ST301 Revision Notes: Theorems, Definitions, and Remarks

Marco Del Vecchio *

May 13, 2017

*M.del-Vecchio@warwick.ac.uk

Contents

1 Fundamentals

Definition 1 (Loss Function)**.** *A loss function $L(d,\theta)$, $d \in D, \theta \in \Theta$, maps the event $\{(d,\theta)|d \in D, \theta \in \Theta\}$, according to which we take decision d and the outcome is θ onto a real number intuitively representing the loss associated with the event.*

Definition 2 (Pay-off Function)**.** *We call $R(d,\theta) = -L(d,\theta)$ a pay-off function.*

Definition 3 (Probability Function)**.** *Denote by $p(\theta|d,x)$ the posterior distribution of an outcome $\theta \in \Theta$ given the that D.M. has taken decision $d \in D$ and observed some data $x \in \mathcal{X}$.*

Definition 4 (Expected Monetary Value(E.M.V.))**.** *Given $p(\theta|d,x), \theta \in \Theta, d \in D, x \in \mathcal{X}$, the expected monetary value of taking decision $d \in D$ is defined as*

$$\bar{f}(d) = \mathbb{E}_{p(\theta|d,x)}(f(d,\theta)) = \begin{cases} \sum_{\theta\in\Theta} f(d,\theta)p(\theta|d,x) & \textit{discrete case} \\ \int_{\theta\in\Theta} f(d,\theta)p(\theta|d,x)d\theta & \textit{continuous case} \end{cases}$$

where $f \in \{L,R\}$. Further, we say that a decision maker (D.M.) follows the E.M.V. strategy if he chooses a decision $d^* \in D$ such that $\bar{L}(d)$ is minimised or equivalently such that $\bar{R}(d)$ is maximised.

Definition 5 (Bayes Decision)**.** *We call a decision $d^* \in D$ which solves either*

$$\arg\min_{d\in D} \bar{L}(d)$$

or

$$\arg\max_{d\in D} \bar{R}(d)$$

a Bayes decision

Theorem 1 (Bayes' Theorem)**.** *Let $p(\cdot)$ denote a probability function. Then,*

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)},$$

provided $p(y) > 0$.

Definition 6 (Conditional Independence)**.** *The components of a random vector $\boldsymbol{X} = (X_1,\ldots,X_m)$ are said to be conditionally independent with respect to a random variable Y written*

$$\coprod_{j=1}^{m} X_j|Y$$

if and only if for each i, $1 \leq i \leq m$

$$\mathbb{P}(\boldsymbol{X} = x|Y) = \prod \mathbb{P}(X_i = x_i|Y)$$

Definition 7 (Naive Bayes Assumption)**.** *Let $\boldsymbol{X} = (X_1,\ldots,X_m)$ be a random vector and Y a random variable. Then, the assumption that*

$$\mathbb{P}(\boldsymbol{X} = x|Y) = \prod \mathbb{P}(X_i = x_i|Y)$$

is called the Naive Bayes Assumption.

2 Decision Trees

Definition 8 (Decision Tree)**.** *A decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. We distinguish between three more kinds of nodes:*

- *decision node: a node indicating that a decision has to be made by the D.M. whose outgoing edges indicate the possible decisions.*
- *chance node: a node indicating the draw of a random variable.*
- *leaf node: the last node of a branch indicating the outcome of following that branch of the tree.*

Definition 9 ((Proper) Terminal Pay-off)**.** *A leaf node in a decision tree indicates the pay-off (outcome) associated with following the branch which terminates with said leaf node. Depending on whether the leaf node takes into account the cost of taking decisions, we refer to it as proper terminal pay-off (if it does) or terminal pay-off (if it does not).*

Definition 10 (Expected Value Of Perfect Information (E.V.P.I.))**.** *We define the expected value of perfect information as*

$$\mathbb{E}_{p(\theta|d,x)}\Big(\max_{d\in D} R(d,\theta)\Big|\theta\Big) - \max_{d\in D^0}\mathbb{E}_{p(\theta|d,x)}(R(d,\theta))$$

where D^0 represents the set of decisions which cost the least and give us the least amount of information. That is, the E.V.P.I. is the difference between the maximum expected pay-off under perfect information and the maximum expected pay-off, under uncertainty, obtained by following the E.M.V strategy but without experimenting.

