Information Theory - notes

Viviana Arrigoni
Internet of Things 2024-2025
Computer Science Department, Sapienza University of Rome

1 Entropy as the limit of lossless compression

Definition 1.1. Let $t \in (0,1)$. The binary entropy of t is:

$$h(t) = t \log \frac{1}{t} + (1-t) \log \frac{1}{1-t}. \tag{1}$$

The entropy is the measure of how chaotic and unpredictable a system is. The domain of the binary entropy can be extended by continuity to 0 and to 1, where its value is 0. It has a maximum point in 0.5, where its value is 1.

We can extend the definition of entropy to non binary discrete random variables and their probability distributions.

Definition 1.2. Let X be a random variable (R.V), $X \in \mathcal{X} = \{x_1, \dots, x_n\}$, and $\mathbb{P}[X = x_i] = p_i$ $\forall i = 1, \dots, n$. The entropy of X is:

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i). \tag{2}$$

Notice that $H(X) = -\sum_{i=1}^{n} p_i \log(p_i) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$.

Definition 1.3. Let X and Y be two random variables, $X \in \mathcal{X} = \{x_1, \dots, x_n\}$, $Y \in \mathcal{Y} = \{y_1, \dots, y_m\}$. The joint entropy of X and Y is:

$$H(X \wedge Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}[X = x \wedge Y = y] \log \frac{1}{\mathbb{P}[X = x \wedge Y = y]}. \tag{3}$$

Definition 1.4. The information content or surprisal of an event E is:

$$I(E) = \log \frac{1}{\mathbb{P}[E]}. \tag{4}$$

The surprisal of an event grows with the unlikelihood of the event. The entropy of a random variable is the expected value of the surprisal of its possible outcomes.

Example 1.1. Let X be the random variable representing the outcome of a dice roll. Assuming the dice is fair, then each possible outcome has equal probability $\frac{1}{6}$, meaning that X follows the uniform probability distribution. If we roll two dice, a blue and a red one, and if we call R the random variable representing the outcome of the red dice roll, and B the random variable representing the outcome of the blue dice roll, then each possible couple $(R = r, B = b)$ $r, b = 1, \dots, 6$ is equally likely. The number of possible pairs is 36, each with probability $\frac{1}{36}$. Nevertheless, if we call $S = R + B$, then S does not follow the uniform distribution since some values of S are more likely than others. For example, there are 6 possible outcomes such that the sum of the two dice equals 7, whereas only one pair $(R = 6, B = 6)$ results in $S = 12$. The surprisal for $S = 7$ is $I(S = 7) = -\log \frac{6}{36} = \log \frac{1}{6} \sim 2.6$, whereas the surprisal for $S = 12$ is $I(S = 12) = -\log \frac{1}{36} \sim 5.2$.
The entropy of the variable X following the uniform distribution over the outcomes of a dice roll is:

$$H(X) = 6 \cdot \frac{1}{6} \log(6) \sim 2.6. \tag{5}$$

The entropy of the uniform probability distribution is maximal. In fact, if a random variable follows a very skewed distribution, then it is very easy to predict its outcome. Instead, if it follows the uniform distribution, its outcomes are purely random.

1.1 Binary Encodings

Let $\mathcal{X}$ be a finite alphabet. Let $\mathcal{M} = \{\text{words of a language on } \mathcal{X}\}$ and $\mathcal{M}^* = \{\text{sequences of words in} \mathcal{M}\}$. Notice that $|\mathcal{M}| < \infty$, while $|\mathcal{M}^*| = \infty$.

Definition 1.5. A variable length binary encoding is an injective function $f : \mathcal{M} \to \{0,1\}^*$ that assigns a binary string to each word in $\mathcal{M}$.

Definition 1.6. The extension by concatenation of a variable length binary encoding f is $f^* : \mathcal{M}^* \to \{0,1\}^*$ such that for each $\underline{m} = (m_1, \ldots, m_n)$, $m_i \in \mathcal{M}$, $f^*(\underline{m}) = (f(m_1), \ldots, f(m_n))$.

