A Problem In Information Transmission

Alice and Bob participate in a game show where a prize of great value is randomly located at one of 256 equiprobable locations on a map that is a 16-by-16 grid. The actual location of the prize is revealed to Alice in the game show studio. She must transmit it to Bob so he can find it before other contestants. She has at her disposal sequences of binary symbols: zeros and ones.

Discussion Questions:

1. What are possible ways that Alice could use a binary sequence to convey the location of the prize?

2. What is the most efficient method (fewest symbols)?

3. How do we know the proposed method is the most efficient?

4. Suppose the prize is likely to be found at four specific locations of the possible 256 locations each with a 12.5% probability and at any of the other 252 locations each with a probability of 0.1984%. Would the most efficient method for transmitting the location change? If so, what would be the better method?

5. Suppose noise is present in the medium used to transmit Alice's message to Bob where zeros become ones and ones become zeros with a probability of 5%. Is there a way to convey the message more reliably through this corrupted medium? Is it perfect? How do we know?

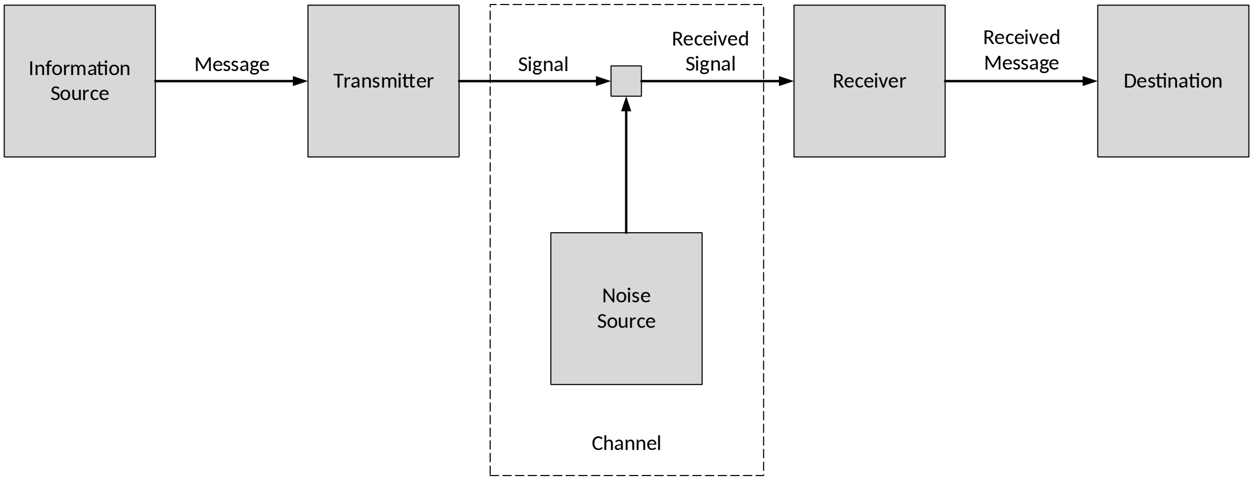
Information Theory

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have* <u>*meaning*</u>*.*
>
> Claude Shannon

This is how Claude Shannon began his seminal work "A Mathematical Theory Of Communication" published in 1948 which is the basis for understanding how all information transmission systems work. Familiar examples include telegraphy, voice telephony, broadcast and cable television, satellite communication, fiber optics, digital storage and retrieval as well as cellular genetics and quantum mechanics. This list could go on and on!

Communication System Model

Shannon showed that all information transmission systems have the same five-element model:



<u>Information Source</u>

The *information source* produces a message or a sequence of messages to be communicated to the destination. These messages can be in a variety of forms:

- Sequences of letters (telegraph)
- Continuous function of time (radio)
- Continuous function of time and space (black and white television)
- Multiple continuous functions of time and space (color television)
- Sequences of symbols (RAM, magnetic disks, CDs/DVDs)
- Combinations of all of the above (television with audio)

Transmitter

The *transmitter* operates on the message in some way to produce a signal suitable for transmission over the channel. Examples include:

- Morse code (telegraph)
- AM/FM systems (broadcast radio and television)
- IQ modulation systems (digital communication systems)

Channel

The *channel* is the medium used to transmit the signal from the transmitter to the receiver. During transmission the signal may be perturbed by noise. In this case the received signal will necessarily differ from the transmitted signal. Example of channels include:

- A pair of wires (telegraph line, digital bus)
- Coaxial cable (transmission line)
- A band of radio frequencies (cellular phone, satellite links, wireless LAN)
- A beam of light (fiber optics)

Receiver

The *receiver* usually performs the inverse operation of the transmitter, constructing the received message from the received signal. In the presence of noise the received message may or may not be the same as the original message.

