Probability theory

In this summary, we will be examining a lot of stochastic systems. Stochastic systems deal with probabilities. So, let's dive into the probability theory first.

1 The probability distribution function

1.1 Definition of the probability distribution function

Very important in probability theory is the **probability distribution function** (PDF) $f(u)$. This function has as limits $f(-\infty = 0)$ and $f(\infty) = 1$. It also increases: if $u < v$, then $f(u) \leq f(v)$. Finally, PDFs are also **right continuous**. To find out what this means, we examine some discontinuity v in the graph of $f(u)$. Now let's approach this point v from the right. The value which we get is $f(u+)$. Right continuous functions now must have $f(u) = f(u+)$.

We can make a distinction between **continuous** and **discrete** PDFs. Continuous PDFs usually have a continuous shape: the value of $f(u)$ more or less gradually increases from 0 to 1. For **continuous PDFs** we also have

$$f(u) = \int_{-\infty}^{u} p(v)\,dv, \qquad \text{where} \qquad f(\infty) = \int_{-\infty}^{\infty} p(v)\,dv = 1. \tag{1.1}$$

In the above equation, $p(u)$ is the **probability density function**.

Discrete PDFs are rather different. The graph of $f(u)$ now takes the shape of a staircase. The points where $f(u)$ jumps up are denoted by u_n.

$$f(u) = \sum_{u_n < u} p(n), \qquad \text{where} \qquad f(\infty) = \sum p(n) = 1. \tag{1.2}$$

Now, $p(n)$ is called the **probability frequency function**.

1.2 Examples of probability distribution functions

Several examples of PDFs exist. We'll examine a few now. The **Bernoulli distribution** has a discrete PDF. Given the parameter q (satisfying $0 \leq q \leq 1$), the distribution is defined by

$$p(1) = q \qquad \text{and} \qquad p(0) = 1 - q. \tag{1.3}$$

The **Poisson distribution** is discrete as well. Given the parameter λ (satisfying $\lambda \in \mathbb{R}_+$), it is defined by

$$p(k) = \lambda^k \frac{e^{-\lambda}}{k!}. \tag{1.4}$$

In this equation, we must have $k \in \mathbb{N} = \{0, 1, \ldots\}$.

The **gamma distribution** is continuous. Its parameters are λ and r and satisfy $\lambda, r \in \mathbb{R}_+$. The distribution is defined by

$$p(v) = \frac{v^{r-1}}{\lambda^r} \frac{e^{-\frac{v}{\lambda}}}{\Gamma(r)}, \qquad \text{where} \qquad \Gamma(r) := \int_0^{\infty} v^{r-1} e^{-v}\,dv. \tag{1.5}$$

The function $\Gamma(r)$ is known as the **gamma function**. (By the way, the ':=' means 'is per definition'.)

However, the most important distribution is the **Gaussian distribution**, also known as the **normal distribution**. This continuous distribution has as parameters a mean vector m and a variance matrix

Q. m satisfies $m \in \mathbb{R}^n$ while Q satisfies both $Q \in \mathbb{R}^{n \times n}$ and $Q = Q^T > 0$. (With $Q > 0$ we mean that Q is **strictly positive definite**, which in turn demands that $x^t Q x > 0$ for all vectors x. This, in turn, demands that all eigenvectors of Q are positive.) The distribution is now defined by

$$p(v_1, v_2, \ldots, v_n) = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} e^{-\frac{1}{2}(v-m)^T Q^{-1}(v-m)}. \tag{1.6}$$

But why is this distribution so important? Well, let's suppose that we have a number of independent distributions. If we add these distributions up and normalize them, then the **central limit theorem** claims that the resulting distribution will converge to a Gaussian distribution. The more distributions are added up, the close the result will be a to a Gaussian distribution. And since many phenomena in real life are the result of sums of distributions, we can use the Gaussian distribution to approximate them.

2 Measurable spaces and probability spaces

2.1 σ-algebras

Let's examine a set Ω. A σ-algebra F on Ω is a collection of subsets of Ω, satisfying three important rules.