Notice that f^* might be non-injective. For instance if $\mathcal{M} = \{m_1, m_2, m_3\}$, a possible variable length binary encoding f is $f(m_1) = 0$, $f(m_2) = 1$, $f(m_3) = 01$. Notice that f is injective, while f^* is not, in fact: $f^*(m_1 m_2) = f^*(m_3) = 01$.

Definition 1.7. A variable length binary encoding f defined over the set $\mathcal{M}$ is prefix-free if $\forall m, m' \in \mathcal{M}$, $m \neq m'$ it holds that $f(m) \not\prec f(m')$, where $\not\prec$ means "is not a prefix of".

Definition 1.8. Let $\underline{x}$ and $\underline{y}$ be two strings, then $\underline{x} \prec \underline{y}$ ($\underline{x}$ is a prefix of $\underline{y}$) if $\underline{x} = \underline{y}$ or $\exists \underline{z} \in \{0,1\}^*$ such that $\underline{y} = \underline{x}\underline{z}$. For instance, $\underline{x} = 00$ is a prefix of $\underline{y} = 0001$ ($\underline{z} = 01$).

Definition 1.9. A binary encoding f is uniquely decodable (U.D.) if its extension by concatenation f^* is injective.

Observe that a prefix-free encoding is uniquely decodable, whereas the vice-versa is not true in general. For instance, $\mathcal{M} = \{m_1, m_2\}$, $f(m_1) = 0$, $f(m_2) = 01$, f^* is injective but f is not prefix-free. Hence, being prefix-free is a stronger condition than unique decodability.

Why do we like it when an encoding is uniquely decodable?
Let f be the following encoding: $f(m_1) = 00$, $f(m_2) = 01$, $f(m_3) = 000$, $f(m_4) = 1$. If a sender sends the sequence $m_1 m_2$, the receiver sees the string 0001. If the sender sends the sequence $m_3 m_4$, the receiver still sees the string 0001. There is no way the receiver can distinguish the original string that is being sent. This causes ambiguity in the communication. If f^* is injective then the receiver can restrict the codomain of f^*, that is $\{0,1\}^*$, to just the subset of binary strings that are the image of f^*. In this way, f^* is bijective and hence invertible, meaning there is a 1-to-1 mapping between sent and received strings, avoiding ambiguity.

Why do we like it when an encoding is prefix-free?
In some channels, such as the telephone line, communication begins when a valid number is dialled. If the telephone number of user 1 is a prefix of the telephone number of user 2, there is no way to call user 2 because as soon as you dial the first digits of its number, the call to user 1 begins.

Definition 1.10. Let P be any probability distribution over a set of words $\mathcal{M}$, and let $f : \mathcal{M} \to \{0,1\}^*$ be a prefix-free binary encoding. The average length of f with respect to P is:

$$\sum_{m \in \mathcal{M}} P(m)|f(m)| \tag{6}$$

where $|f(m)|$ is the length of the string $f(m)$.

Lemma 1.1. (Kraft's inequality) Let f be a prefix-free binary encoding. Then:

$$\sum_{m \in \mathcal{M}} 2^{-|f(m)|} \leq 1. \tag{7}$$

The intuition behind this result is that prefix-free binary encodings do not encode words to binary sequences that are very short because they are the prefixes of many other strings. This inequality allows us to provide one of the most powerful results in Information Theory.

Theorem 1.2. Let $f : \mathcal{M} \to \{0,1\}^*$ be a prefix-free binary encoding. Let P be a distribution on $\mathcal{M}$. Then, the average length of the encoding f is lower-bounded by the entropy of P:

$$\sum_{m \in \mathcal{M}} P(m)|f(m)| \geq H(P). \tag{8}$$

Proof.