Destination

The *destination* is the person or thing for whom the message is intended.

Types of Communication Systems

Communication systems may be classified into three categories: *discrete, continuous* and *mixed*.

In a discrete system both the message and the signal are a sequence of discrete symbols. An example of a discrete system is a telegraph where the message is a sequence of letters and the signal is a sequence of dots, dashes and spaces. A more modern example is a digital data bus where the message is a sequence of ones and zeros and the signal is a sequence of discrete binary-valued voltages.

In a continuous system both the message and the signal are continuous functions of time, e.g. radio and television.

A mixed system has both discrete and continuous variables, e.g. digital transmission of speech.

Definition of Information

Let E be an event that occurs with probability $P(E)$. The amount of *information* received if event E occurs is defined to be:

$$I(E)=\log_x \frac{1}{P(E)}=-\log_x P(E)$$

The unit of information depends on the logarithm base. If the base is 2 then the unit of information is the *bit*, a contraction of the term ‘binary information digit” and first coined by John Tukey, a colleague of Shannon’s at Bell Labs. If the base is *e* then the unit is the *nat* and if it is 10 then the unit is the *hartley* named in honor of Ralph Hartley who first suggested a logarithmic measure for information and whose work in information theory laid the groundwork for Shannon.

Information theorists work predominantly with bits and omit the base 2 subscript in the logarithmic function as it is assumed.

Zero-Memory Source

Suppose an information source produces a sequence of symbols chosen from a fixed finite *source alphabet* $S=\{s_1, s_2, \ldots, s_q\}$. Each symbol s_i in the sequence is statistically independent and occurs with probability $P(s_i)$. Such a source is called a *zero-memory source* and is completely described by the source alphabet S and the associated symbol probabilities.

The amount of information obtained from the occurrence of symbol s_i is:

$$I(s_i)=\log \frac{1}{P(s_i)}=-\log P(s_i)$$

This happens with probability $P(s_i)$ so the average amount of information per source symbol is the weighted average of all the source symbols:

$$\sum_S P(s_i) I(s_i)$$

This quantity, the average amount of information per source symbol, is called the *entropy* $H(S)$ of the zero-memory source:

$$H(S)=\sum_S P(s_i)\log \frac{1}{P(s_i)}=-\sum_S P(s_i)\log P(s_i)$$

Shannon noted the similarity in the definitions of entropy for information theory and statistical mechanics.

Entropy may be interpreted either as the average amount of information per symbol provided by the source or as the average amount of uncertainty per symbol that is removed once a source symbol is observed.

If an information source produces symbols at a definite rate of R_S symbols per second then the entropy of the source measured in bits per second is:

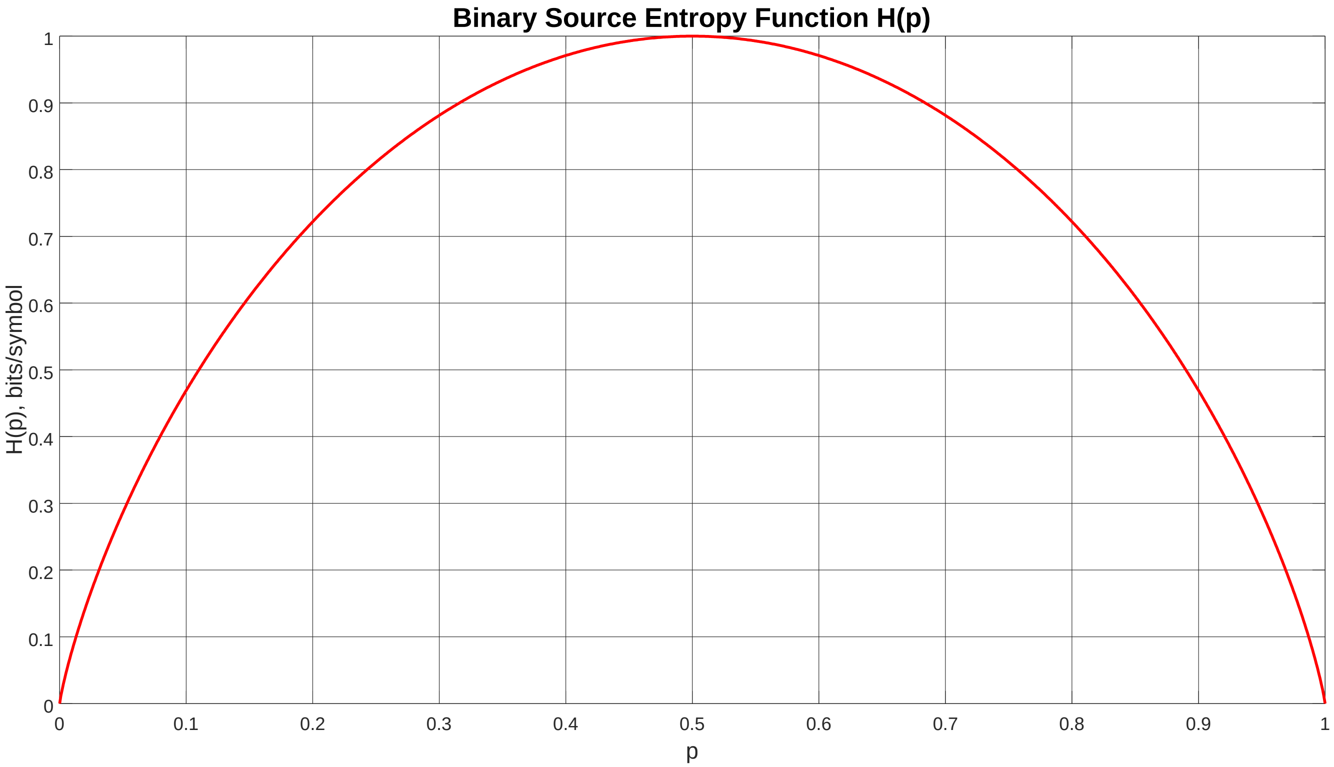
$$H'(S)=R_S\cdot H(S)$$

An important example of a zero-memory source is the zero-memory binary source. The source alphabet is simply $S=\{0,1\}$. The probability of a 0 symbol being emitted by the source is p so the probability of a 1 symbol being emitted is therefore $1-p$.

The entropy function $H(S)$ is:

$$H(S)=-p\log p-(1-p)\log(1-p)$$

This particular entropy function occurs so often in information theory it is also referred to as $H(p)$ to highlight the functional dependence on the probability variable p. A plot of $H(p)$ is shown below:



Calculation of $H(0)$ and $H(1)$ involves evaluating $0\log 0$ which by definition is zero.

Note that if the output of the binary source is certain (either $p=0$ or $p=1$) then the source provides no information. The maximum average amount of information provided by this source is 1 bit/symbol which occurs if and only if the two output symbols are equiprobable ($p=0.5$).

Markov Source

The zero-memory source is too restrictive for many applications of interest. A more general type of source is one where the occurrence of a symbol s_i depends on a finite number m of preceding symbols. Such a source is called an m-th order Markov source. It is fully specified by a source alphabet of q symbols $S=\{s_1, s_2, \ldots, s_q\}$ and a set of q^{m+1} conditional probabilities:

$$P\left(s_i \middle| s_{j_1}, s_{j_2}, \ldots, s_{j_m}\right) \text{for} \, i=1,2,\ldots,q \, ; \, j_k=1,2,\ldots,q$$

An m-th order Markov source can be represented by a finite-state machine. The q^m states are nodes representing all possible sequences of preceding symbols. The state-to-state transitions are arrows labeled with the conditional probabilities and emitted symbols above.