Remark 1 (On the E.V.P.I.)**.** *The version of the E.V.P.I. given in Definition 10 is characterised by having as baseline the decision which costs the least and provides the D.M. with the minimum amount of information. However, no one stops from changing this baseline. In fact, for instance, it is possible to define the E.V.P.I. as*

$$\mathbb{E}_{p(\theta|d,x)}\Big(\max_{d\in D} R(d,\theta)\Big|\theta\Big) - \max_{d\in D}\mathbb{E}_{p(\theta|d,x)}(R(d,\theta)).$$

That is, as the difference between the maximum expected pay-off under perfect information and the the expected pay-off obtaining by choosing a Bayes Decision.

There is one substantial difference between using the two baselines: using the first baseline does not require the Bayes decision to be known, that is, it does not require us to solve the decision problem.

The handiness of the first baseline becomes obvious when one considers pruning a decision tree. In fact, we can use the definition of E.V.P.I. as given in Definition 10 as an upper bound on the price worth paying for any experimentation. Thus limiting the number of decisions worth considering.

3 Utility Theory

Definition 11 (Gamble)**.** *Let $\langle \boldsymbol{r}, \boldsymbol{p}\rangle$, $\boldsymbol{r} = (r_1 \ldots, r_n), \boldsymbol{p} = (p_i, \ldots, p_i)$ define a gamble where the probability of obtaining reward r_i is $p_i, 1 \leq i \leq n$ with $\sum_i^n p_i = 1$.*

Definition 12 (Certainty Money Equivalent (C.M.E)). *Given a gamble $\langle \boldsymbol{r}, \boldsymbol{\alpha} \rangle$ the C.M.E is the maximum amount of money that a gambler is ready to forfeit in preference to taking the betting scheme. Further, denote the C.M.E of the gamble*

$$B(r^*, r^0, \alpha) := \langle (r^*, r^0), (\alpha, (1-\alpha)) \rangle$$

to be $r(\alpha)$

Definition 13 (Utility Function). *$U(r)$ is said to be a utility function for an achievable reward r if for some $(a, b) \in \mathbb{R}^2, b > 0$,*

$$U(r) = a + b\alpha(r)$$

where $\alpha(r) = r^{-1}(\alpha)$

Further, let x^0 and x^* be the best and worst outcome of attribute $x_i, 1 \leq i \leq n$ respectively. Then, define $\boldsymbol{x}^0$ and $\boldsymbol{x}^*$ as

$$\boldsymbol{x}^0 = (x_1^0, \dots x_n^0)$$
$$\boldsymbol{x}^* = (x_1^*, \dots x_n^*).$$

We say that a utility function U is normalised if

$$U(\boldsymbol{x}^0) = 0, \text{ and } U(\boldsymbol{x}^*) = 1.$$

Axiom 1 (A1). *Gambles giving the same distribution of rewards should be considered equivalent.*

Axiom 2 (A2). *If P_1, P_2, and P are any three gambles then for all $\alpha, 0 < \alpha < 1$*

$$P_1 \preceq P_2 \implies \alpha P_1 + (1-\alpha)P \preceq \alpha P_2 + (1-\alpha)P$$

Axiom 3 (A3). *If P_1, P_2, and P are any three gambles such that $P_1 \preceq P \preceq P_2$ then there exists $\alpha, \beta, 0 < \alpha, \beta < 1$ such that*

$$P \prec \alpha P_2 + (1-\alpha)P_1$$
$$P \succ \beta P_2 + (1-\beta)P_1$$

Theorem 2. *If Axioms A1-A3 hold then there exists a real valued utility function U such that for any rewards r^0, r^*, r satisfying*

$$r \sim \alpha r^* + (1-\alpha)r^0$$

then

$$U(r) = \alpha U(r^*) + (1-\alpha)U(r^0),$$

and for every r such that $r^0 \prec r \prec r^*$,

$$\alpha(r) = \frac{U(r) - U(r^0)}{U(r^*) - U(r^0)}$$

Axiom 4 (A4). *Let P be the gamble $\langle \boldsymbol{r}, \boldsymbol{p} \rangle$ and let $p(r)$ be the probability mass function over the rewards further, assume that there exists r^0 and r^* such that for each $r \in \boldsymbol{r}$, $r^0 \preceq r \preceq r^*$. Then, for each $r \in \boldsymbol{r}$ let*

$$\alpha(r) = \frac{U(r) - U(r^0)}{U(r^*) - U(r^0)}$$

then,

$$P \sim \beta r^* + (1-\beta) r^0$$

where

$$\beta = \sum_{r \in \boldsymbol{r}} \alpha(r) p(r).$$

Remark 2 (On A4)**.** *One implication of Axiom four, is that for any gamble $\langle \boldsymbol{r}, \boldsymbol{p} \rangle$ for which there exists r^0 and r^* in $\boldsymbol{r}$ such that for each $r \in \boldsymbol{r}$, $r^0 \preceq r \preceq r^*$, we can find a two-point gable $\langle (r^*, r^0), (\beta, 1-\beta) \rangle$ equivalent to $\langle \boldsymbol{r}, \boldsymbol{p} \rangle$, where β is as defined above.*