$$\begin{aligned}
&\sum_{m \in \mathcal{M}} P(m)|f(m)| \geq H(P) \\
&\sum_{m \in \mathcal{M}} P(m)|f(m)| - H(P) \geq 0 \\
\text{(definition of entropy)} \quad &\sum_{m \in \mathcal{M}} P(m)|f(m)| - \sum_{m \in \mathcal{M}} P(m) \log \frac{1}{P(m)} = \\
&= \sum_{m \in \mathcal{M}} P(m) \left[|f(m)| - \log \frac{1}{P(m)} \right] = \\
(\log_2 2^a = a) \quad &= \sum_{m \in \mathcal{M}} P(m) \left[\log 2^{|f(m)|} - \log \frac{1}{P(m)} \right] = \\
(\log a - \log b = \log \frac{a}{b}) \quad &= \sum_{m \in \mathcal{M}} P(m) \left[\log(2^{|f(m)|} P(m)) \right] = \\
&= \sum_{m \in \mathcal{M}} P(m) \left[\log \frac{P(m)}{2^{-|f(m)|}} \right] \geq \\
\text{(by log-sum inequality)} \quad &\geq \sum_{m \in \mathcal{M}} P(m) \log \frac{\sum_{m \in \mathcal{M}} P(m)}{\sum_{m \in \mathcal{M}} 2^{-|f(m)|}} \geq \\
\text{(by Kraft's inequality and } \sum_{m \in \mathcal{M}} P(m) = 1) \quad &\geq \log \frac{1}{1} \geq 0.
\end{aligned}$$

Where the log-sum inequality states that if $a = a_1 + \ldots + a_n$ and $b = b_1 + \ldots + b_n$, $a_i, b_j \geq 0$, then:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b} \tag{9}$$

□

The result also holds if we ask for f to be uniquely decodable, not prefix-free. This result says that entropy is the limit to lossless compression. We can compress messages more than the entropy of their probability distribution, but this would result in the loss of part of the meaning of the messages.

Corollary 1.2.1. Equation 8 holds with an equality sign (=) if $f : \mathcal{M} \to \{0,1\}^*$ is such that $P(m) = \frac{1}{2^{-|f(m)|}}$ for each $m \in \mathcal{M}$.

Example 1.2. Assume that your BF lives in Florence, and every day, they send you a message to tell you what the weather is like there. The weather in Florence is sunny (S), foggy (F), rainy (R) and cloudy (C), and all have equal probability $\frac{1}{4}$. If your BF wants to compress the information as much as possible, they need at least 2 bits. A valid prefix-free encoding is $f(S) = 00$, $f(F) = 01$, $f(R) = 10$, $f(C) = 11$. The entropy of this distribution, P, is $H(P) = \frac{1}{4} \cdot 4 \log 4 = 2$. The average length of the encoding f is $4 \cdot \frac{1}{4} \cdot 2 = 2$. Hence, f is maximally compressed, since its average length equals the entropy of P.

Now assume that your BF lives in Milan, where the probability distribution of the weather is $P(S) = \frac{1}{4}$, $P(F) = \frac{1}{2}$, $P(R) = \frac{1}{8}$, $P(C) = \frac{1}{8}$. What is a good choice for this encoding? Remember: most common words should be shorter! Let's study three options:
Option 1. $f(S) = 01$, $f(F) = 1$, $f(R) = 000$, $f(C) = 001$.

Option 2. $f(S) = 000$, $f(F) = 001$, $f(R) = 1$, $f(C) = 01$.
Option 3. $f(S) = 00$, $f(F) = 01$, $f(R) = 10$, $f(C) = 11$.
Notice that all options are prefix-free and well encode the possible messages. Nevertheless, option 1 is the one with the shortest average length, that is $\frac{1}{4}\cdot 2+\frac{1}{2}+\frac{1}{8}\cdot 3+\frac{1}{8}\cdot 3 = \frac{7}{4}$ (the average lengths for option 2 and 3 are $\frac{21}{8}$ and 2, respectively, which are both larger). Furthermore, the average length for the encoding in option 1 is equal to the entropy of P, that is, $H(P) = \frac{1}{4}\log 4+\frac{1}{2}\log 2+\frac{1}{8}\log 8+\frac{1}{8}\log 8 = \frac{7}{4}$.
Option 1 is the best since it assigns the shortest string to the most common messages.

