**Objective:**
This course introduces the concept and basics of information theory and the basics of source and channel encoding/decoding. Understand the key modules of digital communication system. Also, it introduces the meaning of entropy, self and mutual Information. It supports understanding and practicing, design of source encoding, design of the channel encoding and decoding.

**Syllabus:**

<table>
<thead>
<tr>
<th>Event</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Course outline and Review of probability theory - Concept of information theory and coding.</td>
</tr>
<tr>
<td>2</td>
<td>Average information &amp; Entropy – Joint Entropy and Conditional Entropy – Self and Mutual information</td>
</tr>
<tr>
<td>3</td>
<td>Channel capacity - Time rate of information - Capacity of the Binary Symmetric Channel</td>
</tr>
<tr>
<td>4</td>
<td>Continuous channels - Entropy of continuous sources - Entropy maximization - Gaussian sources - Shannon’s formula for capacity.</td>
</tr>
<tr>
<td>5</td>
<td>Bandwidth and S/N of a channel -- Shannon-Hartely theorem of the capacity of AWGN channels.</td>
</tr>
<tr>
<td>6</td>
<td>Source Coding: Universal Codes and Arithmetic Coding - Lossy and Lossless coding - Run Length Coding (RLC).</td>
</tr>
<tr>
<td>7</td>
<td>Variable Length Coding- Huffman coding, Fano coding and shannon coding.</td>
</tr>
<tr>
<td>8</td>
<td>Channel Coding Theorem: Preview, Definitions, and Jointly typical sequences.</td>
</tr>
<tr>
<td>11</td>
<td>Convolutional codes- State diagram-Terrllis diagram.</td>
</tr>
<tr>
<td>12</td>
<td>Viterbi decoding.</td>
</tr>
<tr>
<td>13</td>
<td>Viterbi decoding (Continue) – MAP decoding of Convolutional Codes.</td>
</tr>
<tr>
<td>14</td>
<td>Recursive Systematic Convolutional Codes (RSC)</td>
</tr>
<tr>
<td>15</td>
<td>Turbo Codes. Iterative decoding.</td>
</tr>
<tr>
<td>16</td>
<td>Turbo Codes. Iterative decoding.</td>
</tr>
<tr>
<td>17</td>
<td>Performance of different coded modulation in AWGN channels.</td>
</tr>
<tr>
<td>18</td>
<td>Selected Communication Systems</td>
</tr>
</tbody>
</table>
References:
- M. Schwartz, “Information, transmission, modulation,.....”
- Stremler, “Introduction to communication systems”
- B. P. Lathi, “modern analog and digital communication systems”
- B. Sklar, “Digital communications fundamentals and applications”
- K. Sam, “Digital Communications”
- G. J. Miao, “Signal processing in digital Communications”
- I. Glover, P. Grant, “Digital communications”
Block Schematic Description of a Digital Communication System

In the simplest form, a transmission-reception system is a three-block system, consisting of a) a transmitter, b) a transmission medium and c) a receiver. If we think of a combination of the transmission device and reception device in the form of a ‘transceiver’ and if (as is usually the case) the transmission medium allows signal both ways, we are in a position to think of a both-way (bi-directional) communication system. For ease of description, we will discuss about a one-way transmission-reception system with the implicit assumption that, once understood, the ideas can be utilized for developing / analyzing two-way communication systems. So, our representative communication system, in a simple form, again consists of three different entities, viz. a transmitter, a communication channel and a receiver.

A digital communication system has several distinguishing features when compared with an analog communication system. Both analog (such as voice signal) and digital signals (such as data generated by computers) can be communicated over a digital transmission system. When the signal is analog in nature, an equivalent discrete-time-discrete-amplitude representation is possible after the initial processing of sampling and quantization. So, both a digital signal and a quantized analog signal are of similar type, i.e. discrete-time-discrete-amplitude signals.