Theorem 3. *If Axioms A1-A4 hold then there exists a real valued utility function U such that, given any two gables P_1 and P_2,*

$$\mathbb{E}(U(P_1)) \geq \mathbb{E}(U(P_2)) \Leftrightarrow P_1 \succeq P_2$$

Remark 3 (Eliciting A Utility Function With The Mid-Point Method)**.** *Let $r(0)$ and $r(1)$ be the lowest and highest reward respectively such that $U(r(0)) = 0$ and $U(r(1)) = 1$. Then elicit the following values:*

$$r(1/2) \sim \frac{1}{2} r(0) + \frac{1}{2} r(1)$$
$$r(1/4) \sim \frac{1}{2} r(0) + \frac{1}{2} r(1/2), \quad r(3/4) \sim \frac{1}{2} r(1/2) + \frac{1}{2} r(1)$$

and so on continuing to divide up the reward space so that $U(r(k/2^n)) = k/2^n$ for any positive integer n. Further, since a utility function is increasing in r for $r \in [r(k/2^n), r((k+1)/2^n)]$, it follows that

$$|U(r) - U_n(r)| \leq \frac{1}{2^n}$$

and therefore that

$$|\bar{U}(r) - \bar{U}_n(r)| \leq \frac{1}{2^n}$$

where $U_n(r)$ is the linear interpolation on the points $\{r(k/2^n)\}_{k=1}^{2^n}$. That is, both the elicited utility function and its expected value can be made arbitrarily good.

Definition 14 (Current Wealth Independent)**.** *A D.M. is said to exhibit a current wealth independent behaviour if for a rewards r and current wealth $h > 0$,*

$$r \sim \alpha(h)(r+h) + (1-\alpha(h))(r-h) \implies U(r) = \alpha(h) U(r+h) + (1-\alpha(h)) U(r-h)$$

Definition 15 (Mutually Utility Independent Attributes (M.U.I.A.))**.** *A utility function U is said to have m.u.i. attributes $C(x) = \{x_1, \ldots, x_n\}$ if for every non-empty subset $A(x) \subset \{x_1, \ldots, x_n\}$ and $B(x) \subset \{x_1, \ldots, x_n\}$ such that $A(x) \cup B(x) = C(x)$ and $A(x) \cap B(x) = \emptyset$,*

$$U(x) = a_A(B(x)) + k_A(B(x)) U_A(A(x))$$

with $k_A > 0$. That is, U can be written as a function whose arguments are in $B(x)$, a_A, plus a function whose arguments are in $B(x)$, k_A, times a utility function U_A which only depends on $A(x)$.

Definition 16 (Criterion Weights)**.** *Given a normalised utility function with n attribute, let its criterion weights $k_i, 1 \leq i \leq n$ be defined as*

$$k_i = U(x_1^0, \ldots, x_{i-1}^0, x_i^*, x_{i+1}^0, \ldots, x_n^0).$$

Theorem 4. *If a utility function has m.u.i attributes, the criterion weights associated to it are known, and it has more than three attributes then,*

$$U(x) = \begin{cases} \sum_{i=1}^n k_i U_i(x_i), & \text{if } \sum_{i=1}^n k_i = 1 \\ \frac{1}{k}\big(\prod_{i=1}^n (1 + k k_i U_i(x_i)) - 1\big), & \text{if } \sum_{i=1}^n k_i \neq 1 \end{cases}$$

where k is such that

$$1 + k = \prod_{i=1}^n (1 + k k_i)$$

Remarks:

- *$U_i(x_i)$ is the i^{th} marginal utility of U.*
- *When $\sum_{i=1}^n k_i > 1$ the D.M. prefers to obtain a high utility score on few attributes and a low utility score on the remaining ones rather than obtaining a moderate utility score on all attributes.*
- *When $\sum_{i=1}^n k_i < 1$ the D.M. prefers to obtain a moderate utility score on all the attributes rather than obtaining a high utility score on few attributes and a low utility score on the remaining ones.*

Definition 17 (Value Independent Attributes)**.** *A utility function with attribute vector $\boldsymbol{x}$ is said to have value independent attributes if two distributions of rewards are equally preferred whenever they have identical marginal distributions to each other on all individual attributes.*

Theorem 5. *A utility function has value independent attributes if and only if its criterion weights sum up to one, that is if*

$$\sum_{i=1}^n k_i = 1.$$

Further, a utility function which has value independent attributes is linear.

