# Lecture #3: *Probability Theory and Bayesian Networks*

Scribes: *Isabel Rayas, Yutong Gu, Tejas Bharath, Puranjay Rajvanshi*

## 1 Introduction

The lecture on September 4$^{th}$ provides students an introductory understanding of Probability Theory and Bayesian Networks with examples of how they are applied in the field of robotics. This lecture covered the material relating to probability theory by introducing continuous random variables and contrasting their properties with discrete random variables. Additionally, the Markov Assumption and its impact on probability theory and state estimation was discussed.

The following notes are a collaborative effort to summarize the topics covered in the lecture on September 4$^{th}$. Supplemental material has been included to provide a more comprehensive understanding of these areas. Moreover, real-world examples have been given to relate the topics covered in class with practical application.

## 2 Probability Theory

Probability Theory is a mechanism used to understand and make judgements about a system. Essentially, it defines probabilities as a measure of the likeliness of events occurring in a system. Each events respective probability is represented in a **sample space**, which contains the set of all outcomes within the system. Inside the sample space, any assigned probability should be a non-negative number such that the collection of all possible outcomes in the sample space total 1. Also, if two outcomes cannot occur at the same time, the probability of either outcome occurring is the sum of the probabilities of the individual outcomes.

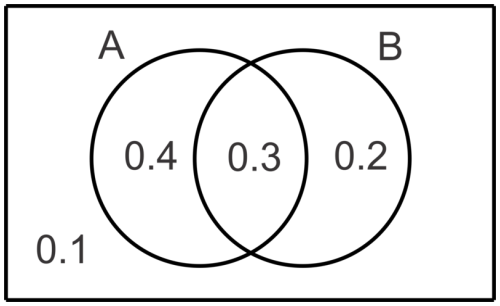


Figure 1: 2 Event Sample Space

For example, Figure 1 shows an example of a two-event sample space with their respective probabilities marked. Note that the sum of all probabilities inside a sample space should total 1 ($0.4 + 0.3 + 0.2 + 0.1 = 1$).

Probability distributions can be split into two types of distributions: **discrete** and **continuous**. Discrete probability distributions describe the probability of each occurrence out of a set of discrete random variables (countable set). Conversely, continuous distributions describe the probability of each occurrence over a range of continuous random variables (infinite set of possible values). Additionally, variables (events in a system) can be either **dependent** or **independent**. Dependent events are events in which the outcome of one is affected by that of another, whereas the outcomes of independent events are not affected by the outcome of the other(s). These distinctions are crucial when making estimations or predictions about the system based on the known probabilities of the events.

## 2.1 Discrete Probability

**Discrete Random Variables (DRVs)** are random variables which map a set of events to a finite set of real numbers. For instance, the possibility that a coin lands as heads may be assigned the value 1 while tails is assigned the value 0, or vice versa. Another example of a discrete random variable is the possible values that a die can take on after a roll (which is 1 to 6). Two important theorems to know for DRVs are the **Total Probability Theorem**, and **Bayes Theorem**.

**Total Probability Theorem**:
$p(x) = \sum_A p(x|A)p(A)$

**Bayes Theorem**:
$p(x \cap A) = p(x|A)p(A)$

These two theorems modified or combined to give us other useful equations such as:

$$p(x|A) = \frac{p(x \cap A)}{p(A)}$$

$$p(x|A) = \frac{p(A|x)p(x)}{p(A)}$$

$$p(x|A) = \frac{p(A|x)p(x)}{\sum_{x'}(p(A|x')p(x'))}$$

The **expectation value, or 1st moment,** of a DRV is more commonly known as the mean of a random variable. Formally, it is defined as the sum of the products $x$ and $p(x)$ for all $x$, where $x$ is the value a DRV may take on and $p(x)$ is the probability of that value occurring. Mathematically, it can be represented as $E[x] = \sum_x xp(x)$. A few properties of the 1st moment are as follows:

$$E[ax] = aE[x]$$

$$E[x + b] = E[x] + E[b]$$

$$E[ax + b] = aE[x] + E[b]$$

## 2.2 Continuous Probability

**Continuous Random Variables (CRVs)** are very similar to DRVs. However, a key difference is that CRVs map events to an infinite, or continuous, set of values. For example, a CRV may be the voltage on a wire at any given moment, or the percent loss of data packets across a noisy channel. These values cannot be defined by a discrete set of values, therefore a CRV is more appropriate. The consequences of CRVs are that, rather than taking a sum of the possible values the random variable may take on, it is instead an integral of a **probability density function (PDF)**[1] $f(x)$, where $x$ is a continuous value that the CRV may take on and $f(x)$ is the probability density of that value occurring. Since, logically, the probability that a CRV will take on an exact number is infinitesimally small, it is necessary to use integrals to define a range where the value can fall within. This results in slightly different equations compared to those defined in the Discrete Probability section.