Notice that we do not always achieve a minimal encoding whose average length is exactly equal to the entropy of the probability distribution over the messages $\mathcal{M}$. In fact, if the probability distribution P is defined as $P(S) = \frac{5}{9}$, $P(F) = \frac{1}{9}$, $P(R) = \frac{2}{9}$, $P(C) = \frac{1}{9}$, we cannot find a prefix-free binary encoding more efficient than the following one: $f(S) = 1$, $f(F) = 001$, $f(R) = 01$, $f(C) = 000$, whose average length is $\frac{5}{9} + \frac{2}{9}\cdot 2 + \frac{2}{9}\cdot 3 = \frac{15}{9} = 1.\bar{6}$. The entropy is $H(P) = \frac{5}{9}\log\frac{9}{5} + \frac{2}{9}\log(9) + \frac{2}{9}\log\frac{9}{2} \sim 1.657$, which is strictly smaller than the average length of f.

***

2 Shannon's capacity theorem for DMC

Definition 2.1. Let X, Y be two random variables, $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. The mutual information of X and Y is defined as:

$$I(X,Y) = H(Y) - H(Y|X) \tag{10}$$

where $H(Y|X)$ is the conditional entropy of Y given X:

$$H(Y|X) = \sum_{x\in\mathcal{X}} \mathbb{P}[X = x] \sum_{y\in\mathcal{Y}} \mathbb{P}[Y = y|X = x] \log \frac{1}{\mathbb{P}[Y = y|X = x]} \tag{11}$$

The mutual information is symmetric ($I(X,Y) = I(Y,X)$), is always non-negative and represents the amount of information that the variables X and Y bring about each other. The higher the mutual information, the more X and Y provide information about one another, whereas the mutual information is minimal (=0) when X and Y are independent (since $H(Y|X) = H(Y)$).

We have seen that we cannot compress messages more than their entropy. Nevertheless, in real life, we hardly compress messages to their entropy when we communicate. This happens because the communication channel is usually noisy, and adding some redundant bits guarantees more robust communication. On the one hand, we want to transmit the least number of bits to encode our messages, meaning that we want to compress the information as much as possible (occupying the channel costs, faster communication). At the same time, we want the communication to be reliable, which can be achieved if we add some redundancy in the encoding. In fact, while the channel allows communication, it also introduces noise, which can introduce errors in the sent signal. Shannon formulated a mathematical model to decode the source of a message while limiting communication errors.

2.1 Discrete Memoryless Channels

In a discrete memoryless channel, the sender encodes messages as sequences of symbols in a finite set $\mathcal{X}$ and sends them through the channel to the receiver, who receives a sequence of symbols, possibly belonging to a different alphabet $\mathcal{Y}$, that is still finite. Since the channel is noisy, some symbols can be altered, and the receiver might receive messages that contain some errors. The memoryless property of a channel means that the introduced errors are independent. The errors introduced by the channel on the transmission of one symbol can be modelled as a stochastic matrix $W(Y|X)$, where X is the sent signal and Y is the received signal. The noise matrix is defined as follows:

$$W(Y|X) = \begin{pmatrix} W(y_1|x_1) & \cdots & W(y_{|\mathcal{Y}|}|x_1) \\ \vdots & \ldots & \vdots \\ W(y_1|x_{|\mathcal{X}|}) & \cdots & W(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|}) \end{pmatrix} \tag{12}$$

Where $W(Y = y|X = x) = \mathbb{P}[\text{receive } y|\text{sent } x]$. Hence, $W(Y|X)$ is a stochastic matrix (each row sums to 1). Since these probabilities are independent at each transmission, the probability that by transmitting a sequence $\underline{x} = (x_1, \dots, x_n)$ the receiver receives a sequence $\underline{y} = (y_1, \dots, y_n)$ is:

$$W^n(Y = \underline{y}|X = \underline{x}) = \prod_{i=1}^{n} W(y_i|x_i) \tag{13}$$

We can formally define a discrete memoryless channel as follows:

Definition 2.2. A discrete memoryless channel is the 3-tuple $(\mathcal{X}, \mathcal{Y}, W)$, where $\mathcal{X}, \mathcal{Y}$ are the input and output set of symbols, respectively, and W is the noise matrix defined in Equation 12 such that Equation 13 holds.