Remark 4 (Eliciting A Utility Function With M.U.I.A.)**.** *It is possible to elicit the criterion weights and the marginal utility functions as follows.*

- *Eliciting the criterion weights: Let*

$$T_i = (x_1^0, \ldots, x_{i-1}^0, x_i^*, x_{i+1}^0, \ldots, x_n^0), \text{ with probability 1}$$
$$T(\alpha) = \begin{cases} \boldsymbol{x}^*, & \text{with prbability } \alpha \\ \boldsymbol{x}^0, & \text{with prbability } 1 - \alpha \end{cases}$$

  Then, k_i is the value of α such that $T_i \sim T(\alpha)$.

- *Eliciting the marginal utility functions: Fix x_j $1 \leq j \neq i \leq n$ to $\bar{x}_j$. Let*

$$T_i = (\bar{x}_1^0, \ldots, \bar{x}_{i-1}^0, x_i, \bar{x}_{i+1}^0, \ldots, \bar{x}_n^0), \text{with prbability } \alpha$$
$$T(\alpha) = \begin{cases} (\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i^*, \bar{x}_{i+1}, \ldots, \bar{x}_n), & \text{with prbability } \alpha \\ (\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i^0, \bar{x}_{i+1}, \ldots, \bar{x}_n), & \text{with prbability } 1 - \alpha \end{cases}$$

  Then, $U_i(x_i)$ is the value of α such that $T_i \sim T(\alpha)$.

4 Extensive and Normal Form Analysis Of A Decision Problem

Definition 18 (Decision Rule)**.** *A decision rule prescribes a unique decision at each point in the decision process at which an action needs to be taken, and it is specified as a function of the information available at that time.*

Definition 19 (Extensive Form Analysis)**.** *The extensive form analysis of a decision problem consists of transforming prior beliefs into posterior ones via Baye's Theorem and use those fixed probabilities alongside the terminal pay-offs to remove bus-optimal branches to identify the Bayes decision rule.*

An extensive from analysis is done in steps:

1. *(Optional) Prune the decision tree using E.V.P.I..*
2. *Use Bayes theorem to calculate posterior beliefs.*
3. *Determine the Bayes decision rule(s) by maximising the expected utility of the D.M..*

Definition 20 (Normal Form Analysis)**.** *The normal form analysis of a decision problem consists of identifying possible Bayes decision rules as a function of algebraic/unknown probabilities.*

A normal from analysis is done in three steps :

1. *Identify and label the set all decision rules in the problem.*
2. *Identify the event(s) $(E_1, \ldots, E_k)$ whose probability vector $p(\boldsymbol{\theta}) = (p(\theta_1), \ldots, p(\theta_k))$ on states $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ (assuming a discrete scenario), is of interest, and evaluate the expected utility $V_i(d_j)$ associated with each possible decision rule d_j for each event $E_i, 1 \leq i \leq k$.*
3. *Identify from $\{\boldsymbol{V}(d_j) = (V_1(d_j), \ldots, V_k(d_j)) | d_j$ being a decision rule$\}$ a subset of decision rules which could be Bayes decision rules for some $p(\boldsymbol{\theta})$.*

Remark 5 (Step 3 of the Normal Form Analysis for $k = 2$)**.** *If k=2 then, a decision rule d_j is strictly preferred to a decision rule d_i, with $j \neq i$ whenever the expected payoff associated with d_j is higher then the one associated with d_i under some fixed probability vector $p(\boldsymbol{\theta})$. That is, when*

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{\theta})}(\boldsymbol{V}(d_j)) > \mathbb{E}_{p(\boldsymbol{\theta})}(\boldsymbol{V}(d_i)) &\implies \\ \sum_{l=1}^{k} V_l(d_j)p(\theta_l) > \sum_{l=1}^{k} V_l(d_i)p(\theta_l) &\implies \\ \bar{G}(d_j) > \bar{G}(d_i). & \end{aligned}$$

where $\bar{G}(d_s) = \sum_{l=1}^{k} V_l(d_s)p(\theta_l)$, $s \in \{i, j\}$. When $k = 2$, this reduces to saying that a decision rule d_j is strictly preferred to a decision rule d_i, with $j \neq i$ under a fixed $p(\theta_1)$ whenever

$$\begin{aligned} V_1(d_j)p(\theta_1) + V_2(d_j)(1 - p(\theta_1)) > V_1(d_i)p(\theta_1) + V_2(d_i)(1 - p(\theta_1)) &\implies \\ -\frac{p(\theta_1)}{1 - p(\theta_1)} < \frac{V_2(d_j) - V_2(d_i)}{V_1(d_j) - V_1(d_i)} & \end{aligned}$$