A key feature of a digital communication system is that a sense of ‘information’, with appropriate unit of measure, is associated with such signals. This visualization, credited to Claude E. Shannon, leads to several interesting schematic description of a digital communication system. For example, consider Fig.1. Which shows the signal source at the transmission end as an equivalent ‘Information Source’ and the receiving user as an ‘Information sink’? The overall purpose of the digital communication system is ‘to collect information from the source and carry out necessary electronic signal processing such that the information can be delivered to the end user (information sink) with acceptable quality’. One may take note of the compromising phrase ‘acceptable quality’ and wonder why a digital transmission system should not deliver exactly the same information to the sink as accepted from the source. A broad and general answer to such query at this point is: well, it depends on the designer’s understanding of the ‘channel’ (Fig. 1) and how the designer can translate his knowledge to design the electronic signal processing algorithms / techniques in the 'Encoder' and ‘decoder’ blocks in Fig. 1. We hope to pick up a few basic yet good approaches to acquire the above skills. However, pioneering work in the 1940-s and 1950-s have established a bottom-line to the search for ‘a flawless (equivalently, ‘error-less’) digital communication system’ bringing out several profound theorems (which now go in the name of Information Theory) to establish that, while error-less transmission of information can never be
guaranteed, any other ‘acceptable quality’, arbitrarily close to error-less transmission may be possible. This ‘possibility’ of almost error-less information transmission has driven significant research over the last five decades in multiple related areas such as, a) digital modulation schemes, b) error control techniques, c) optimum receiver design, d) modeling and characterization of channel and so forth. As a result, varieties of digital communication systems have been designed and put to use over the years and the overall performance have improved significantly.

**Fig. 1** Basic block diagram of a digital communication System

It is possible to expand our basic ‘three-entity’ description of a digital communication system in multiple ways. For example, **Fig.2** shows a somewhat elaborate block diagram explicitly showing the important processes of ‘modulation-demodulation’, ‘source coding-decoding’ and ‘channel encoding – decoding’. A reader may have multiple queries relating to this kind of abstraction. For example, when ‘information’ has to be sent over a large distance, it is a common knowledge that the signal should be amplified in terms of power and then launched into the physical transmission medium. Diagrams of the type in **Figs. 1** and **2** have no explicit reference to such issues. However, the issue here is more of suitable representation of a system for clarity rather than a module-by-module replication of an operational digital communication system.
To elaborate this potentially useful style of representation, let us note that we have hardly discussed about the third entity of our model. the ‘channel’. One can define several types of channel. For example, the ‘channel’ in Fig.2 should more appropriately be called as a ‘modulation channel’ with an understanding that the actual transmission medium (called ‘physical channel’), any electromagnetic (or otherwise) transmission- reception operations, amplifiers at the transmission and reception ends and any other necessary signal processing units are combined together to form this ‘modulation channel’.

- **Source of information:**
  Analog: the same as sine wave or continuous.
  Digital: sampled data such that the values of the signal are being digitized by 0 or 1.

- **Source coding:**
  Input to this block may be one of the following 6:
  1- Analog
  2- Discrete
  3- Digital
  4- Sampled data
  5- Continuous
  6- Quantized
But the output signal of this stage should be digital.
The question is why digital?!
1- Better maintenance in digital than in analog, because the tolerance of
digital components is much more better than the analog.
2- Cheaper.
3- Adapting (changing the response due to the noise interface using
computer program).
4- Programmable and portable.
5- Higher security protection because the interchangeability of code is much
more flexible than in analog.
“Analog waves or signals suffer difficulties in coding and that channel
coding is available only in digital system”.

- **Modulation**
  Advantages:
  1- Reach the wave or signal to move away possible point.
  2- Reduces the size of the using antenna.
  3- Transfer more frequencies on the channel at the same time by using
multiplexing.

- **Types of the channels:**
  1- Twin wire
  2- Coaxial cable
  3- Fiber optics
  4- Satellite
  5- Microwaves