**Total Probability Theorem**:
$p(x) = \int_A p(x|A)p(A)$

[1] DRVs also have probability density functions; however, they are in the discrete domain. For example, a die would have a discrete probability density function $f(x) = 1/6$ for $x \in \{1, 2, 3, 4, 5, 6\}$

**Bayes Theorem**:
$p(x \cap A) = p(x|A)p(A)$

These two theorems modified or combined to give us other useful equations such as:
$p(x|A) = \frac{p(x \cap A)}{p(A)}$

$p(x|A) = \frac{p(A|x)p(x)}{p(A)}$

$p(x|A) = \frac{p(A|x)p(x)}{\int_{x'}(p(A|x')p(x'))}$

The expectation value, or 1st moment, of a CRV is written as $E[x] = \int_x xp(x)$. The properties of the 1st moment are the same as that of DRVs:
$E[ax] = aE[x]$

$E[x + b] = E[x] + E[b]$

$E[ax + b] = aE[x] + E[b]$

An important distribution to know of for CRVs is the **Normal Distribution, or Gaussian Distribution**. This distribution is a commonly used distribution found often in the real world. It has the following probability density function.
$f(x) = \frac{1}{(\sqrt{2\pi\omega^2})} e^{\frac{-(x-\mu)^2}{2\omega^2}}$

Fortunately, this PDF can be simplified to simply writing $N(x; \mu, \omega^2)$

## 2.3 Random Vectors

When there are multiple random variables that we are interested in which depend on each other, it may be easier to represent them as a vector. These are what is called **random vectors**. For example, let

$$\mathbf{X} = \begin{vmatrix} X_1 \\ X_2 \\ X_3 \end{vmatrix}$$

where $X_1$, $X_2$, $X_3$ represent the different positions an object may be at any point in time in the x, y, and z direction, respectively. Then, **X** is the random vector with a **joint probability density function, or multivariate function** $f(X_1 = a, X_2 = b, X_3 = c)$ which describes the possibility density that $X_1$, $X_2$, and $X_3$ will take on the values $a$, $b$, and $c$,

respectively.

Random vectors also follow some of the same rules as DRVs and CRVs such as the total probability of every combination of values a random vector may take on must equal 1. For discrete random variables this means that
$\sum_{X_1=-\infty}^{+\infty} \sum_{X_2=-\infty}^{+\infty} \cdots \sum_{X_n=-\infty}^{+\infty} f(\mathbf{X}) = 1.$

The expectation vector for a discrete random vector is:
$\overline{\mathbf{X}} = E[\mathbf{X}] = \sum_{X_1=-\infty}^{+\infty} \sum_{X_2=-\infty}^{+\infty} \cdots \sum_{X_n=-\infty}^{+\infty} \mathbf{X} * f(\mathbf{X}).$

For CRVs, it is the same, except that it is an series of integrals instead of a summations.

For a multivariate normal distribution $N$, we have the following probability density function:

$$p(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} \quad exp\left(\frac{(\mathbf{X}-\bar{\mathbf{X}})^T \Sigma^{-1} (\mathbf{X}-\bar{\mathbf{X}})}{2}\right) \tag{1}$$

where $\Sigma$ is the covariance matrix and $n$ is the number of random variables in $\mathbf{X}$. The input here, $\mathbf{X}$, is a random vector, but note that the output of the multivariate function is a scalar.

The **covariance matrix** is a matrix whose element in the i,j position is the covariance between the i-th and j-th elements of a random vector. Using the notation where $\mathbf{X}$ is an *n-length* random vector, the covariance matrix is calculated by the following function:

$$\Sigma = [\Sigma_{ij}] = E[(\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^T] \tag{2}$$

This can also be expanded to integral form and can be written as the following:

$$\Sigma = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^T * f(\mathbf{X}) dx_1 dx_2 ... dx_n \tag{3}$$

Note that resulting covariance matrix is always both symmetric matrix and positive semidefinite, so the inverse exists.