The output of a Markov source is a *stochastic* process. Because the conditional probabilities that specify the Markov source are time-invariant the process is also *stationary*. The probabilities that the source is in its various states can be calculated from the conditional probabilities. These state probabilities are called the *stationary distribution*.

We restrict ourselves to considering only *ergodic* Markov sources. Simply put an ergodic source is one where every sufficiently long sequence produced by that source has the same statistical properties (e.g. symbol frequencies). With certainty averages for a specific long sequence are identical to those same averages over an ensemble of possible long sequences.

The entropy of a Markov source can be determined as follows: For each state S_i there is a set of probabilities $P(s_j \vee S_i)$ for producing the various possible symbols s_j. Thus there is an entropy H_i for each state. The entropy of the Markov source is then defined as the average of these H_i weighted by the probabilities of the occurrences of the states. The weights are the stationary distribution $P(S_i)$ of the source.

$$H(S)=\sum_{i=1}^{q^m} P(S\text{¿¿}i)H_i=-\sum_{i=1}^{q^m} P(S\text{¿¿}i)\sum_{j=1}^{q} P(s_j \vee S_i)\log P(s_j \vee S_i)\text{¿¿}$$

Markov sources are of interest to information theorists because of their ability to mathematically model human language.

Stationary Distribution

When working with Markov sources the stationary distribution is needed to determine the entropy of the source. The tedious calculations to find it can be facilitated with a transition probability matrix and matrix algebra.

Suppose we have an m-th order Markov source with source alphabet $S=\{s_1, s_2, \ldots, s_q\}$. The state machine representing this source has q^m states denoted by S_i. The set of probabilities $P(s_k, S_j \vee S_i)$ represent the chances given state S_i of transitioning to state S_j while generating the symbol s_k. Note $i, j=1,2,\ldots,q^m$ while $k=1,2,\ldots,q$. Many of these probabilities must be zero because the generation of a particular s_k from a state S_i restricts the possible states S_j to only q choices of the q^m total states.

If the source is ergodic the stationary probability $P(S$¿¿ $j)$¿ of being in state S_j must be the sum of the probabilities of transitioning to state S_j from any other state weighted by the probability of being in that other state:

$$P(S_j)=\sum_{i=1}^{q^m} P(s_k, S_j | S_i) P(S_i) \, \text{for} \, j=1,2,\ldots,q^m$$

Define the q^m-by-q^m state transition probability matrix $[P]$ where $[P]_{i,j}=P(s_k, S_j \vee S_i)$ and the 1-by-q^m stationary distribution row vector $w]$ where $w]_j=P(S_j)$. The q^m summations above can then be written concisely as a matrix equation:

$$w]=w][P]$$

$$w][P-I]=0]$$

This q^m-by-q^m system of equations does not have a unique solution for $w]$. There is however an additional constraint that allows for a unique solution. The individual stationary probabilities sum to unity since the source must with certainty be in one of the q^m states:

$$\sum_{j=1}^{q^m} P(S_j)=\sum_{j=1}^{q^m} w]_j=1$$

Replacing any one of the q^m state probability equations with this constraint will result in a unique solution for the stationary distribution.

To introduce this constraint into the matrix equation for $w]$ define a new matrix $[Q]$ where the first column of the matrix $[P-I]$ has been replaced with all ones. Further define a row vector $b]$ consisting of a one followed by q^m-1 zeros. The matrix equation to solve for $w]$ becomes:

$$w][Q]=b]$$

$$w]=b][Q]^{-1}$$

If a row vector $h]$ is formed that contains the q^m state entropies H_i which are solved for separately then the source entropy $H(S)$ is the dot product of $w]$ and $h]$:

$$H(S)=\sum_{i=1}^{q^m} P(S\text{¿¿}i)H_i=\text{¿}w]h]^T\text{¿¿}$$

Extended Source

When discussing the properties of information sources it is often useful to deal with blocks of source symbols rather than individual source symbols.

If S is a source with an alphabet of q symbols $S=\{s_1,s_2,\ldots,s_q\}$ then the n-th extension of S is denoted by S^n and has an alphabet of q^n symbols $S^n=\{\sigma_1,\sigma_2,\ldots,\sigma_{q^n}\}$ where each σ_i is a sequence of the s_i's of length n.