Definition 2.3. An encoder $\mathcal{C}^n \subseteq \mathcal{X}^n$ is a set of distinct sequences of length n that can be transmitted through a channel. It maps messages to sequences of symbols, also known as "codewords". $|\mathcal{C}^n|$ is the number of different messages of n symbols that can be sent through the channel.

Definition 2.4. A decoder $\varphi_n : \mathcal{Y}^n \to \mathcal{C}^n$ is a function that associates received messages of length n to sequences in $\mathcal{C}^n$.

The decoder defines the correct mapping between received messages and sent messages. If the receiver receives a symbol $y \in \mathcal{Y}$, the sent symbol is supposed to be the $x \in \mathcal{X}$ such that $\varphi(y) = x$. To be good, the communication must be such that the probability that a received symbol y_i associated with the correct symbol x_i is high. At the same time, we want to compress the information as much as possible to send more informative messages. Unfortunately, these two quantities are inversely proportional: if a symbol is strongly compressed (i.e., it encodes many bits, meaning that the "bits per symbol" rate is high), alterations due to noise can cause the misunderstanding of the sent message at the receiver side. Let's now formalize the concept of transmission rate and error.

Definition 2.5. The transmission rate of an encoder $\mathcal{C}^n$ is defined as:

$$\frac{1}{n} \log |\mathcal{C}^n| \tag{14}$$

and represents the number of bits carried per symbol.

An **error** occurs when the transmitter sends a symbol x and the receiver receives a symbol y such that $\varphi(y) \neq x$, meaning that $y \notin Im_{\phi}^{-1}(x)$. We can extend this concept to sequences of symbols $\underline{x} \in \mathcal{C}^n$ and $\underline{y} \in \mathcal{Y}^n$.

Example 2.1. Letìs consider the DMC with input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{a, b\}$, and noise matrix:

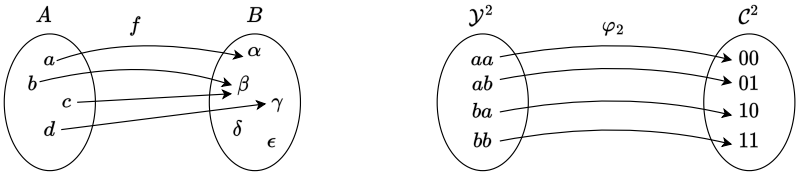
$$W(Y|X) = \begin{array}{c} \\ W(Y|X=0) \\ W(Y|X=1) \end{array} \begin{array}{c} \begin{array}{cc} W(Y=a|X) & W(Y=b|X) \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix} \end{array} \tag{15}$$

Let us consider sequences of length 2, and let the decoder be $\varphi_2(aa) = 00$, $\varphi_2(ab) = 01$, $\varphi_2(ba) = 10$ and $\varphi(bb) = 11$. If the sender sends $\underline{x} = 00$ and the receiver receives $\underline{y} = aa$, no error occurred, since $aa \in Im_{\varphi_2}^{-1}(00)$. Notice the probability of this event is $\mathbb{P}[aa|00] = 0.8^2$. Instead, if the sender sends $\underline{x} = 00$ and the receiver receives $\underline{y} = ab$, an error occurred, since $ab \notin Im_{\varphi_2}^{-1}(00)$. The probability of this event is $\mathbb{P}[ab|00] = 0.8 \cdot 0.2$. Figure 1 shows a schematic view of this.

Definition 2.6. The error probability for $\mathcal{C}^n$ and φ^n over the DMC $(\mathcal{C}^n, \mathcal{Y}^n, W^n)$ is:

$$W^n(\overline{Im_{\varphi_n}^{-1}(\underline{x})}|\underline{x}) = 1 - W^n(Im_{\varphi_n}^{-1}(\underline{x})|\underline{x}), \tag{16}$$

where $\overline{Im_{\varphi_n}^{-1}(\underline{x})} = \mathcal{Y}^n \setminus Im_{\varphi_n}^{-1}(\underline{x})$.



