Probability Theory

Jonathan Lawry

Department of Engineering Mathematics
University of Bristol

Axioms of Probability Theory

- For probability measures we denote the uncertainty measure $\mu = P$.
- The axioms of probability measures are then given by:
- **P1:** $P(W) = 1$ and $P(\emptyset) = 0$
- **P2:** If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
- Clearly then probability measures are simply *additive uncertainty measures*.
- From **P1** and **P2** the following well known properties follow:
- **General Additivity** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Complement** $P(A^c) = 1 - P(A)$
- Proofs...?

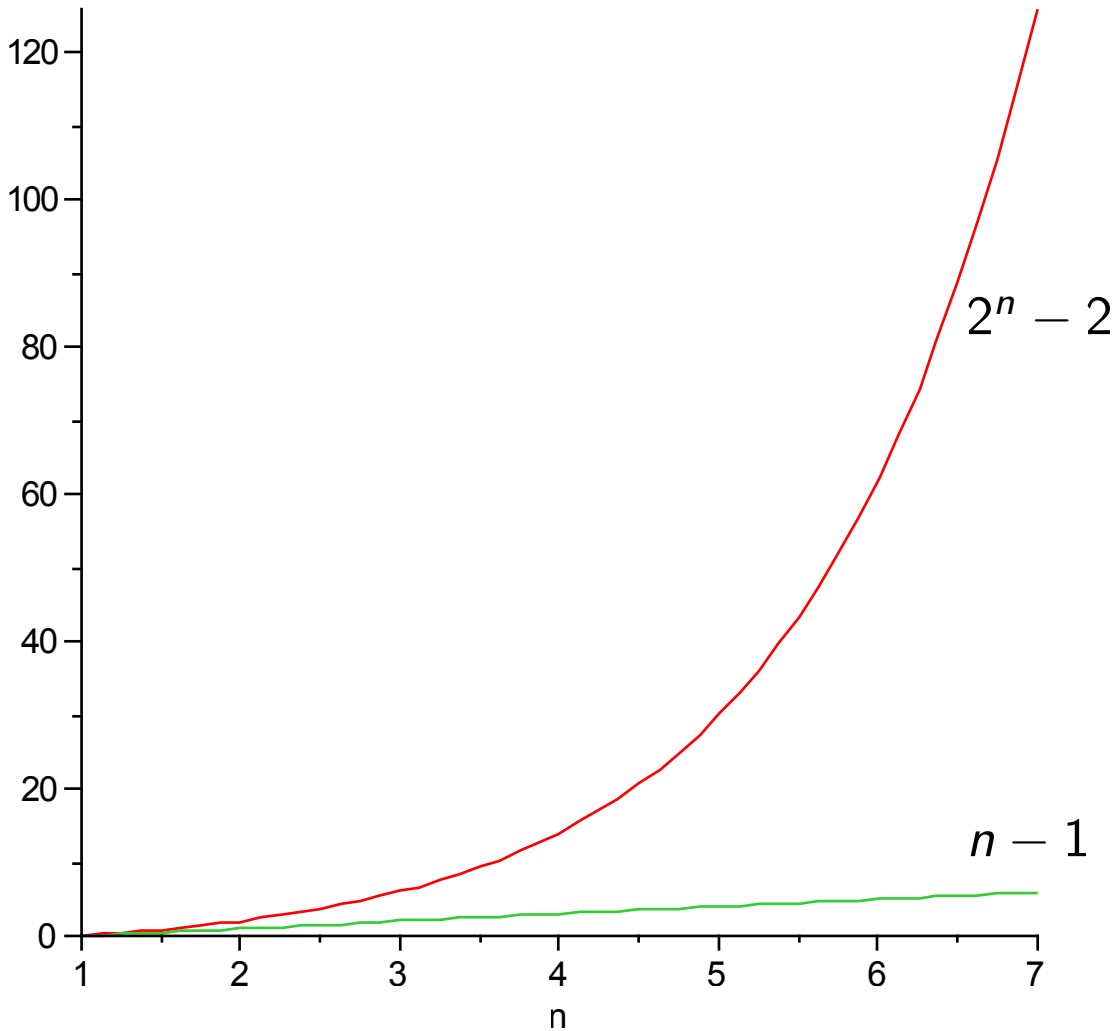
Probability Distributions

- Suppose $A = \{w_1, \ldots, w_k\}$ so that $A = \{w_1\} \cup \ldots \cup \{w_k\}$
- By Additivity **P2** it follows that $P(A) = P(\{w_1\}) + \ldots + P(\{w_k\})$
- We can define a *probability distribution* $p : W \to [0, 1]$ where $p(w_i) = P(\{w_i\})$ so that:
- $\sum_{w \in W} p(w) = 1$ and for any set $A \subseteq W$ we have that $P(A) = \sum_{w \in A} p(w)$.
- Hence, the probability measure P is uniquely defined by the probability distribution p.
- This comes directly from the assumption of additivity.

Information Requirements

- To define an uncertainty measure in general requires the agent to define $\mu(A)$ for every subset A of W, with the exceptions of W itself and $\emptyset$ which must have belief values 1 and 0 respectively.
- If W has n elements then the agent needs to specify $2^n - 2$ values.
- To define a probability measure the agent needs only define a probability distribution and therefore only needs to specify $n - 1$ values (since the probability distribution must sum to 1).

Information Requirements: 2



Updating Probabilities

- Suppose an agent has defined a probability measure P on W, for the lecture attendance problem.
- For suppose they then learn that 'all the female students are present'. How should they then update their probability measure?
- If it is certainly true that 'all the female students are present' then all possible worlds not in the set $\{\langle 0,1,1,0\rangle, \langle 1,1,1,0\rangle, \langle 0,1,1,1\rangle, \langle 1,1,1,1\rangle\}$ should be given probability 0.
- Since there is no other information available to the agent the remaining possible worlds (i.e. those in the above set) are given probabilities proportional to the original measure P but renormalised so as to sum to 1.

Conditional Probability