- For a zero-memory source if the probabilities associated with the various s_i are P_i then the probabilities associated with the various σ_i are $P(\sigma_i)=P_{i_1}P_{i_2}\ldots P_{i_n}$ since each symbol in the sequence is independent.
- For an m-th order Markov source the conditional symbol probabilities for the various s_i are $P(s_i|s_{j_1},s_{j_2},\ldots,s_{j_m})$. The n-th extension of this source is also a Markov source but of order μ where $\mu=\lceil m/n \rceil$, the smallest integer greater than or equal to m/n. The conditional probabilities for the extended source are $P(\sigma_i|\sigma_{j_1},\sigma_{j_2},\ldots,\sigma_{j_\mu})$.

In either case the entropy of the extended source is:

$$H(S^n)=nH(S)$$

Each symbol from the original source S adds on average $H(S)$ bits of information. The sequence of n symbols from the extended source S^n therefore provides on average $nH(S)$ bits of information.

Adjoint Source

Using the stationary distribution of an m-th order Markov source it is possible to calculate the probability distribution of the individual source symbols $S=\{s_1,s_2,\ldots,s_q\}$. These q numbers, $P_1,P_2,\ldots,P_q$, are called the first-order (unconditional) symbol probability distribution of the source. The *adjoint source* to S, written $\acute{S}$, is the zero-memory information source with a source alphabet identical to S but with symbol probabilities $P_1,P_2,\ldots,P_q$.

The adjoint to a zero-memory source is the source itself.
It can be demonstrated that the entropy of the adjoint source $\acute{S}$ is always greater than or equal to the entropy of the original source S with equality occurring when S is a zero-memory source:

$$H(\acute{S})\geq H(S)$$

The two sources have identical first-order symbol probabilities. They differ only in the fact that S has additional constraints expressed by the conditional symbol probabilities imposed on its output sequences. These constraints serve to decrease the average amount of information from the original source relative to the adjoint source $\acute{S}$.

Extended Adjoint Source

The adjoint $\acute{S}^n$ of an extended source S^n can also be defined. Let $P(\sigma_1), P(\sigma_2), \ldots, P(\sigma_{q^n})$ be the first-order (unconditional) probabilities of the σ_i, the symbols of the n-th extension of an m-th order Markov source.

The entropy of the adjoint to a source is always greater than or equal to the entropy of the source itself. The entropy of an extended source is n times the entropy of the original source. Putting these facts together:

$$H(\acute{S}^n) \geq H(S^n) = nH(S)$$

Dividing by n:

$$\frac{H(\acute{S}^n)}{n} \geq H(S)$$

It can be shown in the limit as $n \to \infty$:

$$\lim_{n \to \infty} \frac{H(\acute{S}^n)}{n} = H(S)$$

This result is important later when we are seeking an efficient method to code information source symbols.

Digital Signal Transmission

In modern digital communications systems information source symbols must be transformed to waveforms suitable for transmission over the channel. This transformation process includes some or all of the following steps: formatting, source coding, encryption, channel coding and modulation.

Formatting

The original form of most communicated data is textual or analog. Textual data is character coded using one of several standard formats. Examples include the American Standard Code of Information Interchange (ASCII), the Extended Binary Coded Decimal Interchange Code (EBCDIC), Baudot and Hollerith. Analog data is sampled, quantized and then digitally encoded by means of an analog-to-digital conversion (ADC) process. In both cases the result of formatting is the original source information is represented by a sequence of binary symbols.

Source Coding

Source coding is the operation of forming efficient representations of information source symbol sequences. Efficient representation permits reduction (compression) in memory or bandwidth resources required to store or transmit source data. The possible efficiency gains depend on the statistical correlation among successive information source symbols. Lossless source coding schemes result in no loss of original source information while lossy source coding schemes can achieve much greater efficiencies but with corresponding losses of original source information.

Encryption

Encryption is the process where information source data is encoded in such a way that only authorized parties can access it. Encryption is not a subject of this course.