In the following, we introduce a few short tables, which may help a reader
to recapitulate some relevant issues of electrical communications. **Table 1** lists
some of the important events which have contributed to the developments in
electrical communication. **Table 2** presents different frequency bands with
typical applications that are commonly used for the purpose of electrical
communications. This table is very useful for our subsequent lessons. **Table 3**
mentions frequency ranges for a few popular broadcast and communication
services. **Table 4** gives an idea of typical centre frequencies and the nominal
bandwidths that are available for five frequency bands. It is important to note
that larger bandwidths are available when the operating frequency bands are
higher. **Table 5** provides an idea of typical power losses of several physical
transmission media at representative operating frequency. It may be noted that
all transmission media are not equally suitable at all frequencies. An important
factor other than the power loss in a physical medium is its cost per unit length.
<table>
<thead>
<tr>
<th>Year / Period</th>
<th>Achievements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1838</td>
<td>Samuel F. B. Morse demonstrated the technique of telegraph</td>
</tr>
<tr>
<td>1876</td>
<td>Alexander Graham Bell invents telephone</td>
</tr>
<tr>
<td>1897</td>
<td>Guglielmo Marconi patents wireless telegraph system. A few years earlier, Sir J. C. Bose demonstrated the working principle of electromagnetic radiation using a ‘solid state coherer’</td>
</tr>
<tr>
<td>1918</td>
<td>B. H. Armstrong develops super heterodyne radio receiver</td>
</tr>
<tr>
<td>1931</td>
<td>Teletype service introduced</td>
</tr>
<tr>
<td>1933</td>
<td>Analog frequency modulation invented by Edwin Armstrong</td>
</tr>
<tr>
<td>1937</td>
<td>Alec Reeves suggests pulse code modulation (PCM)</td>
</tr>
<tr>
<td>1948-49</td>
<td>Claude E. Shannon publishes seminal papers on ‘A Mathematical Theory of Communications’</td>
</tr>
<tr>
<td>1956</td>
<td>First transoceanic telephone cable launched successfully</td>
</tr>
<tr>
<td>1960</td>
<td>Development of Laser</td>
</tr>
<tr>
<td>1962</td>
<td>Telstar 1, first satellite for active communication, launched successfully</td>
</tr>
<tr>
<td>1970-80</td>
<td>Fast developments in microprocessors and other digital integrated circuits made high bit rate digital processing and transmission possible; commercial geostationary satellites started carrying digital speech, wide area computer communication networks started appearing, optical fibers were deployed for carrying information through light., deep space probing yielded high quality pictures of planets</td>
</tr>
<tr>
<td>1980-90</td>
<td>Local area networks (LAN) making speedy inter-computer data transmission became widely available; Cellular telephone systems came into use. Many new applications of wireless technology opened up remarkable scopes in business automation.</td>
</tr>
<tr>
<td>1990-2000</td>
<td>Several new concepts and standards in data network, such as, wireless LAN (WLAN), AdHoc networks, personal area networks (PAN), sensor networks are under consideration for a myriad of potential applications.</td>
</tr>
</tbody>
</table>

Table 1 Some milestones in the history of electrical communications
Table 2. Electromagnetic bands with typical applications

Any radio operation at 1GHz or beyond (upto several tens of GHz) is also termed as ‘microwave’ operation.

<table>
<thead>
<tr>
<th>Frequency Band</th>
<th>Wavelength</th>
<th>Name</th>
<th>Transmission Media</th>
<th>Some Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 – 30 KHz</td>
<td>100–10 Km</td>
<td>Very Low Frequency (VLF)</td>
<td>Air, water, copper cable</td>
<td>Navigation, SONAR</td>
</tr>
<tr>
<td>30–300 KHz</td>
<td>10 Km–1 Km</td>
<td>Low Frequency (LF)</td>
<td>Air, water, copper cable</td>
<td>Radio beacons, Ground wave communication</td>
</tr>
<tr>
<td>300KHz – 3 MHz</td>
<td>1 Km – 100 m</td>
<td>Medium Frequency (MF)</td>
<td>Air, copper cable</td>
<td>AM radio, navigation, Ground wave communication</td>
</tr>
<tr>
<td>3 MHz – 30 MHz</td>
<td>100 m–10 m</td>
<td>High Frequency (HF)</td>
<td>Air, copper and coaxial cables</td>
<td>HF communication, Citizen’s Band (CB) radio, ionosphere communication</td>
</tr>
<tr>
<td>30MHz – 300 MHz</td>
<td>10 m – 1 m</td>
<td>Very High Frequency (VHF)</td>
<td>Air, free space, coaxial cable</td>
<td>Television, Commercial FM broadcasting, point to point terrestrial communication</td>
</tr>
<tr>
<td>300 MHz – 3 GHz</td>
<td>1 m–10 cm</td>
<td>Ultra High Frequency (UHF)</td>
<td>Air, free space, waveguide</td>
<td>Television, mobile telephones, satellite communications,</td>
</tr>
<tr>
<td>3GHz – 30 GHz</td>
<td>10 cm–1 cm</td>
<td>Super / Extra High Frequency (SHF / EHF)</td>
<td>Air, free space, waveguide</td>
<td>Satellite communications, wireless LAN, Metropolitan Area network (WMAN), Ultra Wideband communication over a short distance</td>
</tr>
<tr>
<td>30 GHz – 300 GHz</td>
<td>1 cm – 1 mm</td>
<td></td>
<td></td>
<td>Mostly at experimental stage</td>
</tr>
<tr>
<td>30 Tera Hz – 3000 Tera Hz</td>
<td>10 μm – 0.1μm (approx)</td>
<td>Optical</td>
<td>Optical fiber</td>
<td>Fiber optic communications</td>
</tr>
</tbody>
</table>
Table .3  A few popular frequency bands