In turn, by plotting the values of $V_i(d_j)$, $1 \leq i \leq k$ on a 2-d plot, we can see that a decision d_j is a Bayes decision in the sense that it maximises $\bar{G}(d_j)$ for some probability distribution over $\boldsymbol{\theta} = (\theta_1, \theta_2)$ whenever it lies on the North-East boundary of the convex hull formed by the points in (V_1, V_2)-space.

5 Sensitivity And Probability

Definition 21 (Empirically Well Calibrated Forecaster)**.** *We say that a forecaster is empirically well calibrated on a set of n predictions if the total number of times the forecaster quoted the probability of an event happening, q, the proportion of times the event has happened, $\hat{q}$, is equal to q. This being true for all unique values of q that the forecaster quotes.*

Definition 22 (Scoring Rule)**.** *A loss function $L(a,q), a \in \{0,1\}, 0 \leq q \leq 1$, is called a scoring rule if it is used to penalise bad probability forecasts q.*

Definition 23 (Proper Scoring Rule)**.** *A scoring rule $L(a,q)$ which minimises the expected loss related to the scoring rule*

$$\bar{L}(q|p) = pL(1,q) + (1-p)L(0,q)$$

when $q = p$, and the D.M. follows the E.M.V. strategy, is called a (strictly) proper scoring rule.

Definition 24 (Brier Score)**.** *The Brier score is defined as*

$$L(a,q) = (a-q)^2.$$

Further, the Brier score is a proper scoring rule.

Definition 25 (Logarithmic Score)**.** *The Logarithmic score is defined as*

$$L(a,q) = \begin{cases} -log(q) & a = 1 \\ -log(1-q) & a = 0 \end{cases}$$

Further, the Logarithmic score is a proper scoring rule.

Definition 26 (Empirical Score)**.** *Let $\{(a_i, q_i) | 1 \leq i \leq n\}$ denote n pairs of outcome and probability predictions, then the forecaster's empirical score S_n is given by*

$$S_n = \sum_{i=1}^{n} L(a_i, q_i)$$

when $q = p$, is called a (strictly) proper scoring rule.

Theorem 6. *If the proper scoring used is $L(a,q) = (a-q)^2$, known as Brier score, the forecaster's empirical score will be at least as low is he replaces his quoted probabilities q_i with $\hat{q}_i$ where $\hat{q}_i$ are the observed probabilities (sample proportions).*

Proof. Suppose a forecaster quotes m probabilities $q_1 \ldots, q_m$ such that

$$0 \leq q_1 < q_2 < \ldots, < q_m \leq 1,$$

where q_i is quoted $n_i > 0$ times $1 \leq i \leq m$ such that

$$\sum_{i=1}^{m} n_i = n$$

where n is the total number of forecasts. Further, let $a_i(j)$ denote the outcome of the j^{th} period $1 \leq j \leq n_i$ for which the forecaster quoted a probability q_i, $1 \leq i \leq m$.

Finally, denote the proportion of times that the event for which the probability q_i, $1 \leq i \leq m$ was quoted has happened as $\hat{q}_i$ so that

$$\hat{q}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} a_i(j).$$

Then, fi we are using the Brier score has our scoring rule, the empirical score is given by

$$S_n(\boldsymbol{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (a_i(j) - q_i)^2$$

where

$$\begin{aligned}
\sum_{j=1}^{n_i} (a_i(j) - q_i)^2 &= \sum_{j=1}^{n_i} (a_i(j) - \hat{q}_i + \hat{q}_i - q_i)^2 \\
&= \sum_{j=1}^{n_i} (a_i(j) - \hat{q}_i)^2 + \sum_{j=1}^{n_i} (\hat{q}_i - q_i)^2 \\
&= \sum_{j=1}^{n_i} (a_i(j) - \hat{q}_i)^2 + n_i(\hat{q}_i - q_i)^2
\end{aligned}$$

since $2(\hat{q}_i - q_i) \sum_{j=1}^{n_i} (a_i(j) - \hat{q}_i) = 0$. Hence,

$$S_n(\boldsymbol{q}) = S_n(\hat{\boldsymbol{q}}) + n_i(\hat{q}_i - q_i)^2 \implies S_n(\boldsymbol{q}) \geq S_n(\hat{\boldsymbol{q}})$$

with equality when

$$n_i(\hat{q}_i - q_i)^2 = 0 \implies \hat{q}_i = q_i.$$

□

6 Bayesian Networks And Relevance

Definition 27 (Parent and Child)**.** *A vertex X is a parent of of a vertex Y and Y is a child of vertex X in a directional graph $\mathcal{G}$ if and only if there is a directed edge from X to Y in $\mathcal{G}$.*

