1 Uncertainty, Probability and Bayesian Networks

1.1 Uncertainty and Probability

1.1.1 Motivation: Why Uncertainty?

Classical logic assumes that agents know the "whole truth" (logical statements are true or false). However, in the real world, agents must deal with **uncertainty** due to:

- **Partial Observability:** We cannot see the state of the entire world (e.g., road state, other drivers' plans).
- **Noisy Sensors:** Information received may be incorrect (e.g., wrong traffic reports).
- **Uncertainty in Action Outcomes:** Actions are stochastic (e.g., a flat tire, accident).
- **Modeling Complexity:** It is impossible to model every single factor (The *Qualification Problem*).

Ignorance Types

- **Laziness**: Listing all exceptions is too tedious.
- **Theoretical Ignorance**: The underlying mechanisms are not fully understood (e.g., perfect weather modeling).
- **Practical Ignorance**: The rules are known, but the specific data for a situation is missing.

1.1.2 Probability Theory Basics

Probability provides a way to summarize uncertainty. It represents a **degree of belief**, not necessarily a degree of truth.

Kolmogorov's Axioms These axioms constrain the probabilistic beliefs an agent can reasonably hold.

1. $0 \leq P(a) \leq 1$ (All probabilities are between 0 and 1).
2. $P(false) = 0$, $P(true) = 1$ (Necessarily true propositions have probability 1).
3. $P(a \lor b) = P(a) + P(b) - P(a \land b)$ (Probability of a disjunction).

The Dutch Book Theorem

Proposed by Bruno de Finetti (1931). It states that if an agent holds a set of beliefs that violate the axioms of probability, a betting strategy (a "Dutch Book") can be constructed against them such that the agent is **guaranteed to lose money** regardless of the outcome.

Random Variables Instead of dealing with raw events, we use **Random Variables (RVs)** to describe the world.

- **Boolean:** $X \in \{true, false\}$ (e.g., $hasUmbrella$).
- **Discrete:** Finite set of values (e.g., $Weather \in \{sunny, rain, cloudy\}$). Values must be *exhaustive* and *mutually exclusive*.
- **Continuous:** Infinite domain (e.g., *Temperature*).

1.1.3 Distributions and Inference

Joint Probability Distribution The joint distribution $P(X_1, \ldots, X_n)$ assigns probabilities to every possible combination of values for all random variables.

- It allows us to answer *any* question about the domain.

- **Problem:** The size of the table grows exponentially ($O(d^n)$ for n variables of domain size d).

Marginalization (Summing Out) We can extract the distribution of a subset of variables by summing out the others.

$$P(Y) = \sum_z P(Y, z)$$

This is how we recover simple probabilities from the joint distribution.

Conditional Probability Represents beliefs given evidence.

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Using the **Product Rule**, we can rewrite the joint probability:

$$P(a, b) = P(a|b)P(b) = P(b|a)P(a)$$

The Chain Rule

Used to decompose a joint distribution into a product of conditional probabilities:

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

1.1.4 Bayes' Rule

Derived from the product rule ($P(x, y) = P(x|y)P(y) = P(y|x)P(x)$).

$$P(Hypothesis|Evidence) = \frac{P(Evidence|Hypothesis) \cdot P(Hypothesis)}{P(Evidence)}$$

- $P(H|E)$: **Posterior** (Probability of hypothesis after seeing evidence).
- $P(E|H)$: **Likelihood** (Probability of evidence assuming hypothesis is true).
- $P(H)$: **Prior** (Initial probability of hypothesis).
- $P(E)$: Marginal Likelihood (Normalization constant).

Example: The AIDS Test (Base Rate Fallacy) Consider a test for a disease:

- $P(pos|sick) = 0.99$ (Sensitivity)
- $P(neg|healthy) = 0.995$ (Specificity), so $P(pos|healthy) = 0.005$.
- $P(sick) = 0.0001$ (Prior - Base rate).

If you test positive (pos), what is the probability you are sick ($sick$)?

$$P(sick|pos) = \frac{P(pos|sick)P(sick)}{P(pos)}$$

Where $P(pos) = P(pos|sick)P(sick) + P(pos|healthy)P(healthy)$.

$$P(sick|pos) = \frac{0.99 \cdot 0.0001}{(0.99 \cdot 0.0001) + (0.005 \cdot 0.9999)} \approx 0.0194$$

Lesson: Even with a reliable test, if the disease is rare, a positive result often implies a low probability of actually having the disease.