<table>
<thead>
<tr>
<th>Name / Description</th>
<th>Frequency Range</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Broadcast Radio</td>
<td>540 KHz – 1600 KHz</td>
<td>Commercial audio broadcasting using amplitude modulation</td>
</tr>
<tr>
<td>FM Broadcast Radio</td>
<td>88 MHz – 108 MHz</td>
<td>Commercial audio broadcasting using frequency modulation</td>
</tr>
<tr>
<td>Cellular Telephony</td>
<td>806 MHz – 940 MHz</td>
<td>Mobile telephone communication systems</td>
</tr>
<tr>
<td>Cellular Telephony and Personal Communication Systems (PCS)</td>
<td>1.8 GHz – 2.0 GHz</td>
<td>Mobile telephone communication systems</td>
</tr>
<tr>
<td>ISM (Industrial Scientific and Medical) Band</td>
<td>2.4 GHz – 2.4835 GHz</td>
<td>Unlicensed band for use at low transmission power</td>
</tr>
<tr>
<td>WLAN (Wireless Local Area Network)</td>
<td>2.4 GHz band and 5.5 GHz</td>
<td>Two unlicensed bands are used for establishing high speed data network among willing computers</td>
</tr>
<tr>
<td>UWB (Ultra Wide Band)</td>
<td>3.7 GHz – 10.5 GHz</td>
<td>Emerging new standard for short distance wireless communication at a very high bit rate (typically, 100 Mbps)</td>
</tr>
</tbody>
</table>

Table .4  Some Carrier frequency values and nominal bandwidth that may be available at the carrier frequency

<table>
<thead>
<tr>
<th>Frequency band</th>
<th>Carrier frequency</th>
<th>Approx. Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long wave Radio [LF]</td>
<td>100KHz</td>
<td>~ 2 KHz</td>
</tr>
<tr>
<td>Short wave [HF]</td>
<td>5MHz</td>
<td>100KHz</td>
</tr>
<tr>
<td>VHF</td>
<td>100MHz</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Micro wave</td>
<td>5GHz</td>
<td>100 MHz</td>
</tr>
<tr>
<td>Optical</td>
<td>$5 \times 10^{14}$ Hz</td>
<td>10 GHz – 10 THz</td>
</tr>
</tbody>
</table>

Table .5  Typical power losses during transmission through a few media

<table>
<thead>
<tr>
<th>Transmission medium</th>
<th>Frequency</th>
<th>Power loss in [dB/km]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twisted copper wire [16 AWG]</td>
<td>1 KHz</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>100KHz</td>
<td>3.0</td>
</tr>
<tr>
<td>Co-Axial Cable [1cm dia.]</td>
<td>100KHz</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>3 MHz</td>
<td>4.0</td>
</tr>
<tr>
<td>Wave Guide</td>
<td>10 GHz</td>
<td>1.5</td>
</tr>
<tr>
<td>Optical Fiber</td>
<td>$10^{14} – 10^{16}$ Hz</td>
<td>&lt;0.5</td>
</tr>
</tbody>
</table>
**Information Theory:**

**Def:** Information theory is a subject that deals with information and data transmission from one point to another. The concept of information is related to probability. Any signal that conveys information must be unpredictable (random), but not vice versa, i.e. not any random signal conveys information. (Noise is a random signal conveying no information).