1. If the set A is in F ($A \in F$), then the complement A^c is also in F ($A^c \in F$). (In other words, F is closed with respect to complementation.)
2. Let's examine a set of sets $\{A_1, A_2, \ldots, A_n\}$ such that all A_i are in F. Now let's take the union of all these sets. This union must now also be in F. In an equation, we have $A_1 \cup A_2 \cup \ldots \cup A_n \in F$.
3. The set Ω is in F. (And thus, due to rule 1, also the empty set $\emptyset$ is in F.)

Examples of σ-algebras include $\{\emptyset, \Omega\}$ and $\{\emptyset, A, A^c, \Omega\}$ for every set $A \in \Omega$.

A tuple (Ω, F), consisting of a set Ω and a σ-algebra F on Ω, is called a **measurable space**. A σ-algebra G on Ω consisting of subsets of the σ-algebra F (thus satisfying $G \subseteq F$) is called a sub-σ-algebra.

2.2 Probability measures

Suppose we have a measurable space (Ω, F). Let's examine a function $P : F \to \mathbb{R}_+$. (In other words, the function P takes as input elements of F and as output it gives elements of $\mathbb{R}_+$.) Also examine any disjoint set of sets $\{A_1, A_2, \ldots, A_n\}$ such that all A_i are in F. (With disjoint, we mean that A_i and A_j (with $i \neq j$) have no elements in common: $A_i \cap A_j = \emptyset$.) We now say that P is σ-additive if

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_n). \tag{2.1}$$

If we also have that $P(\Omega) = 1$, then we say that P is a **probability measure**. We also say that the triple (Ω, F, P) is a **probability space**. Such a probability space has several interesting properties.

1. $P(\emptyset) = 0$.
2. If $A_1 \subseteq A_2$, then $P(A_1) \leq P(A_2)$.
3. $P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} P(A_i)$ for any combination of sets $A_1, \ldots, A_n$.
4. For any $A \in F$, we have $0 \leq P(A) \leq 1$.

3 Random variables

3.1 What is a random variable?

Let's suppose we have some experiment, but we don't know its outcome x yet. We can then define x as a **random variable**. If some **event** ω_i occurs, x will have the value $x(\omega_i)$, whereas if some other event ω_j occurs, x will have the value $x(\omega_j)$. The events ω_i are part of the **event space** Ω. x is thus a function from Ω to $\mathbb{R}$ ($x : \Omega \to \mathbb{R}$).

A rather basic example of a random variable is the **indicator function**. The indicator function $I_A(\omega)$ of a subset $A \in \Omega$ is defined as

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A. \end{cases} \tag{3.1}$$

A **simple** random variable is a finite linear combination of indicator functions of measurable sets. In other words, if we have a certain combination of sets $A_1, \ldots, A_n \in F$, then the random variable

$$x = \sum_{i=1}^{n} c_i I_{A_i} \tag{3.2}$$

is a simple random variable.

3.2 PDFs and σ-algebras of a random variable

Every random variable x has a PDF $f_x(u)$ attached to it. Generally speaking, the PDF $f_x(u)$ is the probability that $x(\omega) < u$. If we combine this with our knowledge on probability spaces, we find that

$$f(u) = P(\{\omega \in \Omega | x(\omega) \leq u\}) = P(A) \qquad \text{with} \qquad A = \{\omega \in \Omega | x(\omega) \leq u\}. \tag{3.3}$$

What does the above equation mean? Well, we first look at all events $\omega \in \Omega$ for which $x(\omega) \leq u$. We denote the set of all these events by A. The value of $f(u)$ now equals the value of the probability measure $P(A)$.

Let's examine a random variable defined on the measurable space (Ω, F). We denote the set of all possible values of x by X. We now say that x takes values in the measurable space (X, G). Here, the set G has a relationship with F. In fact, for every set $A \in G$, we have

$$x^{-1}(A) := \{\omega \in \Omega | x(\omega) \in A\} \in F. \tag{3.4}$$

We can now also define $x^{-1}(G)$, according to

$$x^{-1}(G) := \{x^{-1}(A) | \forall A \in G\}. \tag{3.5}$$

Note that we now must have $x^{-1}(G) \subseteq F$. However, it is not necessarily true that $x^{-1}(G) = F$. But it can be shown that $x^{-1}(G)$ is a σ-algebra. We define this σ-algebra as $F^x := F(x) := x^{-1}(G)$. We say that F^x is the **σ-algebra generated by** x.