Definition 28 (Ancestor)**.** *A vertex Z is an ancestor of a vertex Y in $\mathcal{G}$ if $Z = Y$ or if there exists a directed path in $\mathcal{G}$ from Z to Y.*

Definition 29 (Ancestral Set)**.** *Let X be a subset of vertices $V(\mathcal{G})$ of $\mathcal{G}$. The ancestral graph of X denoted $A(X)$ is the set of all vertices in $V(\mathcal{G})$ that are ancestors of a vertex in X.*

Definition 30 (Ancestral Graph)**.** *The ancestral graph $\mathcal{G}(A(X))$ has as vertex set $A(X)$ and as edge set $\{e = X_e \to Y_e \in E(\mathcal{G}) | X_e, Y_e \in A(X)\}$.*

Definition 31 (Mixed Graph)**.** *A graph is said to be mixed if it contains both directed and undirected edges.*

Definition 32 (Moralised Graph). *The moralised graph $\mathcal{G}^M$ of $\mathcal{G}$ has the same vertex set $\mathcal{G}$ but has an undirected edge between any two vertices $X_i, X_j \in V(\mathcal{G})$ when there is no directed edge between them in $\mathcal{G}$ bu they are parents of the same child in $\mathcal{G}$.*

Definition 33 (Decomposable Graph). *A graph $\mathcal{G}$ is said to be decomposable if the moralised graph of $\mathcal{G}$ is equal to $\mathcal{G}$.*

Definition 34 (Skeleton). *The skeleton $\mathcal{S}(\mathcal{H})$ of a mixed graph $\mathcal{H}$ is a graph with same vertex set as $\mathcal{H}$ and an undirected edge between X_i and X_j if and only if there is a directed edge between X_i and X_j in $\mathcal{H}$.*

Definition 35 (Pattern). *The pattern $\mathcal{P}$ of a DAG $\mathcal{G}$ is a mixed graph with the same vertices as $\mathcal{G}$ and a directed edge from X_i to Y replaced with an undirected edge from X_i to Y if and only if no other parent X_j of Y is not connected to or form X_i by an edge. That is, if and only if Y's parents are all married with each other.*

Definition 36 (Essential Graph). *The essential graph $\mathcal{E}$ derived from a pattern graph $\mathcal{P}$ is a graph such that an edge is directed if and only if changing its direction would lead to a different pattern graph.*

Definition 37 (Triangulated Graph). *A graph $\mathcal{G}$ is said to be triangulated if all the parents of a node have been married with a directed edge until $\mathcal{G}$ becomes decomposable.*

Definition 38 (Clique). *A clique of an undirected graph $\mathcal{H}$ is a subset of nodes in which every node is connected to every other node (maximally complete).*

Definition 39 (Running Intersection Property). *Let $\{C_j | j = 1, \ldots, m\}$ be the cliques of a decomposable DAG $\mathcal{G}$ and let the separators $\{B_j | j = 1, \ldots, m\}$ be defined as*

$$B_j = C_j \cap C^{j-1}$$

where $C^{j-1} = \cup_{i=1}^{j-1} C_i$ is the set of all components of C_j appearing in a clique listed before C_j. Then if,

$$B_j \subset C_{j_s}$$

for some index j_s such that $1 \leq j_s < j$ then we say that the cliques have the running intersection property.

Theorem 7. *A decomposable graph has cliques that can be indexed so that they exhibit the running intersection property.*

Definition 40 ((Pre-)Junction Tree). *A pre-junction tree $\mathcal{PJ}$ of a decomposable DAG $\mathcal{G}$ is a directed tree with vertices representing the cliques of $\mathcal{G}$ and a directed edge from C_i to C_j if C_i is the mother of C_j. A junction tree $\mathcal{P}$ of a decomposable DAG $\mathcal{G}$ is an undirected tree with vertices representing the cliques of $\mathcal{G}$ and an undirected edge from C_i to C_j if C_i is the mother of C_j.*

Definition 41 (Conditional Irrelevance). *We say a measurement X is irrelevant for predicting Y given the measurement Z if the D.M. believes that once he has learned about Z then X will provided no further information about Y.*

Definition 42 (Irrelevance Statement). *A statement involving a conditional irrelevance statement is called an irrelevance statement. An a valid irrelevance statement must satisfy three properties:*

1. *Functional Dependence:*

$$Y \perp\!\!\!\perp X|(Z,Y)$$

This defines what we mean by knowing something. Given Z and Y, learning about X will not give use any more information about Y because we know Y already.