* What’s information?
  - Message
  - Article
  - Painting
  - Music
  - ......

For different purpose you can have different definitions for Information.

- C.E. Shannon developed an entropy theory for information, orienting at data storage and transmission:

  **Information means removal of uncertainty.**

Namely, information means how much uncertainty you should remove to make sure of some particular issue.

- For example, in a coin toss, we have the probability \( p(H) = 50\% \) to get a 'Head' and \( p(T) = 50\% \) to get a 'Tail'. That means if we want an 'H' in a toss, the possibility of success is 50\%, or, the uncertainty is 50\%. Thus, we say: if our aim is undoubtedly to obtain an 'H', we should remove the 50\% uncertainty. This ‘removal’ gives the required information.

The details for how to realize this idea will be given later.

- This theory has proven to be a successful and powerful tool with applications in many areas, not only in IT, but also in gambling theory and strategy of financial market.

**Self Information:**

Suppose that the source of information produces finite set of messages \( x_1, x_2, x_3 \ldots x_n \) with probabilities \( P(x1), p(x2), p(x3), \ldots p(xn) \) and such that \( \sum_{i=1}^{n} p(X_i) = 1 \). The amount of information gained by knowing that the source produces the message \( x_i \) is related with \( p(xi) \) as following:
1- Information is zero if \( p(x_i) = 1 \) (certain event).

2- Information increases as \( p(x_i) \) decreases to zero.

3- Information is +ve quantity.

The function that relates \( p(x_i) \) with information of \( x_i \) is denoted by \( I(x_i) \) and is called self information of \( x_i \).

The log function shown satisfies all previous three points hence:

\[
I(x_i) = - \log_a p(x_i)
\]

1- If "a" = 2 (this is mostly used in digital communication) then \( I(x_i) \) has the units of bits.

2- If "a" = e = 2.71828, then \( I(x_i) \) has the units of nets.

3- If "a" = 10, then \( I(x_i) \) has the units of decade.

Recall that \( \log_a x = \frac{\ln x}{\ln a} \)
• Let $X$ be a probability space (space of all samples), and $x \in X$ a sample. [E.g., drawing red/white balls from a bag forms $X$ and a try gives a result $x$.]

**Examples:**
- Coin toss: H or T. So the probability space $X = \{H, T\}$, and a sample $x$ can be $x = H$ or $x = T$;
- Throwing 6-sided dices: 1, 2, 3, 4, 5, 6. So $X = \{1, 2, 3, 4, 5, 6\}$, and $x = 1, 2, 3, 4, 5$ or 6.
- Other examples, such as weather: (rain, warm), (rain, cool), (dry, warm) or (dry, cool)......

• $p(A)$: Let $A \subseteq X$ be the probability of an event $A$. Then the probability of the event $A$ is
  \[ p(A) = \sum_{x \in A} p(X) \]  
  (1)

  Apparently: $p(X) = 1$, $p(\emptyset) = 0$, and $0 \leq p(A) \leq 1$.

**Example** (dice-throwing): $X = \{1, 2, 3, 4, 5, 6\}$, and $x \in X$. $A$ can be:
- $A = \{x \mid x \leq 4\}$, i.e. $A = \{x \mid x = 1, 2, 3, 4\}$
- or $A = \{x \text{ odd}\}$, i.e. $A = \{x \mid x = 1, 3, 5\}$

• $p(A + B)$: If two events $A$ and $B$ are disjoint (i.e., mutually exclusive), then
  \[ p(A + B) = p(A) + p(B), \quad A, B \in X. \]  
  (2)

**Example:** For $A = \{1, 2\}$, $B = \{5, 6\}$, there is $A + B = \{1, 2, 5, 6\}$.

• $p(AB)$: The probability of “both the events $A$ and $B$ happen” $p(AB)$ can be understood as ‘$p(A \cap B)$’.
  (3)

**Example:** For $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$, there is $A \cap B = \{3, 4\}$.

• $p(A \cup B)$: If $A$ and $B$ are just two events in $X$ (disjoint or not), then
  \[ p(A \cup B) = p(A) + p(B) - p(AB). \]  
  (4)

  Apparently: $p(A \cup B) \leq p(A) + p(B)$, because $0 \leq p(AB) \leq 1$.