- For $A, B \subseteq W$ the conditional probability of A given B is defined by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Notice that if $P(B) = 0$ then $P(A|B)$ is undefined. Some theories of probability avoid this problem by taking conditional probabilities as primitives.
- E.g. the conditional probability that 'at least one male student is present' given that 'all the female students are present' is:

$$\frac{P(\{\langle 1,1,1,0\rangle, \langle 0,1,1,1\rangle, \langle 1,1,1,1\rangle\})}{P(\{\langle 0,1,1,0\rangle, \langle 1,1,1,0\rangle, \langle 0,1,1,1\rangle, \langle 1,1,1,1\rangle\})}$$

Bayes Theorem

- Suppose that we have discovered that a component in a jet engine has failed and we hypothesise that it is due to a fault of *type d*.
- Let E = component has failed and H = fault of *type d*.
- We require $P(H|E)$, but we only know the following:
- An earlier study of components with fault *type d* estimates that 80% will fail i.e. $P(E|H) = 0.8$.
- This compares to a failure rate of only 40% amongst those components without fault *type d* i.e. $P(E|H^c) = 0.4$
- From this information we can apply Bayes Theorem to obtain $P(H|E)$.

Bayes Theorem: 2

- **Bayes Theorem** follows trivially from the definition of conditional probability:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(E|H)P(H)}{P(E)}$$

- $P(E|H)$ corresponding to the probability of the evidence given the hypothesis is called the *likelihood*.
- $P(H)$ is called a *prior* probability (more to follow on this).
- $P(E)$ can be determined from the *total theorem of probability*.

Theorem of Total Probability

- Notice that for any sets $A, B \subseteq W$, $A = (A \cap B) \cup (A \cap B^c)$.
- Also notice that $(A \cap B) \cap (A \cap B^c)$ so by additivity $P(A) = P(A \cap B) + P(A \cap B^c)$.
- By the definition of conditional probability $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B^c) = P(A|B^c)P(B^c)$.
- This gives us the theorem of total probability:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

- Or in Bayes Theorem: $P(E) = P(E|H)P(H) + P(E|H^c)P(H^c)$

Bayes Theorem: 3

- For the jet engine example we have by the theorem of total probability that:
  $P(E) = 0.8P(H) + 0.4(1 - P(H)) = 0.4(P(H) + 1)$
- Hence,

$$P(H|E) = \frac{0.8P(H)}{0.4(P(H) + 1)} = \frac{2P(H)}{P(H) + 1}$$

- So we need to know $P(H)$ corresponding to the prior probability of the hypothesis.
- But on what basis can such prior probabilities be determined?