The correlation coefficient can be represented as a standardized form of covariance by

using the following formula:

$$\rho_{i,j} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} * \Sigma_{jj}}} \tag{4}$$

# 3 Dynamical Systems and State Estimation

In a dynamical system, there is a dependence on time. Generally, in robotics, the robot can do one of two things at any given time: observe or actuate. The environment the robot is in can change due to the actions the robot takes or due to external factors, both of which may time-dependent. For a robust robotic system, these changes should be modeled and accounted for.

We define the following random variables for a dynamical system:

$x$: state- This is the true state of the world. It contains all the information that an omniscient observer could know, and therefore also all the information that the robot could need to make a decision. Note that this does not mean that the true state is necessarily *known* to the robot (in fact, it almost never is).

$z$: observation- This is any information about the world that the robot collects. Some examples common in robotics could include images, LIDAR sensor readings, pressure sensor readings, or audio input. Observations are important because the robot generally does not have knowledge of the true state; rather, it has a belief of the likelihood of each possible state being the true state. The observations often have some level of uncertainty associated with the information about the state they provide.

$u$: input- This is the control commands that are sent to the robot. They are often thought of as actions taken. Examples could include moving forward, turning left, closing the grasp of the robot hand, or pushing a button.

Because robot knowledge is imperfect, its belief over the states may be influenced by the observations it receives. The input may change the environment, which may then change the state, which may affect the observations, etc.

## 3.1 The Markov Assumption

The Markov assumption simplifies many of the equations described above, and allows us to compute beliefs over states much quicker. It is an assumption that the current state is only dependent on the previous state and the previous input. In other words,

$$p(x_t|x_{0:t-1}, u_{0:t}) = p(x_t|x_{t-1}, u_t) \tag{5}$$

Note that the Markov assumption is an example of conditional independence, since $x_t$ is independent of all other states, inputs, and observations, given $x_{t-1}$ and $u_t$.

This is also known as the **transition function**, as it describes the probability of transitioning to the state at time $t$ from the state at time $t-1$. The transition function can specify the dynamics of a system, for example in the case of a car.

The transition function shows that the current state depends on the previous state and the previous input, but *not the previous observation*. This makes sense because the state $x$ is the true world state, and nothing our sensors report will change the true state. We use our sensors and their observations to help us estimate $x$, but our measurements do not inherently change the world. Otherwise, having a bad camera could change a robot's position!

The **observation function** under the Markov assumption similarly states that the probability of getting an observation at time $t$ only depends on the previous state.

$$p(z_t|x_{0:t}, z_{1:t-1}, u_{1:t-1}) = p(z_t|x_t) \tag{6}$$

Receiving an observation reduces the variance in the belief of the robot; for example, if you are unsure where you are but you get an image of your front door, your belief that you are in front of your house will increase while the belief that you are at either of the houses next door will decrease.

### 3.2 State Estimation

State estimation is the problem of determining the current state, given the history of $z$ (observations) and $u$ (inputs). We denote the probability of being in the state $x_t$ given this history as our **belief over $x$**, or $bel(x_t)$:

$$bel(x_t) = p(x_t|z_{1:t}, u_{1:t}) \tag{7}$$

Using Bayes' rule for 3 random variables, we can expand this further:

$$= \frac{p(z_t|x_t, z_{1:t-1}, u_{1:t})p(x_t|z_{1:t-1}, u_{1:t})}{p(z_t|z_{1:t-1}, u_{1:t})} \tag{8}$$

We call the normalizing term in the denominator $\eta$ and further marginalize over $x_{t-1}$, recalling that marginalizing the probability of $A$ given $C$ over $B$ can be done by $P(A|C) =$

$\sum_B P(A|B,C)P(B|C)$:

$$= \eta p(z_t|x_t) \int_{-\infty}^{\infty} p(x_t|x_{t-1}, z_{1:t-1}, u_{1:t}) p(x_{t-1}|z_{1:t-1}, u_{1:t}) dx_{t-1} \tag{9}$$

Finally, we apply the Markov assumption here to simplify:

$$= \eta p(z_t|x_t) \int_{-\infty}^{\infty} p(x_t|x_{t-1}, u_t) p(x_{t-1}|z_{1:t-1}, u_{1:t-1}) dx_{t-1} \tag{10}$$

If we look at the last term in this equation, we notice that it is equivalent to $bel(x_{t-1})$. This leads us to the recursive equation for our belief over the states:

$$bel(x_t) = \eta p(z_t|x_t) \int_{-\infty}^{\infty} p(x_t|x_{t-1}, u_t) bel(x_{t-1}) dx_{t-1} \tag{11}$$

We can use this belief update equation to do inference on a dynamical system, which we can often represent as a graphical network.