$Im(f) = \{\alpha, \beta, \gamma\}$

$Im_f(a) = \{\alpha\}$ $Im_f^{-1}(\alpha) = \{a\}$

$Im_f(b) = \{\beta\}$ $Im_f^{-1}(\beta) = \{b, c\}$

$Im_f(c) = \{\beta\}$ $Im_f^{-1}(\gamma) = \{d\}$

$Im_f(d) = \{\gamma\}$

$Im_{\varphi_2}(aa) = \{00\}$ $Im_{\varphi_2}^{-1}(00) = \{aa\}$

$Im_{\varphi_2}(ab) = \{01\}$ $Im_{\varphi_2}^{-1}(01) = \{ab\}$

$Im_{\varphi_2}(ba) = \{10\}$ $Im_{\varphi_2}^{-1}(10) = \{ba\}$

$Im_{\varphi_2}(bb) = \{11\}$ $Im_{\varphi_2}^{-1}(11) = \{bb\}$

Figure 1: Decoding

In simpler words, the error probability is the sum of the probabilities of receiving a sequence $\underline{y}$ given that the sent sequence is $\underline{x}$ such that $\varphi_n(\underline{y}) \neq \underline{x}$, meaning that $\underline{y}$ does not decode $\underline{x}$.
Our goal is to maximise the transmission rate (how informative messages are), while minimizing the error probability. Nevertheless, we cannot change W^n, as the channel cannot be controlled by the transmitter and the receiver. The channel is given and is noisy. What we can choose, is the way the transmitter communicates its symbols in $\mathcal{X}$, i.e. the probability distribution over $\mathcal{X}$. Apparently, it seems impossible to communicate through a noisy channel with an arbitrarily small error probability. Nevertheless, Shannon proved otherwise.

Definition 2.7. The maximum error probability is the maximum error that can occur when communicating a sequence of symbols in $\mathcal{C}^n$: $\mathsf{e}(W^n, \mathcal{C}^n, \mathcal{Y}^n) = \max_{\underline{x} \in \mathcal{C}^n} W^n(\overline{Im_{\varphi_n}^{-1}(\underline{x})}|\underline{x})$

Definition 2.8. R is an achievable rate for the DMC $(\mathcal{C}^n, \mathcal{Y}^n, W^n)$ if $\exists$ a sequence $\{\mathcal{C}^n, \varphi_n\}_{n \in \mathbb{N}}$ such that:

$$\mathsf{e}(W^n, \mathcal{C}^n, \mathcal{Y}^n)0, \quad \limsup_{n \to \infty} \frac{1}{n} \log |\mathcal{C}^n| \geq R \tag{17}$$

An achievable transmission rate is the number of bits per symbol that can be encoded with an error probability that goes to 0 when the sequences of transmitted symbols grow in length. It is easy to see that all achievable rate are finite, $R < \infty$, hence the set of all achievable rates has a maximum.

Definition 2.9. The capacity of a DMC with noise matrix W is the maximum achievable rate over the channel:

$$C(W) = \max\{R : R \text{ is an achievable rate}\}. \tag{18}$$

So the capacity of the channel, expressed in bits per symbols, is a maximum achievable rate that can be used for transmission with an error probability that goes to 0, meaning, in a reliable way.

Definition 2.10. The Shannon's capacity of channel with noise matrix W is the maximum mutual information between the input signal X and the output signal Y, that is:

$$C^*(W) = \max_{P(Y|X) \sim W(Y|X)} I(X, Y). \tag{19}$$

Theorem 2.1. (Shannon's capacity theorem for DMC.) It holds that $C(W) = C^*(W)$.

This result revolutionized the modern communication, and it says that, in order to balance the tradeoff between the error probability and the amount of information transmitted, we can pack information in the sent messages with a number of bits that is equal to the maximum mutual information between input and output signals. Notice that this is a bound: we can achieve a transmission with error probability that goes to zero with lower rates, too. Nevertheless, we cannot exceed the mutual information if we want the transmission to be reliable.