Prior Probabilites

- In order to apply the conditional probability formula the agent must already have defined a probability measure (i.e. a *prior*).
- Assuming the conditional probability formula is the only mechanism available to the agent inorder to update probabilities on the basis of new information, then at some point the agent must specify a probability measure on the basis of *no information*.
- **Laplace's principle of insufficient reason:** In the absence of any other information all possible worlds should be assumed to be equally probable i.e. the probability distribution should be *uniform*.

Transformation Invariance

- Sometimes assuming a uniform prior gives you different answers if you transform the problem.
- Let $W = \{w_1, w_2, w_3\}$ and X be a random variable with values in $\{1, 2\}$.
- Suppose we have no information concerning the distribution of p or the definition of X.
- By the principle of insufficient reason we should take $p(w_1) = p(w_2) = p(w_3) = \frac{1}{3}$.
- Also, applying the principle to X we have $P(X = 1) = P(X = 2) = \frac{1}{2}$.
- But $P(X = 1) = P(\{w : X(w) = 1\}) = 0$ or $= \frac{1}{3}$ or $= \frac{2}{3}$ or $= 1$.

Geometric Representation

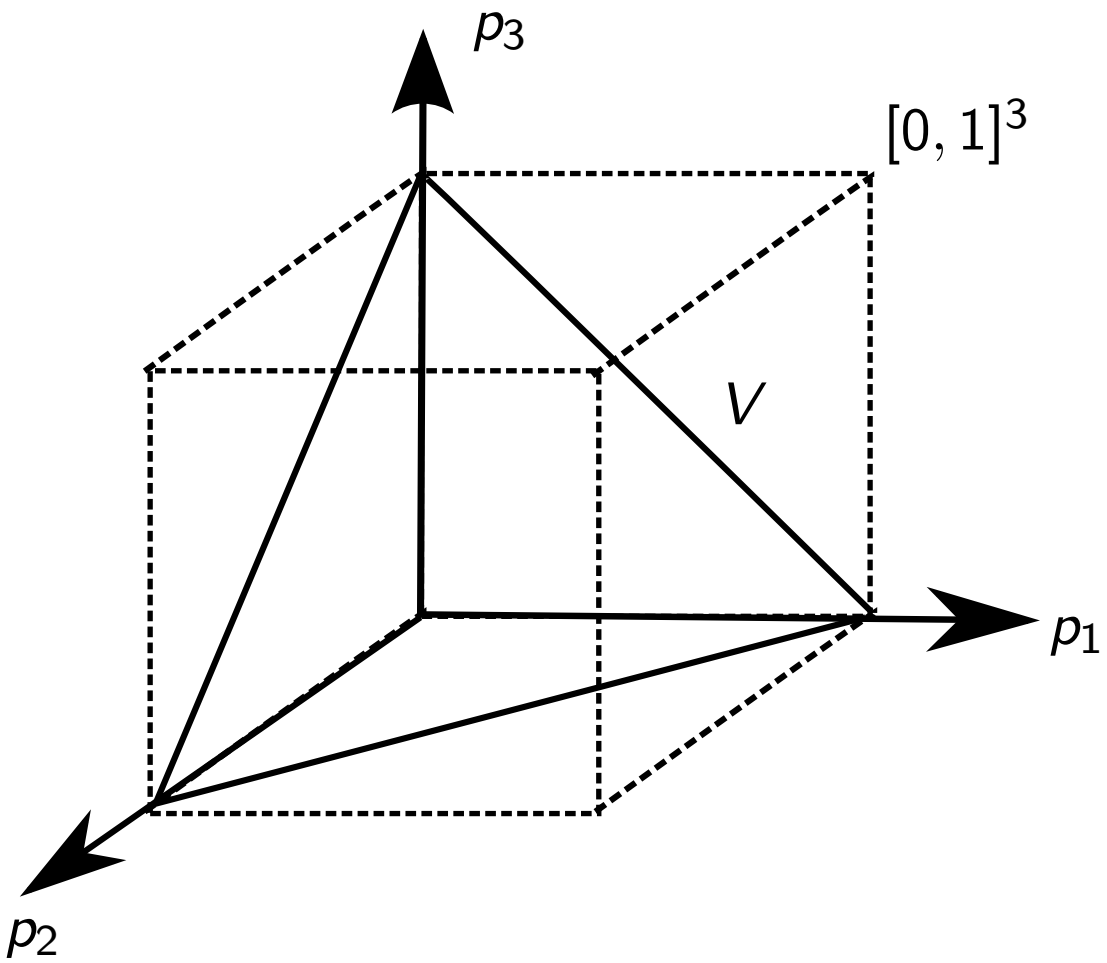
- Let $W = \{w_1, \ldots, w_n\}$ then a probability measure on 2^W can be represented as a vector $\langle p_1, \ldots, p_n \rangle$ where $p(w_i) = p_i$.
- Then the set of all possible distributions corresponds to