**Example:**
For $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$, there is $A \cup B = \{1, 2, 3, 4, 5, 6\}$.
• Calculation of $p(x)$:

$$p(x) = \frac{N(x)}{N}$$  \hspace{1cm} (5)

Where $N(x)$ denotes the frequencies (number of times) of “$x$ happens”, and $N$ the number of all possible outcomes. Consequently

$$p(A) = \frac{N(A)}{N} = \sum_{x \in A} \frac{N(x)}{N} = \sum_{x \in A} p(x).$$ \hspace{1cm} (6)

**Example:** In coin toss, suppose we do 10 times of throw, and get 6 H’s, then $p(H) = 6/10 = 0.60$. If we do 100 times of throw and get 55 H’s, then $p(H) = 55/100 = 0.55$. If we do 1000 times of throw and get 510 H’s, then $p(H) = 510/1000 = 0.51$. The limit of $p(H)$ is 0.5 when the times of throws go to infinity.

• Let $(X, p)$ be a probability space. Let $B$ be a subset of $X$ such that $p(B) \neq 0$. Consider an experiment where outcomes that are not in $B$ do not count. Then for each $x \in X$ the conditional probability of $x$ given $B$ is defined as $p(x|B)$:

$$p(x|B) = \begin{cases} \frac{p(x)}{p(B)}, & \text{if } x \in B, \\ 0, & \text{if } x \notin B. \end{cases}$$  \hspace{1cm} (7)

It is seen that $\sum_{x \in X} p(x|B) = 1$, because

$$\sum_{x \in X} p(x|B) = \sum_{x \in B} p(x|B) + \sum_{x \notin B} p(x|B) = \sum_{x \in B} \frac{p(x)}{p(B)} = \frac{p(B)}{p(B)} = 1.$$

This means that the sum of all the conditional probabilities on the same condition $B$ is 1.

• $p(A|B)$:

$$p(A|B) = \frac{p(AB)}{p(B)} \hspace{1cm} \text{[regarded as } p(A|B) = \frac{p(A \cap B)}{B} \text{]}.$$ \hspace{1cm} (8)

Understanding:

$$p(A|B) = \sum_{x \in A \cap B} p(x|B) + \sum_{x \notin A \cap B} p(x|B) = \sum_{x \in A \cap B} \frac{p(x)}{p(B)} + 0 = \frac{p(AB)}{p(B)}.$$ \hspace{1cm} (9)

One can understand “conditional probability” as: the probability of “$A$ happens” under the condition of “$B$ happens”. 

**Example:** (example of weather for conditional probability)

<table>
<thead>
<tr>
<th></th>
<th>Warm</th>
<th>Cool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>Dry</td>
<td>0.70</td>
<td>0.15</td>
</tr>
</tbody>
</table>

(10)

Apparently,

\[ p(\text{Cool}) = p(\text{Rain and Cool}) + p(\text{Dry and Cool}) = 0.10 + 0.15 = 0.25. \]

\[ p(\text{Dry}) = p(\text{Dry and Warm}) + p(\text{Dry and Cool}) = 0.70 + 0.15 = 0.85 \]

But

\[ p(\text{Dry|Cool}) = \frac{p(\text{Dry and Cool})}{p(\text{Cool})} = \frac{0.15}{0.25} = 0.60 < 0.85 = p(\text{Dry}), \]

which means that, when the temperature is low, it is more unlikely to be dry. Similarly,

\[ p(\text{Warm}) = p(\text{Rain and Warm}) + p(\text{Dry and Warm}) = 0.05 + 0.70 = 0.75, \]

\[ p(\text{Dry}) = p(\text{Dry and Warm}) + p(\text{Dry and Cool}) = 0.70 + 0.15 = 0.85, \]

\[ p(\text{Dry|Warm}) = \frac{p(\text{Dry and Warm})}{p(\text{Warm})} = \frac{0.70}{0.75} = 0.93 > 0.85 = p(\text{Dry}), \]

saying that when the temperature is high it is more likely to be dry. Hence, temperature affects humidity.