2.2 Continuous channels with Gaussian noise

Spoiler: The sensor layer of the IoT is usually a wireless sensor network (WSN), meaning that the communication between devices is wireless. Wireless communication uses electromagnetic waves to encode bits and to transmit a signal from a transmitter to a receiver. The waves are characterized by different parameters. One of them is their amplitude, that is proportional to how powerful the signal strength is. Another key feature of waves is that, when they interfere, they sum up together. Since all the objects emit electromagnetic radiations, the channels that the wireless signals travel through are usually very noisy, and noise has an additive and independent behaviour. This means that, if we transmit a signal X and receive a signal Y, then it holds that $Y = X + Z$, where Z is the noise, and X and Z are independent. Furthermore, the noise usually has a Gaussian behaviour with 0 mean, meaning that $Z \sim \mathcal{N}(0, \sigma^2)$. The probability density function for such a random variable is: $\phi(z) = \frac{e^{-\frac{z^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$, where σ is the standard deviation of the distribution.
To formulate the Shannon's capacity theorem for this kind of channels, we need to tune $P(X)$ such that the mutual information between X and Y is maximal. Since the mutual information is defined as the difference between the entropy of Y and the entropy of Y given X, we need to extend the concept of entropy in the context of continuous variables.
Thankfully, this can be done easily by defining the differential entropy. It is not important to see its definition (it is the straightforward extension of Definition 1.2 from the discrete case to the continuous case - with an integral instead of a sum). What we need to know is that the entropy is maximal when it is indeed normally distributed! In particular, the entropy of a normally distributed random variable with variance σ^2 is:

$$\frac{1}{2}\log(2\pi e\sigma^2). \tag{20}$$

Furthermore, it holds that if $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Z \sim \mathcal{N}(0, \sigma_Z^2)$, and if X and Z are independent, then also $Y = X + Z$ follows the normal distribution, its mean is 0 and its variance is $\sigma_Y^2 = \sigma_X^2 + \sigma_Z^2$.

Now we have all the ingredients for computing $I(X,Y) = H(Y) - H(Y|X) = H(Y) - H(X + Z|X)$. Notice that, since X and Z and independent, then $H(X + Z|X) = H(Z)$ (easy to prove). Hence, $I(X,Y) = H(Y) - H(Z)$. It holds that:

$$\begin{aligned}
H(Y) - H(Z) &= \frac{1}{2}\log(2\pi e\sigma_Y^2) - \frac{1}{2}\log(2\pi e\sigma_Z^2) = \\
&= \frac{1}{2}\log(\frac{2\pi e\sigma_Y^2}{2\pi e\sigma_Z^2}) = \\
&= \frac{1}{2}\log(\frac{\sigma_Y^2}{\sigma_Z^2}) = \\
&= \frac{1}{2}\log(\frac{\sigma_X^2 + \sigma_Z^2}{\sigma_Z^2}) = \\
&= \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})
\end{aligned}$$

The variance σ_X^2 is the power of the signal X (we will see that electromagnetic waves carry energy), and analogously for Z. The quantity $\frac{\sigma_X^2}{\sigma_Z^2}$ is the signal to noise ratio (SNR), representing the ratio between the power of the signal and the power of the noise. The higher the ratio, the more the signal is strong and can be heard by the receiver. By assuming that the bandwidth of the channel is B, it follows from the sampling theorem (known as Nyquist-Shannon theorem), that the conversion from an analog signal (i.e., an electromagnetic wave) into a digital signal (i.e., a sequence of bits) can be done perfectly if the analog signal is sampled often enough, and in particular, it has to be sampled twice the bandwidth. Hence, by integrating the above expression, we get the Shannon-Hartley Theorem:

$$C(W) = B\log(1 + \text{SNR}). \tag{21}$$

This capacity is expressed in bits per second, and says that the speed of transmission depends on the bandwidth of the channel, on the strength of the signal, and on the strength of the noise in the channel.