$$V = \{\vec{p} \in [0,1]^n : \sum_{i=1}^{n} p_i = 1\}$$

- V is an $n-1$ convex polytope of $[0,1]^n$
- Convex means that if $\vec{p}, \vec{q} \in V$ then $\lambda\vec{p} + (1-\lambda)\vec{q} \in V$ for any $\lambda \in [0,1]$.
- The centre of mass of V is $\langle \frac{1}{n}, \ldots, \frac{1}{n} \rangle$.

Geometric Representation:2

- Let $W = \{w_1, w_2, w_3\}$ so that
  $V = \{\langle p_1, p_2, p_3 \rangle \in [0,1]^3 : p_1 + p_2 + p_3 = 1\}$



Justification for Probability

- Is it possible to identify compeling reasons why a *rational* agent should use a belief function satisfying properties P1 and P2. (i.e. can we justify why uncertainty measures should be probability measures)?
- Here we consider two possible approaches:
- The physicist R.T Cox proposed a set of commmon sense properties that an agent's measure of belief should satisfy and then proved that any such measure is *isomorphic* to (a rescaling of) a probability measure.
- Emerging from the work of de Finetti, Ramsey, Kemeny and Shimony is the idea that an agent's level of belief in a proposition should be identified with their willingness to bet on this proposition.

Cox's Justification

- Cox proposes that an agent's measure of belief should satisfy the following informal properties:
- *The plausibility of a statement is a real number and is dependent on information we have related to the statement.*
- *Plausibilities should vary sensibly with the assessment of plausibilities in the model.*
- *If the plausibility of a statement can be derived in many ways, all the results must be equal.*
- These postulates are two informal, however, to prove the required theorem and hence Cox's formalizes them as follows:

Cox's Postulates

- **Cox1:** The agent defines a conditional measure $\mu(\bullet|\bullet) : 2^W \times 2^W \to [0,1]$ so that $\mu(A|B)$ gives the measure of belief in A given that B holds.
- **Cox2:** If $A \neq \emptyset$ then $\mu(\emptyset|A) = 0$ and $\mu(A|A) = 1$.
- **Cox3:** If $B \neq \emptyset$ then $\mu(A^c|B) = S(\mu(A|B))$ where $S : [0,1] \to [0,1]$ is a decreasing function satisfying $\forall x, \forall y \in [0,1]$ $S(S(x)) = x$ and $y.S(\frac{x}{y}) = S(x)S(\frac{S(x)}{S(y)})$.
- **Cox4:** If $B \cap C \neq \emptyset$ then $\mu(A \cap B|C) = F(\mu(A|B \cap C), \mu(B|C))$ where $F : [0,1]^2 \to [0,1]$ is strictly increasing (in both coordinates), continuous on $(0,1]^2$ and associative (i.e $\forall x, y, z \in [0,1]$ $F(x, F(y,z)) = F(F(x,y),z))$