3.3 The characteristic function

Consider a random variable x with PDF $f_x(u)$. The **expectation** $E[x]$ of this random variable can now be found using

$$E[x] = \int_{-\infty}^{\infty} v p_x(v)\, dv \quad \text{(for continuous)} \qquad \text{and} \qquad E[x] = \sum v_n p_x(v_n) \quad \text{(for discrete)}. \tag{3.6}$$

The function $E[.]$ is called the **expectation function**. We use it to define the **characteristic function** $c_x : \mathbb{R}^n \to \mathbb{C}$ of a random variable x, according to

$$c_x(u) = E[e^{iu^T x}] = \int_{-\infty}^{\infty} e^{iuv} p(v)\, dv. \tag{3.7}$$

In the above equation, $i = \sqrt{-1}$ denotes the complex variable. The characteristic function is quite convenient. If you have it, you can find the corresponding PDF, and vice versa.

3.4 Gaussian random variables

Previously, we have seen the PDF of a Gaussian distribution. Any random variable $x : \Omega \to \mathbb{R}^n$ with such a PDF is called a **Gaussian random variable** with parameters m and Q. This is denoted by $x \in G(m, Q)$. The characteristic function of x has the form

$$c_x(u) = E[e^{iu^T x}] = e^{iu^T m - \frac{1}{2} u^T Q u}. \tag{3.8}$$

Let's examine several Gaussian random variables $x_1, \ldots, x_n$. We can put them together in a vector $x^T = [x_1^T \; \ldots \; x_n^T]$. If the new random vector x is also Gaussian (thus satisfying $x \in G(m, Q)$ for some m, Q), then we say that $x_1, \ldots, x_n$ are **jointly Gaussian**.

Gaussian random variables have several nice properties. Let's examine a few.

- Every linear combination $y = Ax + b$ of a Gaussian random variable is also a Gaussian random variable. In fact, if $x \in G(m, Q)$, then $y \in G(Am + b, AQA^T)$.
- Let's examine two jointly Gaussian random variables x and y. We now have

$$\begin{bmatrix} x \\ y \end{bmatrix} \in G\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{bmatrix} \right), \qquad \text{where} \qquad Q_{xy} = E\left[(x - m_x)(y - m_y)^T \right] = Q_{yx}^T. \tag{3.9}$$

  If $Q_{xy} = 0$, then F^x and F^y are independent, and vice versa. In other words, when Gaussian random variables are uncorrelated, they are also independent, and vice versa.
- Independent Gaussian random variables are always jointly Gaussian. (The converse is of course not always true.)
- If $y \in G(m, Q)$ and $S = S^T$, then $E[y^T S y] = \text{tr}(SQ) + m^T S m$. (The function tr(.) is the trace of the matrix: the sum of the diagonal elements.)

4 Conditional expectation

4.1 Properties of conditional expectation

Let's examine a measurable space (Ω, F). Also examine a sub-σ-algebra G of F. We now define the **conditional expectation** of x given G, denoted by $E[x|G]$, as the random variable $E[x|G]$ that is both G measurable and satisfies

$$E[xI_A] = E[E[x|G] I_A] \tag{4.1}$$

for every set $A \in G$. By the way, the random variable $E[x|G](\omega)$ is G measurable if

$$\{\omega \in \Omega | E[x|G](\omega) \leq r\} \in G \quad \text{for all } r \in \mathbb{R}. \tag{4.2}$$

There are several properties of the conditional expectation. We will examine a few.