1.2 Bayesian Networks

1.2.1 Independence

To avoid the exponential explosion of the joint distribution, we utilize **Independence**.

- **Independence:** $P(X, Y) = P(X)P(Y)$ or $P(X|Y) = P(X)$.
- **Conditional Independence:** X and Y are independent given Z if $P(X|Y, Z) = P(X|Z)$.

Example: *Age* and *Gender* are independent. *Cancer* is independent of *Age* and *Gender* **given** *Smoking*.

1.2.2 Bayesian Networks (BN)

A Bayesian Network is a data structure to represent dependencies compactly.

- **Structure:** A Directed Acyclic Graph (DAG).
- **Nodes:** Random variables $X_1, \ldots, X_n$.
- **Edges:** Directed edge $X_i \to X_j$ indicates direct influence.
- **Parameters:** Each node X_i has a Conditional Probability Table (CPT) quantifying $P(X_i|Parents(X_i))$.

Semantics The full joint distribution is defined as the product of the local conditional distributions:

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

Local Markov Assumption

Each variable X_i is conditionally independent of its non-descendants, given its parents.

1.3 Exact Inference in Bayesian Networks

1.3.1 The Inference Task

Given evidence $E = e$ and a query variable X, we want to compute $P(X|e)$.

$$P(X|e) = \frac{P(X, e)}{P(e)} \propto \sum_{y} P(X, e, y)$$

Where y are the *hidden* variables (neither query nor evidence).

1.3.2 Variable Elimination

A systematic method to perform summation. Instead of computing the full joint (exponential) and then summing, we push sums inwards to factor out terms.

Algorithm Steps:

1. **Factorize:** Write the joint distribution as a product of CPTs.
2. **Order:** Choose an elimination order for hidden variables.
3. **Sum Out:** For each variable Z to be eliminated:
    - Collect all factors containing Z.
    - Multiply them.
    - Sum over the values of Z.
    - Replace the old factors with the new factor (result).

Example Walkthrough (Abstract): Factors: $P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$. Eliminate v:

$$f_v(t) = \sum_v P(v)P(t|v)$$

New set of factors: $f_v(t), P(s), P(l|s)\ldots$ Proceed sequentially.

1.3.3 Complexity

NP-Hardness

Inference in Bayesian Networks is **NP-Hard**. This is proven via reduction to 3-SAT (Boolean Satisfiability). Even approximate inference with bounded error is NP-Hard.

1.4 Approximate Inference: Sampling

Since exact inference is hard, we use stochastic simulation (Monte Carlo). We draw N samples and estimate probabilities by counting.

1.4.1 Direct Sampling (Empty Network)

Used when there is no evidence.

1. Sort variables topologically.
2. Sample X_1 from $P(X_1)$.
3. Sample X_2 from $P(X_2|Parents(X_2))$ (using value sampled for parents).
4. Repeat until all variables are sampled.

1.4.2 Rejection Sampling

Used for computing $P(X|e)$.

1. Generate samples from the empty network.
2. **Reject** (discard) any sample that does not match the evidence e.
3. Estimate $P(X|e)$ by counting frequencies in the remaining samples.

Drawback: If the evidence is rare, we reject almost all samples, making it inefficient.

1.4.3 Markov Chain Monte Carlo (MCMC)

Instead of generating independent samples from scratch, the system wanders through the state space. The state is the current assignment of all variables.

Markov Blanket

The Markov Blanket of a node consists of:

- Its Parents.
- Its Children.
- Its Children's Parents.

A node is conditionally independent of *all other nodes* in the network given its Markov Blanket.

Gibbs Sampling Algorithm To estimate $P(X|e)$:

1. Fix evidence variables to their observed values e.
2. Initialize non-evidence variables randomly.

3. Loop:
    - Pick a non-evidence variable Z_i.
    - Resample Z_i from $P(Z_i | MarkovBlanket(Z_i))$.
    - Record the state.

This process creates a Markov Chain that converges to the true posterior distribution (stationary distribution) if the chain is irreducible, aperiodic, and ergodic.