Cox's Theorem

- If **Cox1**,. . .,**Cox4** hold then there is a continuous, strictly increasing surjective function $g : [0, 1] \to [0, 1]$ such that $g(\mu(\bullet|W))$ satisfies **P1** and **P2**, and for $A, B \subseteq W$

$$g(\mu(A|B)) \times g(\mu(B|W)) = g(\mu(A \cap B|W))$$

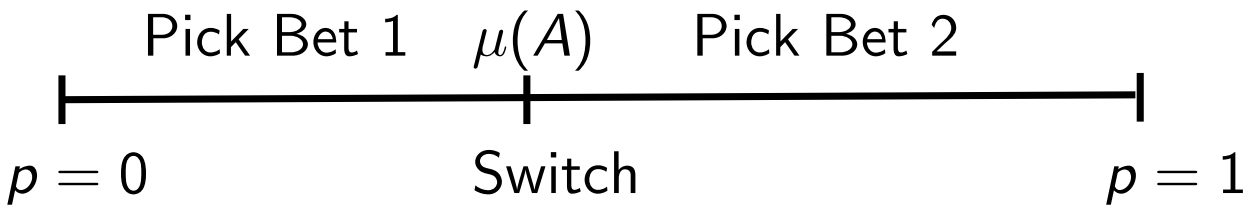
- This version of the theorem is actually due to Paris and Halpern who detected problems with Cox's original proof.
- For example, the functional equations required for S and the associativity of F where not explicitly required by Cox.

Cox's Postulates Considered

- **Cox3** states that the agent's belief in a proposition should increase has his/her belief in the negation of the proposition decreases.
- Cox justifies **Cox4** by using the following example of a runner:
- Suppose there is a runner with properties C. Then the agent's belief that he can run to a distant place B and return A, should only be a function (F) of the belief that he will get there $\mu(B|C)$ and that having got there he will return $\mu(A|B \cap C)$.
- The other properties of S and F are harder to justify except for the *idempotence* of S which allows for double negation i.e. $\mu((A^c)^c|B) = \mu(A|B)$.

Betting Justification

- Let $\chi_A : W \to \{0, 1\}$ be the membership (characteristic) function of set $A \subseteq W$.
- Now consider two bets defined for $0 \leq p \leq 1$, $S > 0$ and $A \subseteq W$ as follows:
- **Bet 1:** Gain $S(1 - p)$ if A is true and lose Sp if A is false.
- **Bet 2:** Lose $S(1 - p)$ if A is true and gain Sp if A is false.



Rational Belief

- If $p \leq \mu(A)$ then the agent picks bet 1 and his/her gain is $S(1-p)\chi_A(w^*) - Sp(1-\chi_A(w^*)) = S(\chi_A(w^*) - p)$
- If $p > \mu(A)$ then the agent picks bet 2 and his/her gain is $Sp(1-\chi_A(w^*)) - S(1-p)\chi_A(w^*) = -S(\chi_A(w^*) - p)$
- We say that μ is *rational* if there do not exist $S_i, T_j > 0$, $A_i, B_j \subseteq W$, $p_i < \mu(A_i)$, $q_j > \mu(B_j)$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$ such that for all possible worlds $w \in W$

$$\sum_{i=1}^{n} S_i(\chi_{A_i}(w) - p_i) - \sum_{j=1}^{m} T_j(\chi_{B_j}(w) - q_j) < 0$$

- If μ is rational in the above sense then μ satisfies P1 and P2 and hence is a probability measure.

Joint Probability Distributions

- Given random variables $X_1, \ldots, X_n$ their joint distribution is given by:

$$P(X_1 = x_1, \ldots, X_n = x_n) =$$

$$P(\{w : X_1(w) = x_1, \ldots, X_n(w) = x_n\})$$

- If X_i has k_i values then specifying the joint distribution requires $(\prod_{i=1}^{n} k_i) - 1$ values.
- For example if X_i are all binary variables then we must specify $2^n - 1$ values.