2. *Symmetry:*

$$Y \perp\!\!\!\perp X|Z \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

3. *Perfect Composition: for any set of disjoint measurement X,Y,Z,W,*

$$X \perp\!\!\!\perp (Y,Z)|W \Leftrightarrow X \perp\!\!\!\perp Y|(W,Z) \text{ and } X \perp\!\!\!\perp Z|W$$

Theorem 8 (Conditional Independent Statement Via Factorisation)**.** *Let $X_1,\ldots,X_n$ be random variables such that their joint probability function is given by $p(X_1,\ldots,X_n) = p(\boldsymbol{X})$ then,*

$$p(\boldsymbol{X}) = p_1(X_1)p_2(X_2|X_1)\ldots p_n(X_n|X_1,\ldots,X_{n-1}).$$

Suppose that for $2 \leq i \leq n$, $p_i(X_i|X_1,\ldots,X_{i-1}) = P(X_i|\boldsymbol{X}_{Q_i})$ where $\boldsymbol{X}_{Q_i} \subseteq \{X_1,\ldots,X_{i-1}\}$. Now, let $\boldsymbol{X}_{R_i} = \{X_1,\ldots,X_{i-1}\} \setminus \boldsymbol{X}_{Q_i}$ then these factorisation imply for $2 \leq i \leq n$ that

$$X_i \perp\!\!\!\perp \boldsymbol{X}_{R_i}|\boldsymbol{X}_{Q_i}.$$

Definition 43 (Directed Acyclic Graph (DAG))**.** *A graph $\mathcal{G}$ associated with $n-1$ irrelevance statements that has vertices $V(\mathcal{G}) = \{X_1,\ldots,X_n\}$ and an edge from X_i to X_j if and only if $X_i \in \boldsymbol{X}_{Q_j}$, where $\boldsymbol{X}_{Q_j}$ are the parents of X_j and X_j is a child of each $X_i \in \boldsymbol{X}_{Q_j}$, is called a DAG.*

Definition 44 (Influence Diagram (ID))**.** *An Influence Diagram is a DAG together with the set of irrelevance statements that defines it.*

Definition 45 (Equivalent IDs)**.** *Two IDs are said to be equivalent if they imply the same set of irrelevance statements.*

Theorem 9. *Two IDs are equivalent if they have the same patter graph.*

Remark 6 (On the Essential Graph)**.** *Every undirected edge in an essential graph $\mathcal{E}$ is associated with two equivalent IDs whose DAGs differ in the direction of this edge.*

Definition 46 (Bayesian Network (BN))**.** *A Bayesian Network is an Influence Diagram together with the factorisation of the joint probability function generated from the irrelevance statements.*

Theorem 10. *If a DAG $\mathcal{G}^*$ is created form a valid DAG $\mathcal{G}$ by adding an edge from X_j to X_i, $1 \leq j < i \leq n$, then $\mathcal{G}^*$ is also a valid DAG.*

Proof. By construction, we have that

$$\boldsymbol{X}_{Q_i^*} = \{\boldsymbol{X}_{Q_i}, X_j\}$$
$$\boldsymbol{X}_{R_i} = \{\boldsymbol{X}_{R_i^*}, X_j\}$$

with $\boldsymbol{X}_{R_i^*} = \{X_1, \ldots, X_{i-1}\} \setminus \boldsymbol{X}_{Q_i^*}$. Then,

$$X_i \perp\!\!\!\perp \boldsymbol{X}_{R_i} | \boldsymbol{X}_{Q_i} \Leftrightarrow X_i \perp\!\!\!\perp \{\boldsymbol{X}_{R_i^*}, X_j\} | \boldsymbol{X}_{Q_i}$$

Since $\mathcal{G}$ is valid, by Perfect Composition we have that

$$X_i \perp\!\!\!\perp \{\boldsymbol{X}_{R_i^*}, X_j\} | \boldsymbol{X}_{Q_i} \implies X_i \perp\!\!\!\perp \boldsymbol{X}_{R_i^*} | \{\boldsymbol{X}_{Q_i}, X_j\} \Leftrightarrow X_i \perp\!\!\!\perp \boldsymbol{X}_{R_i^*} | \boldsymbol{X}_{Q_i^*}.$$

