Acting Under Uncertainty

- Agent’s knowledge provides a degree of belief ➔ Probability Theory
  - Probability Theory provides the degree to which an agent believes a statement given all plausible alternative situations that are indistinguishable from the current situation.
  - Probability Theory has the same ontological commitment as logic
    - Facts either hold or do not hold in the world
  - Probabilities depend on the evidence (precepts) the agent has acquired.

- Agents must have preferences towards the outcomes of different plans.
  - **Utility Theory** – Theory of preference based on an object’s utility.
    - Preference indicates that different agents might have different goals so while goals might seem misguided, they are not necessarily irrational.
  - **Decision Theory** – Combination of Probability and Utility Theory.
    - **Principle of Maximum Expected Utility** - An agent is rational iff it chooses the action that yields the highest expected utility (averaged over all outcomes).

- **Belief State** – a representation of probabilities of all possible actual states of the world.

Probability Theory

- **Language of Probability** – ascribes degrees of belief to propositions.
  - **Random Variable** – refers to part of the world with initial unknown status
  - **Domain** – the values a random variable can take on.
    - Boolean – true/false
    - Discrete – countable, mutually exclusive, and exhaustive.
    - Continuous – uncountable
  - **Atomic Event** – a specific complete specification of the world about which the agent is uncertain.
    - The set of atomic events is mutually exclusive and exhaustive ➔ forms a partition
    - Any atomic event entails true/false for every proposition.
    - Any proposition is equivalent to the disjunction of all atomic events that entail the proposition as true.

- **Prior Probability** P(x) – the degree of belief of proposition x in the absence of any evidence about other propositions.

- **Probability Distribution**, P(X) – the vector of probabilities ascribed to each of the possible states of proposition X.

- **Joint Probability Distribution**, P(X₁,X₂,…,Xₙ) – the probabilities of all combinations of values of variables X₁, X₂, ..., Xₙ.

- **Full Joint Probability Distribution** – includes the complete set of variables for the environment.

- **Probability Density Function**, p(x)dx – The probability of a continuous variable on the interval [x,x+dx] in the limit as dx ➔ 0.
• **Conditional Probability** \( P(x | y) \) – the degree of belief of proposition \( x \) given the evidence \( y \).

• **Product Rule:** \( P(X, Y) = P(X | Y)P(Y) \)

• **Kolmogrov’s Axioms**
  1. All probabilities are between 0 and 1:
     \[ 0 \leq P(a) \leq 1 \]
  2. True propositions have probability 1; false propositions have 0:
     \[ P(\text{true}) = 1 \quad P(\text{false}) = 0 \]
  3. Probability of a disjunction is given by
     \[ P(a \lor b) = P(a) + P(b) - P(a \land b) \]

• Using these Axioms we can derive other important rules:
  o Let a discrete variable \( X \) have a domain \( \{x_1, x_2, \ldots, x_n\} \) or \( \chi \):
    \[ \sum_{i=1}^{n} P(X = x_i) = 1 \quad \text{or} \quad \int_{\chi} P(x) \, dx = 1 \]
  o The probability of a proposition is equal to the sum of the probabilities in which it holds; the set \( e(a) \) for proposition \( a \).
    \[ P(a) = \sum_{e, e(a)} P(e_i) \]

• In probability, statements do not refer directly to the world, but rather to an agent’s belief about the world, so why can’t agent’s beliefs violate the probability axioms?
  o **de Finetti Theorem** – If Agent 1 expresses a set of degrees of belief that violate Kolmogorov’s Axioms, then there is a combination of bets Agent 2 can place that guarantees that Agent 1 will lose money every time.
  o Decision making is an unavoidable process (not making a decision is a decision).

• **Probability Philosophy (Briefly)**
  o **Frequentist** – probabilities are results of repeated experiments.
  o **Objectivist** – probabilities come from a propensity of objects to act in a certain way.
  o **Subjectivist** – probabilities are a way of characterizing beliefs.
  o **Principle of Indifference** – Propositions that are syntactically ‘symmetric’ w.r.t. evidence should be given equal probability.
  o **Inductive Logic** – a logic capable of computing the correct probability for any proposition from any collection of observations. Unclear if it exists.

**Probabilistic Inference** – computation of posterior probability of query proposition from observed evidence.

• **Marginalization (Conditioning)** – variables other than the query variable are summed out in order to obtain the probability of the query variable:
joint:  \( P(Y) = \sum_z P(Y, z) \)

conditional:  \( P(Y) = \sum_z P(Y \mid z) P(z) \)

- **Normalization** – introduction of constant \( \alpha \) that normalizes the distribution to 1 instead of explicitly computing the constant probabilities that do not depend on the variable being calculated.

- **Combinatorial Explosion** – Full Joint Table is size \( O(2^n) \) and requires \( O(2^n) \) time to be processed.

**Marginal Independence** – variables independent of each other can be factored:
\[
P(X, Y) = P(X) P(Y) \quad P(X \mid Y) = P(X)
\]

- If a complete set of variables can be divided into independent subsets, then the full joint distribution can be factored into separated joint distributions on those subsets.