Marginal Distributions

- Marginal distributions of each variable can be calculated from the joint distribution according to:

$$P(X_1 = x_1) =$$

$$\sum_{x_2 \in \Omega_2} \sum_{x_3 \in \Omega_3} \cdots \sum_{x_n \in \Omega_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- **Conditional Distributions:** The conditional distribution of X_1 given X_2 is defined by:

$$P(X_1 = x_1 | X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)}$$

Independence

- Random variable X_1 is independent of X_2 if $P(X_1 = x_1 | X_2 = x_2) = P(X_1 = x_1)$
- Random variables $X_1, \ldots, X_n$ are independent if $P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$
- In this case the number of values which must be specified in order to define the joint distribution is $\sum_{i=1}^{n}(k_i - 1)$
- Suppose that $X_1, \ldots, X_n$ are binary variables so that $k_i = 2$ for $i = 1, \ldots, n$
- Then the independent case requires n values to be specified where as the fully dependent case requires $2^n - 1$.

Conditional Independence

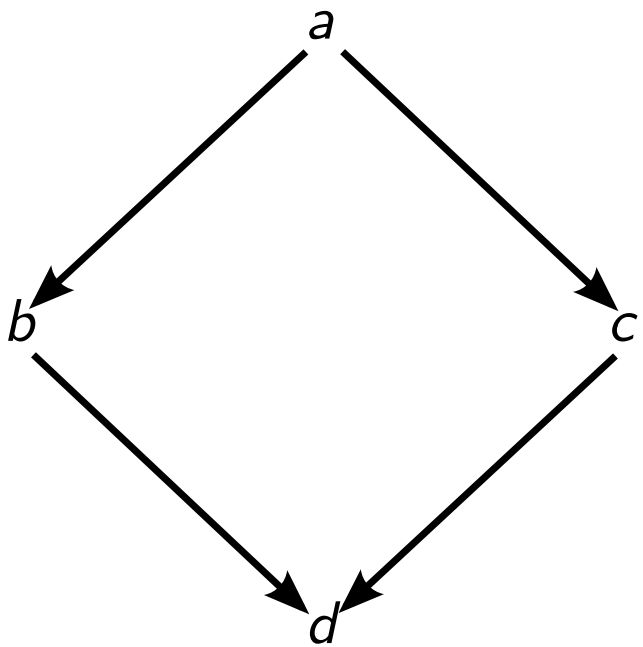
- A full independence assumption considerably simplifies probabilistic reasoning but is often unrealistic.
- Let U, V, W be exclusive subsets of $\{X_1, \ldots, X_n\}$. Then the variables in U are said to be conditionally independent of the variables in V given the variables in W if $P(U|V, W) = P(U|W)$
- For example, X_1 and X_2 are conditional independent of X_3 given X_4 and X_5 if $P(X_1, X_2|X_3, X_4, X_5) = P(X_1, X_2|X_4, X_5)$
- If we know that values of X_4 and X_5 then knowing the value of X_3 gives no additional information about the values of X_1 and X_2.

Conditional Independence:2

- **Example:** Let X_1 =‘the last bus is due to arrive at 12pm’, X_2 =‘the last bus has not yet arrived’, X_3 =‘I will be able to catch the last bus’
- X_3 is conditionally independent of X_1 given X_2
- Because...
- The time the last bus is due to arrive is not relevent to whether or not I will be able to catch it given that I know if it has not arrived yet.
- The notion of conditional independence is used in *graphical models* of probability to provide computationally feasible representations of joint probability distributions.

A Little Graph Theory

- **Directed Graph:** A directed graph is an ordered pair (V, E) where V is a set of vertices (nodes) and E is a binary relation on V encoding the edges.
- A directed graph (V, E) is acyclic if there is no sequence of nodes $v_1, \ldots, v_n$ where $v_1 = v_n$ and $(v_i, v_{i+1}) \in E$ for $i = 1, \ldots, n-1$
- **Example:** $V = \{a, b, c, d\}$ and $E = \{(a, b), (a, c), (b, d), (c, d)\}$



Bayesian Networks

- Bayesian networks are graphical models that utilise the notion of conditional independence to provide a compromise between the extreme cases of total dependence and total independence of random variables.
- This is achieved by assuming independence when *possible* and dependence when *necessary*.
- A number of *causal relationships* between the random variables are specified.
- We then distinguish between direct and indirect causes.