Furthermore, we can discuss a special case of conditional probability — independence.

- **Independence:** If

\[ p(A|B) = p(A), \]

(11)

then the events \(A\) and \(B\) are called independent. An equivalent definition for independence of \(A\) and \(B\) is

\[ p(AB) = p(A \cap B) = p(A) p(B), \]

(12)

because it leads to (11):

\[ p(A|B) = \frac{p(A) p(B)}{p(B)} = p(A). \]

(13)

One can understand “independent events \(A\) and \(B\)” in conditional probability as: the condition of “\(B\) happens” does not affect the happening of \(A\), hence

\[ p(A|B) = p(A). \]

On the contrary, if the happening of \(B\) makes the happening of \(A\) easier/harder, then the events \(A\) and \(B\) are not independent.
**Combined Experiment:**

- In this section we will develop further understanding for \( p(A|B) = \frac{p(AB)}{p(B)} \). The key idea is to focus on individual points \( x \in A \subseteq X \), \( y \in B \subseteq Y \), and write out \( p(A|B) \) with \( p(x|B) \) and \( p(x|y) \).

- Let us check out some typical examples first:
  1. Coin toss (twice): Each toss has a result H or T. Hence \( X = \{H,T\} \) and \( Y = \{H,T\} \). Combining these two experiments together, one has a joint experiment with the results \( (x, y) = \{HH, HT, TH, TT\} \), \( x \in X \), \( y \in Y \). This result can also be expressed as

        | H | T |
        |---|---|
        | H | HH |
        |   | TT |
        | T | TH |

  2. Weather: For 4 typical days we test/classify them with two experiments with different criteria: humidity and temperature. These two experiments form a joint experiment, with the result:

        | Warm | Cool |
        |-----|------|
        | Rain | R.W. |
        | Dry  | D.W. |
        |      | R.C. |
        |      | D.C. |

D.W., D.C.}, \( x \in \) Humidity, \( y \in \) Temperature.

In these examples, we regard \( (x, y) \in X \times Y \) as one outcome of a joint experiment, then this outcome has a probability named joint probability distribution.

- \( p(x, y) \) — Joint probability distribution (or product distribution) for \( X, Y \):
  - simultaneously choosing \( x \) from \( X \) and \( y \) from \( Y \). Apparently
  
  \[
  \sum_{x, y} p(x, y) = 1
  \]

- For \( p(x, y) \), if summing up \( y \) and \( x \) respectively, we obtained the so-called marginal distributions.

- \( p(x) \) and \( p(y) \) — Marginal distributions:
  
  \[
  p(x) = \sum_{y \in Y} p(x, y), \quad p(y) = \sum_{x \in X} p(x, y).
  \]

  This is equivalent to the following notation:
  
  \[
  p(x) \equiv p(x, Y), \quad p(y) \equiv p(X, y).
  \]

- In the example of coin toss, one has the following probability distributions

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>T</th>
<th>( p(X) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>( p(HH) = 0.25 )</td>
<td>( p(HT) = 0.25 )</td>
<td>( p(HH) + p(HT) = 0.5 )</td>
</tr>
<tr>
<td>T</td>
<td>( p(TH) = 0.25 )</td>
<td>( p(TT) = 0.25 )</td>
<td>( p(TH) + p(TT) = 0.5 )</td>
</tr>
<tr>
<td>( p(Y) )</td>
<td>( p(HH) + p(TH) = 0.5 )</td>
<td>( p(HT) + p(TT) = 0.5 )</td>
<td>MARGINAL</td>
</tr>
</tbody>
</table>
**Conditional Probability:**

- Let \((x, y) \in X \times Y\) form a joint experiment, and \(x, y \in X \times Y\). One can define the conditional probability \(p(x|y)\) as

\[
p(x|y) = \frac{p(x, y)}{p(y)}.
\]