## 4 Bayesian Networks

Also know as **belief network**, **probabilistic network**, **causal network**, and **knowledge map**, a Bayesian network is used to represent a full joint probability distribution in a concise manner. The need of Bayesian network arises from the idea of having a complicated domain with $n$ different proposition variables: a joint probability distribution would require $2^n$ numbers for it to be fully specified, so if $n$ is large, then it won't be feasible for us to create a full joint probability distribution. This is where Bayesian network comes in; the intuition is that there's almost always some independence between variables, so we don't have to know all $2^n$ variables to find out what's happening in our environment.

Also called as sentential variable, is a variable which can either be true or false

A Bayesian network is a set of nodes (or variables) connected with directed arcs, forming an acyclic graph. The full specification is as follows:

1. Each node corresponds to a random variable, which may be discrete or continuous.
2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y , X is said to be a parent of Y. The graph has no directed cycles (and hence is a directed acyclic graph, or DAG).
3. Each node $X_i$ is conditionally independent of its own descendants, given its parents.

To strengthen our understanding let's consider an example. Suppose we have a robot that looks like this:

Figure 2: https://images.app.goo.gl/aMzkoGPCq5X6bVYj9

Usually when the robot has sufficient battery, the robot's light is turned on, and it's able to move from one place to another. If the robot's battery is low, then its light is turned off and it loses its mobility as well. A broken wheel could also result in the robot's loss of mobility, even if the battery is full.
Here we can see that there are 4 proposition variables: low battery, light off, no mobility, and broken wheel. For a joint probability distribution we'd need $2^4$ assignments.
Instead, we can observe that the mobility of our robot is dependent on "low battery" and "broken wheel" whereas the robot's light is only dependent on "low battery" and is completely independent of "broken wheel". Therefore, we can construct the following Bayesian network to represent our environment:

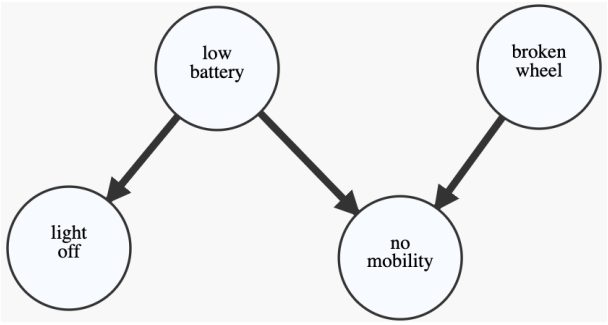


Figure 3: A concise representation of our robot environment

Now we'll discuss all the possible ways three nodes (A,B,C) can be connected in a Bayesian network, and how evidence can be transmitted through them.

### 4.1 Serial Connection

A serial connection consists of A pointing to B, which points to C. So if we get to know what A is then we can gain information about B, which then by the same logic will give us information about C. On the other hand, if we already know what B is, then knowing what A is will have no effect on the information we get regarding C. Similarly, knowing what C is will give us information about A only when we don't know what B is.

For example, let's take our robot environment once again and add another variable to it: the robot makes a sound when it moves. This new addition can give us the following sub-graph:

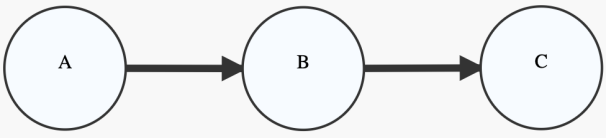


Figure 4: A: low battery, B: no mobility, C: no sound

So, if we know that the robot has low battery, then we can infer that the robot has no mobility, and since it has no mobility, it won't make a sound.
Now let's consider another possibility, what if we already knew that the robot is not moving?
If such was the case, then it would be obvious that the robot would also not make a sound, and our information on whether the battery is high or low would not make any difference to the conclusion we've already drawn.
Similarly, let's take another situation where we only know that the robot is not making a sound. Through this evidence our belief in the possibility that the robot has no mobility increases, which would also increase our belief in the possibility that the robot's battery might be low. But if we already know that the robot is not moving then our knowledge about whether it's making a sound does not affect our belief on the status of the battery.

Therefore, when B is already instantiated, knowledge about A has no effect on our knowledge about C and vice-versa.

$$P(A|B,C) = P(A|B) \tag{12}$$

$$P(C|B,A) = P(C|B) \tag{13}$$

### 4.2 Diverging Connection

A diverging connection works similar to a serial connection, it consists of B pointing to both A and C. So if we get to know what A is then we can gain some information about B, which then will give us some information about C. On the other hand, if we already know what B is, then knowing what A is will have no effect on the information we infer regarding C. Similarly, knowing what C is will give us some information about A only when we don't know what B is.