- Let's examine two random variables x and y that are integrable. (This means that $E[|x|]$ and $E[|y|]$ are finite.) Also suppose that we can write y as

$$y = \sum_{k=1}^{n} c_k I_{A_k}, \tag{4.3}$$

where $A_1, \ldots, A_n$ is a finite partition of Ω. (In other words, the sets $A_1, \ldots, A_n$ are disjoint, but their union equals Ω.) It can now be shown that

$$E[x|F^y] = \sum_{k=1}^{n} d_k I_{A_k} \qquad \text{where} \qquad d_k = \frac{E[xI_{A_k}]}{E[I_{A_k}]}. \tag{4.4}$$

- Let's examine two jointly Gaussian random variables x and y. Assume that $Q_y > 0$. We now have

$$E[x|F^y] = m_x + Q_{xy}Q_y^{-1}(y - m_y), \tag{4.5}$$

$$E[(x - E[x|F^y])(x - E[x|F^y])^T|F^y] = E[(x - E[x|F^y])(x - E[x|F^y])^T] = Q_x - Q_{xy}Q_y^{-1}Q_{xy}^T, \tag{4.6}$$

$$E[e^{iu^Tx}|F^y] = e^{iu^TE[x|F^y] - \frac{1}{2}u^T\tilde{Q}u} \qquad \text{for all } u \in \mathbb{R}^n, \tag{4.7}$$

$$E[e^{iu^TE[x|F^y]}] = e^{iu^Tm_x - \frac{1}{2}u^TQ_{xy}Q_y^{-1}Q_{xy}^Tu} \qquad \text{for all } u \in \mathbb{R}^n. \tag{4.8}$$

In the above equations, we have used the definition $\tilde{Q} := Q_x - Q_{xy}Q_y^{-1}Q_{xy}^T$.

- Conditional expectation is linear. So,

$$E[c_1x_1 + x_2x_2|G] = c_1E[x_1|G] + c_2E[x_2|G]. \tag{4.9}$$

- If $x \leq y$ for all $\omega \in \Omega$, then $E[x|G] \leq E[y|G]$.
- If y is G measurable, then $E[y|G] = y$.
- If $G_1 \subseteq G_2$, then $E[x|G_1] = E[E[x|G_2]|G_1]$. In particular, if we set $G_1 = \{\emptyset, \Omega\}$ and simply write $G_2 = G$, then this reduces to $E[E[x|G]] = E[x]$.
- If F^x and G are independent sub-σ-algebras (with respect to P), then $E[x|G] = E[x]$. Also, F^x and G are independent if and only if for all $u \in \mathbb{R}$, we have $E[e^{iu^Tx}|G] = E[e^{iu^Tx}]$.

4.2 Independence and conditional independence

Let's consider two σ-algebras F_1 and F_2. We say that F_1 and F_2 are independent if $E[x_1x_2] = E[x_1]E[x_2]$ for all $x_1, x_2 : \Omega \to \mathbb{R}$ for which F_1 and F_2 are σ-algebras, respectively.

We can extend this idea to conditional expectations. We say that F_1 and F_2 are **conditionally independent**, given a sub-σ-algebra G, if

$$E[x_1x_2|G] = E[x_1|G]E[x_2|G] \tag{4.10}$$

for all x_1, x_2 with the same conditions as stated earlier. We generally denote this conditional independence by $(F_1, F_2|G) \in CI$. Conditional independence has several properties. In fact, the following four statements are equivalent:

$$(F_1, F_2|G) \in CI, \qquad (F_2, F_1|G) \in CI, \qquad (F_1 \vee G, F_2 \vee G|G) \in CI, \tag{4.11}$$

$$E[x_1|F_2 \vee G] = E[x_1|G] \quad \text{for all } x_1 \text{ with } F_1 \text{ as } \sigma\text{-algebra}. \tag{4.12}$$

Also, if F_1 and $(F_2 \vee G)$ are independent, then also $(F_1, F_2|G) \in CI$.

We can ask ourselves, when are Gaussian random variables conditionally independent? Well, let's consider Gaussian random variables x, y_1 and y_2 with $Q_x > 0$. It can be shown that $(F^{y_1}, F^{y_2}|F^x) \in CI$ if and only if

$$Q_{y_1y_2} = Q_{y_1x}Q_x^{-1}Q_{xy_2}. \tag{4.13}$$