- For the example of Weather:

<table>
<thead>
<tr>
<th></th>
<th>Warm</th>
<th>Cool</th>
<th>(p(X))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>(p(R.W.) = 0.05)</td>
<td>(p(R.C.) = 0.10)</td>
<td>(p(\text{Rain}) = 0.15)</td>
</tr>
<tr>
<td>Dry</td>
<td>(p(D.W.) = 0.70)</td>
<td>(p(D.C.) = 0.15)</td>
<td>(p(\text{Dry}) = 0.85)</td>
</tr>
<tr>
<td>(p(Y))</td>
<td>(p(\text{Warm}) = 0.75)</td>
<td>(p(\text{Cool}) = 0.25)</td>
<td>MARGINAL</td>
</tr>
</tbody>
</table>

Hence

\[
p(\text{Rain}|\text{Warm}) = \frac{p(R.W.)}{p(W)} = \frac{0.05}{0.75} = 0.07, \quad p(\text{Dry}|\text{Warm}) = \frac{p(D.W.)}{p(W)} = \frac{0.70}{0.75} = 0.93,
\]

\[
p(\text{Rain}|\text{Cool}) = \frac{p(R.C.)}{p(C)} = \frac{0.10}{0.25} = 0.40, \quad p(\text{Dry}|\text{Cool}) = \frac{p(D.C.)}{p(C)} = \frac{0.15}{0.25} = 0.60.
\]

- Now \(p(x|B)\) and \(p(A|y)\), with \(x \in A \subseteq X, y \in B \subseteq Y\), can be given by

\[
p(x|B) = \sum_{y \in B} p(x|y), \quad p(A|y) = \sum_{x \in A} p(x|y).
\]
Independence:

- If specially \( x \) and \( y \) are independent, namely \( p(x|y) = p(x) \), then

\[
p(x|y) = \frac{p(x,y)}{p(y)} = p(x) \quad \Rightarrow \quad p(x,y) = p(x)p(y) \quad \forall x \in X, y \in Y.
\]

\( p(x,y) = p(x)p(y) \) can be regarded as an equivalent definition for the independence \( p(x|y) = p(x) \).

When \( A \) and \( B \) independent, one has

\[
p(AB) = \sum_{x \in A, y \in B} p(x,y) = \sum_{x \in A} \sum_{y \in B} p(x)p(y) = \sum_{x \in A} p(x) \sum_{y \in B} p(y) = p(A)p(B).
\]

\[
p(A,B) = p(A)p(B) \quad \Leftrightarrow \quad p(A|B) = p(A)
\]

Example: Coin toss

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>T</th>
<th>( p(X) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>( p(HH) = 0.25 )</td>
<td>( p(HT) = 0.25 )</td>
<td>( p(HH) + p(HT) = 0.5 )</td>
</tr>
<tr>
<td>T</td>
<td>( p(TH) = 0.25 )</td>
<td>( p(TT) = 0.25 )</td>
<td>( p(TH) + p(TT) = 0.5 )</td>
</tr>
<tr>
<td>( p(Y) )</td>
<td>( p(HH) + p(TH) = 0.5 )</td>
<td>( p(HT) + p(TT) = 0.5 )</td>
<td>MARGINAL</td>
</tr>
</tbody>
</table>
Example: A fair die is thrown; find the amount of information gained if you are told that 4 will appear.

Solution:
Since fair die than, \( p(1) = p(2) = \cdots = p(6) = \frac{1}{6} \) then:
\[
I(4) = - \log_2 p(4) = - \log_2 \left(\frac{1}{6}\right) = \ln 2 = 2.5849 \text{ bits .}
\]
(Note that if "a" not given then \( a = 2 \))

Example: A biased coin has \( p(\text{Head}) = 0.3 \). Find the amount of information gained if you are told that a tail appears.

Solution:
\[
P(\text{tail}) = 1 - p(\text{Head}) = 1 - 0.3 = 0.7 \text{ then }
\]
\[
I(\text{tail}) = - \log_2 (0.7) = - \frac{\ln 0.7}{\ln 2} = 0.5145 \text{ bits .}
\]