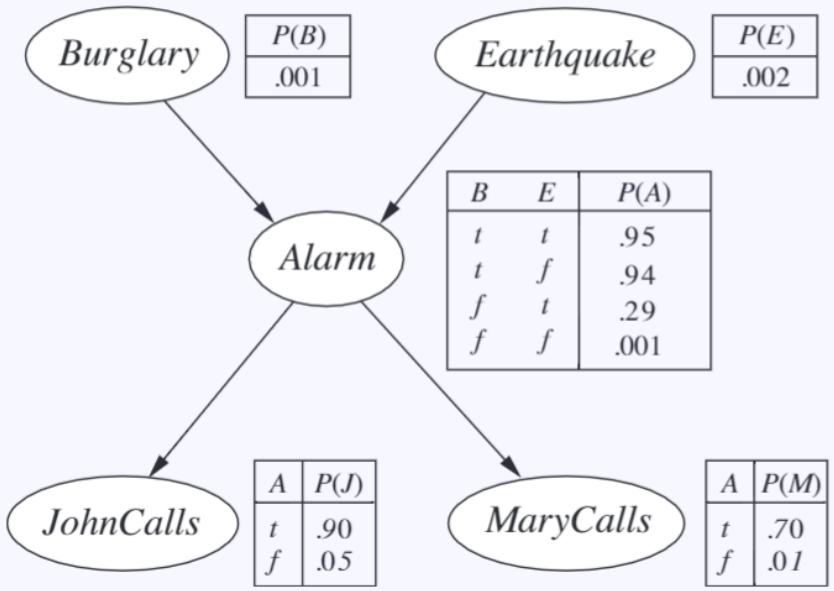
- The conditional probabilities represented in the links involve a smaller number of variables, which makes them easier to estimate and interpret.
- Inference is faster compared to when we use full joint distributions.
- Compact representation of full joint distributions.

Application



Bayesian network with conditional probability tables (CPTs).
B, E, A, J, and M stand for *Burglary, Earthquake, Alarm, JohnCalls*, and *MaryCalls* , respectively.

Application



- Each row in a Conditional Probability Table (CPT) should sum to one.
- For Boolean variables, once you know that the probability of a true value is p, the probability of false must be $1 - p$. We omit the second value in the table.
- The CPT of a variable with k parents has 2^k rows, if the parents are Boolean.

Deriving Value from Knowledge

Let $\{X_1, \ldots, X_n\}$ be the variables of a Bayesian network, and let $\theta(X_i \mid Parents(X_i))$ be the values in the corresponding CPTs. If values $\theta(X_i \mid Parents(X_i))$ are non-negative and each row in the CPTs sums to one, and

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i)),$$

then $\theta(x_i \mid parents(X_i))$ are exactly the conditional probabilities $P(x_i \mid parents(X_i))$. Hence, we can write:

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i)),$$

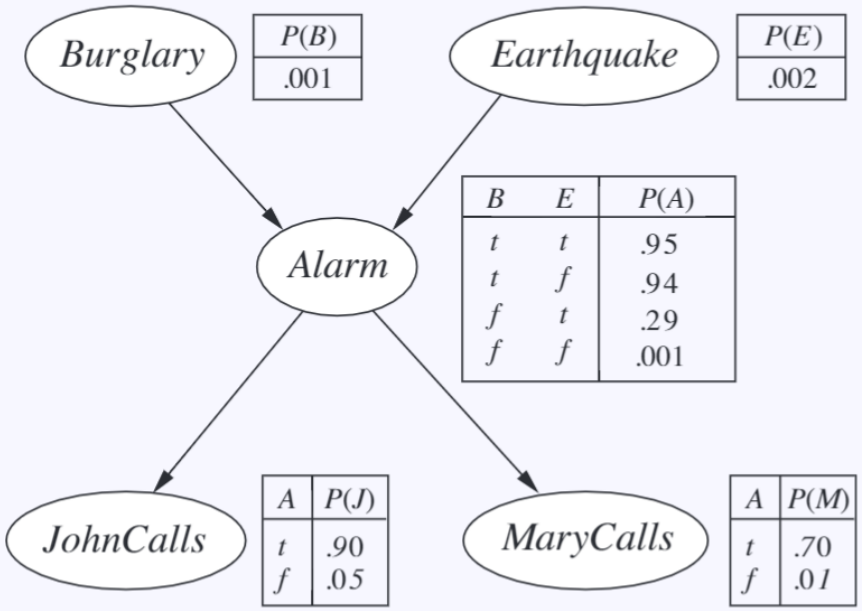
Deriving Value from Knowledge

The full joint distribution

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i)),$$

can be used to answer any query about the domain.

Deriving Value from Knowledge



Probability that both John and Mary call, the alarm has sounded, but neither a burglary nor an earthquake has occurred:

$$\begin{aligned} P(j, m, a, \neg b, \neg e) &= P(j|a)P(m|a)P(a|\neg b \wedge \neg e)P(\neg b)P(\neg e) \\ &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628. \end{aligned}$$

Independence Relations in Bayesian Networks

Given the definition of a Bayesian Network (directed acyclic graph + CPTs), i.e.

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i)),$$

we can prove two types of independence:

1. Each variable is conditionally independent of its non-descendants, given its parents.
2. A variable is conditionally independent of all other variables in the network, given its parents, children, and children's parents (Markov blanket).