□

Theorem 11 (D-Separation Theorem)**.** *Let A, B, and C be any three disjoint subsets of $\{1, \ldots, n\}$ and $\mathcal{G}$ be a valid DAG where the vertices of $\mathcal{G}$ are $V(\mathcal{G}) = \{X_1, \ldots, X_n\}$. Then, if $\boldsymbol{X}_B$ separates $\boldsymbol{X}_C$ from $\boldsymbol{X}_A$ in the skeleton of the moralised ancestral graph of $\mathcal{G}(A(\boldsymbol{X}_{A \cup B \cup C}))$ then,*

$$\boldsymbol{X}_C \perp\!\!\!\perp \boldsymbol{X}_A | \boldsymbol{X}_B.$$

Theorem 12. *Let A, B, and C be any three disjoint subsets of $\{1, \ldots, n\}$. Then if $\boldsymbol{X}_B$ does not separate $\boldsymbol{X}_C$ from $\boldsymbol{X}_A$ in the skeleton of the moralised ancestral graph of $\mathcal{G}(A(\boldsymbol{X}_{A \cup B \cup C}))$ then there exists at least one BN where the assertion $\boldsymbol{X}_C \perp\!\!\!\perp \boldsymbol{X}_A | \boldsymbol{X}_B$ is false*

Theorem 13. *Let $\mathcal{G}$ be a valid DAG of random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$, let $\mathcal{G}^*$ be the skeleton of the moralised graph of $\mathcal{G}$ and assume that for all possible configurations of parents $\{\boldsymbol{X}_{Q_i} | 2 \leq i \leq n\}$ the probability function of $\boldsymbol{X}_{Q_i}$ is $p(\boldsymbol{X}_{Q_i})$, $2 \leq i \leq n$. Then, $p(\boldsymbol{x})$ can be expressed as the product of quotients of probabilities of subsets of the cliques $\{C_1, \ldots, C_m\}$ of $\mathcal{G}^*$.*

Proof. By definition, the factorisation of $p(\boldsymbol{x})$ defined by $\mathcal{G}$ is

$$p(\boldsymbol{x}) = p(x_1) \prod_{i=2}^{n} p(x_i | \boldsymbol{x}_{Q_i}) = \prod_{i=1}^{n} \frac{p(x_i, \boldsymbol{x}_{Q_i})}{p(\boldsymbol{x}_{Q_i})}$$

with $\boldsymbol{x}_{Q_1} = \emptyset$, and where $\boldsymbol{x}_{Q_i} \subseteq \{x_1, \ldots, x_{i-1}\}$, $2 \leq i \leq n$. Now, by definition, a clique of $\mathcal{G}^*$ contains $\{X_i, \boldsymbol{X}_{Q_i}\}$ if all the vertices in $\{X_i, \boldsymbol{X}_{Q_i}\}$ are connected to each other. Well, by definition X_i is connected to all $\boldsymbol{X}_{Q_i}$ because $\boldsymbol{X}_{Q_i}$ are the parents of X_i in $\mathcal{G}$ and therefore also in $\mathcal{G}^*$. Further, after the moralising $\mathcal{G}$, all the parents of X_i which where not married, have been joined by an undirected edge. This implies that all the vertices in $\{X_i, \boldsymbol{X}_{Q_i}\}$ and $\{\boldsymbol{X}_{Q_i}\}$ is connected to every other vertex and therefore $\{X_i, \boldsymbol{X}_{Q_i}\}$ and $\{\boldsymbol{X}_{Q_i}\}$ are cliques. This in turn implies that $p(\boldsymbol{x})$ can be written as a product of quotients of a subsets of cliques of $\mathcal{G}^*$. □

Remark 7 (Using A Junction Tree To Propagate Information)**.** *It is possible to use the junction tree of a DAG $\mathcal{G}$ associated to a Bayesian Network as follows. Suppose the learn the value of a vector $\mathcal{X}$ then, we can propagate the effects of this new information in three steps:*

1. *Update the probability functions of all the cliques where $\mathcal{X}$ appears using Bayes's Theorem.*
2. *Update the probability functions of the separators using Bayes's Theorem, and calculate the multipliers associated with each separator where a multiplier of a separator B_i is defined as*

$$\times_{B_i} = \frac{p_{new}(B_i)}{p_{old}(B_i)}$$

3. *Update the remaining cliques using the appropriate multipliers by calculating*

$$p_{new}(C_i) = \times_{B_k} p_{old}(C_i)$$

for some clique C_i *and separator* B_k.