Following the previous example, let's consider the following variables: "no mobility", "low battery", and "light off". This will give us the following sub-graph:

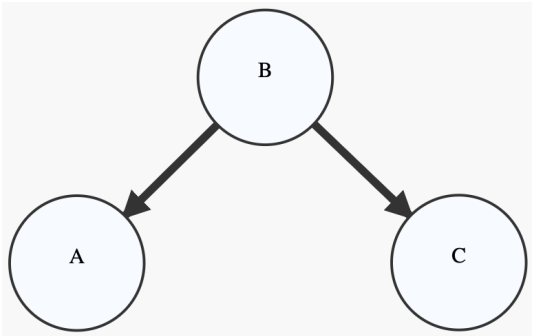


Figure 5: A: no mobility, B: low battery, C: light off

So, if we know that the robot has no mobility, then we can infer that the robot may have low battery, and since it may have low battery, its light might be off.
Now let's consider another possibility, what if we already knew that the robot has low battery?
If such was the case, then it would be obvious that the robot's light is off, and our information on the robot's mobility would not make any difference to the conclusion we've already drawn.
Similarly, let's take another situation where we only know that the robot's light is off. Through this evidence our belief in the possibility that the robot has low battery increases, which would also increase our belief in the possibility that the robot has no mobility. But if we already know that the robot's battery is low, then our knowledge about whether its light is on or off does not affect our belief on the status of its mobility

Therefore, when B is already instantiated, knowledge about A has no effect on our knowledge about C and vice-versa.

$$P(A, C|B) = P(A|B)P(C|B) \tag{14}$$

### 4.3 Converging Connection

A converging connection works a bit differently from the other two. It consists of both A and C pointing to B. So if we get to know what B is then and only then can A and C give information about each other.

Like the previous example, let's consider the following variables: "broken wheel", "no mobility", and "low battery". This will give us the following sub-graph:

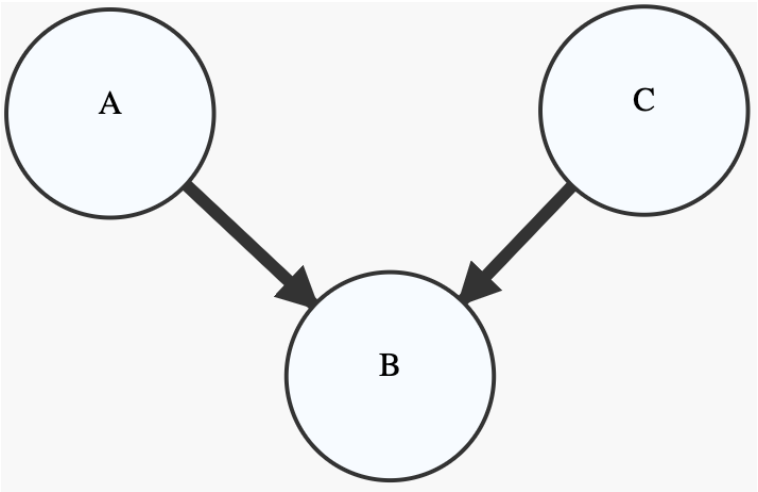


Figure 6: A: broken wheel, B: no mobility, C: low battery

So, even if we know that the robot has a broken wheel, we cannot infer on what's the status of the robot's battery. But if we knew that the robot has no mobility, then we could say that our belief that the robot has a low battery would be decreased because we already know that it has a broken wheel.
Similarly, let's take another situation where we only know that the robot's battery is low. Through this evidence our belief in the possibility that the robot has a broken wheel will not be affected. But if we already know that the robot has no mobility, then our belief that the robot has a broken wheel decreases because we already know that the robot's immobility is likely caused by its low battery.

Therefore, when B is not instantiated, A and C are independent of each other.

$$P(A, C) = P(A)P(C) \tag{15}$$

$$P(A, C|B) \neq P(A|B)P(C|B) \tag{16}$$

# 5 References


- http://ai.stanford.edu/~paskin/gm-short-course/lec1.pdf
- http://www-math.bgsu.edu/~albert/m115/probability/prob_rules
- MIT OCW 6.825 Techniques in Artificial Intelligence
- Artificial Intelligence: A modern approach , by Stuart Russel and Peter Norvig