**Bayes’ Rule** – application of product rule that allows diagnostic beliefs to be derived from casual beliefs:
\[
P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)} \quad P(Y \mid X, e) = \frac{P(X \mid Y, e) P(Y \mid e)}{P(X \mid e)}
\]

- Diagnostic knowledge is often more fragile than casual knowledge \( \Rightarrow \) direct casual (model-based) knowledge provides robustness needed for probabilistic systems to function in the real world.

- **Conditional Independence** – implies that two variables \( X, Y \) are independent given variable \( Z \):
\[
P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z) \quad P(X \mid Y, Z) = P(X \mid Z)
\]

- Given \( n \) symptoms that are conditionally independent given \( Z \), the size of the representation grows as \( O(n) \) instead of \( O(2^n) \).
  - Allows probabilistic systems to scale up.
  - Conditional independence assertions are more common than marginal independence assertions.

- **Naïve Bayes Model** – a single cause \( Y \) directly influences a number of events \( X_i \) that are all conditionally independent given the cause:
\[
P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_i P(X_i \mid Y)
\]
  - Often works in situations where conditional independence does not hold.
14: Probabilistic Reasoning

- **Bayesian Network** (Belief Network, Probabilistic Network, Casual Network, or Knowledge Map)
  - A *directed acyclic graph* (DAG) representing the dependency structure amongst random variables thus providing a concise specification of any full joint probability distribution.
  - **Nodes** – the random variables of the problem (observed variables are shaded).
  - Each node $X_i$ has a conditional probability distribution representing $P(X_i \mid \text{parents}(X_i))$. In the discrete variable case, this can be represented by a Conditional Probability Table (CPT).
  - **Directed Arcs** – represent the dependency of one random variable on another. In the Undirected case, represents interdependency between two variables.
    - The Bayesian Network captures conditional independence relationships in its edges.
    - Probabilities summarize a potentially infinite set of circumstances that are not explicit in the model but rather appear implicitly in the probability.
    - If each variable is influenced by at most $k$ others and we have $n$ random variables, we only need to specify $n \times 2^k$ probabilities instead of $2^n$.
    - More general case of Bayesian Network is “Graphical Model”.

- **Conditional Probability Table (CPT)** - describes the conditional probability of each value of the node for each *conditioning case*, or possible combination of the values of the parent nodes.
  - In general the size of the table is
    \[ \#(X_i) \cdot \prod_{X_j \in \text{parents}(X_i)} \#(X_j) \] 
  where $\#()$ specifies the size of the domain of a variable.
  - The size of the table can be reduced since the total probability for each conditioning case must be 1.

- **Chain Rule:**
  - General:
    \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, X_2, \ldots, X_{i-1}) \]
  - In the case of a BN, if we number the nodes in topological order, the conditional terms in $P(X_i \mid X_1, X_2, \ldots, X_{i-1})$ are all predecessors of $X_i$ and thus, conditional independence reduces this to $P(X_i \mid \text{parents}(X_i))$.
  - Chain Rule in BN:
    \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i)) \]
• Constructing a Bayesian Network
  o The parents of node \( X_i \) should be all nodes that directly influence \( X_i \) from the set of nodes \( X_1, \ldots, X_{i-1} \).
  o The correct order in which to add nodes is to add the “root causes” first, then the variables they influence, and so on until leaves are reached.
  o Slight dependences may not be worth adding due to increased complexity.
  o Constructing Networks for a “Causal Model” will result in specifying fewer numbers, which are often easier to come up with.

• Independence in a Bayesian Network
  o A node is conditionally independent of all non-descendants given its parents.
  o A node is conditionally independent of all other nodes in network given its Markov Blanket (parents, children, and children’s parents).
  o \( d \)-separation – two nodes X and Y in a BN are d-separated if every path between X and Y is blocked.
    ▪ A path between X and Y is blocked if it has any of the following 3 cases for any 3 nodes along the path.
      • head-to-tail with intermediary observed: \( A \perp B \mid C \)
      • tail-to-tail with intermediary observed: \( A \perp B \mid C \)
      • head to head with neither the intermediary nor any of its descendants observed: \( A \perp B \mid \emptyset \)

\[
\begin{array}{cccc}
A & C & \quad & A \quad & C \quad & B \quad & A & C & B \\
\circlearrowleft & \bullet & \circlearrowleft & \circlearrowright & \bullet & \circlearrowright & \circlearrowleft & \bullet & \circlearrowleft
\end{array}
\]

• Canonical distributions: fit a standard pattern with an easily filled in CPT.
  o deterministic node – value is specified exactly by value of parents with no uncertainty.
  o noisy-OR – uncertain ability of each parent to cause the child to be true since the causal relationships may be inhibited.
    ▪ Assumes all possible causes are listed (others can be grouped in a leak node for miscellaneous causes).
    ▪ Assumes the inhibition of each parent is independent of the others.
    ▪ Hence, we need only specify \( O(k) \) parameters instead of \( O(2^k) \) for \( k \) causes: the probability of inhibition of each of the causes.
  o other noisy-operators (MAX, AND, etc).