Formal Definition

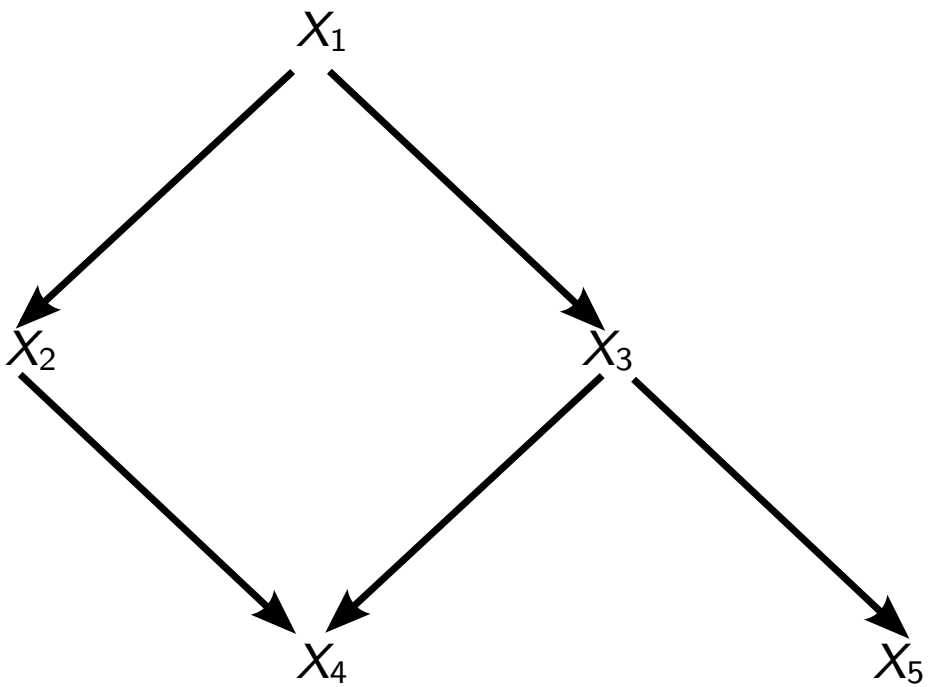
- A bayesian network is a directed graph (V, E) where $V = \{X_1, \ldots, X_n\}$ enumerated such that $(X_j, X_i) \in E$ only if $j < i$ together with a probability distribution on V satisfying:

$$P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|\Pi(X_i))$$

- Where $\Pi(X_i) = \{X_j : (X_j, X_i) \in E\}$ are the *parents* of X_i
- For a bayesian network assume that X_i is conditionally independent of its indirect causes $\{X_1, \ldots, X_{i-1}\} - \Pi(X_i)$ given its direct causes $\Pi(X_i)$.i.e $P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|\Pi(X_i))$

Example Network

- Let $X_1, \ldots, X_5$ be binary random variables into $\{0, 1\}$ such that $X_1 = 1$ iff metastatic cancer is present, $X_2 = 1$ iff total serum calcium increased, X_3=1 iff brain tumour present, $X_4 = 1$ coma present, $X_5 = 1$ iff papilledema present.



Evaluating Distributions

- The Chain Rule: $P(X_1, \ldots, X_n) =$

$$P(X_1)\frac{P(X_1, X_2)}{P(X_1)}\frac{P(X_1, X_2, X_3)}{P(X_1, X_2)}\cdots\frac{P(X_1, \ldots, X_n)}{P(X_1, \ldots, X_{n-1})}$$
$$= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_n|X_1, \ldots, X_{n-1})$$

- Hence, for a Bayesian network:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\Pi(X_i))$$

Complexity

- For each random variable X_i we must store the conditional distribution $P(X_i|\Pi(X_i))$
- This requires that the following number of probability values be specified: $(k_i - 1) \times (\prod_{X_j \in \Pi(X_i)} k_j)$
- Therefore, the total number of values that must be specified for the network is $\sum_{i=1}^{n}(k_i - 1) \times (\prod_{X_j \in \Pi(X_i)} k_j)$
- If X_i are binary random variables then the number of values required is $2^{|\Pi(X_i)|}$
- This lies between n and $2^n - 1$ (Proof?)
- **Example:** For the cancer network: Totally dependent solution requires $2^5 - 1 = 32$ values.
- Bayesian network solution requires $1 + 2 + 2 + 2^2 + 2 = 11$ values.