Channel Coding

Channel coding enables transmitted signals to more reliably withstand the effects of various channel impairments like noise, interference and fading. Channel coding techniques include waveform coding and structured sequences. These methods coupled with high-speed digital signal processing allow system designers to make favorable error-performance, bandwidth and power trade-offs.

Modulation

Modulation deals with the conversion of the sequence of binary digits emerging from the prior steps in the transmission process into electrical waveforms suitable for the channel. Modulation can result in baseband or passband signals.

Baseband signals are low frequency signals (typically DC to a few megahertz) that can be sent over a pair of wires or a coaxial cable. Baseband modulation techniques can be classified as Pulse Coded Modulation (PCM) or M-ary Pulse Modulation. PCM can be further categorized as Non-Return-to-Zero (NRZ), Return-to-Zero (RZ), Phase Encoded or Multilevel Binary. M-ary Pulse Modulation can be further categorized as Pulse-Amplitude Modulation (PAM), Pulse-Position Modulation (PPM) or Pulse-Duration/Width Modulation (PDM/PWM).

Passband signals are high frequency signals (many megahertz and above) that can be sent over a coaxial cable or a radio link by means of antennas. Similar to baseband modulation, passband modulation techniques include binary and multilevel systems. In binary passband systems the amplitude, frequency or phase of a high-frequency carrier is modulated between two states resulting in Amplitude-Shift Keying (ASK), Frequency-Shift Keying (FSK) or Phase-Shift Keying (PSK). In multilevel passband systems the frequency or phase of the carrier can take on more than two states resulting in M-FSK or M-PSK schemes. Additionally two quadrature carriers (IQ) can be used to generate M-ary Quadrature Amplitude Modulation (M-QAM).

Source Coding

Source coding maps all possible sequences of symbols from a source alphabet $S=\{s_1, s_2, \ldots, s_q\}$ to sequences of symbols from a *code alphabet* $X=\{x_1, x_2, \ldots, x_r\}$ comprised of r elements. The important case where $r=2$ is called *binary coding*.

A *block code* maps each source symbol from S into a fixed sequence of symbols from X. These fixed sequences of x_j are called *code words*. The code word corresponding to the source symbol s_i is denoted by X_i. The concept of a block code also includes sequences of source symbols of length n because the code maps each symbol from the n-th extension of the original source into a fixed sequence of code symbols.

A block code is *nonsingular* if all the code words are distinct. A block code is further said to be *uniquely decodable* if and only if the n-th extension of the code is nonsingular for all finite n.

Finally a uniquely decodable code is said to be *instantaneous* if it is possible to decode each word in a sequence without reference to succeeding code symbols. An instantaneous code is also a *prefix* code because for such a code no code word can be the prefix (beginning) of any other code word.

If the probabilities of the various source symbols are $P_1, P_2, \ldots, P_q$ and the lengths of the corresponding code words are $l_1, l_2, \ldots, l_q$ then an *average length* L can be defined:

$$L=\sum_{i=1}^{q} P_i l_i$$

A uniquely decodable code is called *compact* if its average length is less than or equal to the average length of all the other uniquely decodable codes for the same source and code word alphabet. *The fundamental problem of coding information sources is finding compact codes*.

If we further require that the code be instantaneous then it can be shown that the average length has a lower bound:

$$L \geq \frac{H(S)}{\log r}$$

$H(S)$ is the source entropy measured in bits/source symbol. The denominator is a normalizing factor. It expresses the amount of information carried per code symbol and has units of bits/code symbol. L therefore has the expected units of code symbols/source symbol. For binary coding synonyms for code symbol include PCM (Sklar), *binit* (Abramson) or simply binary symbol.

Shannon’s Noiseless Coding Theorem

Consider an m-th order Markov source with alphabet $S=\{s_1, s_2, \dots, s_q\}$ and conditional probabilities $P(s_i|s_{j_1}, s_{j_2}, \dots, s_{j_m})$. The n-th extension S^n of S is a μ-th order Markov source with alphabet $S^n=\{\sigma_1, \sigma_2, \dots, \sigma_{q^n}\}$ and conditional probabilities $P(\sigma_i|\sigma_{j_1}, \sigma_{j_2}, \dots, \sigma_{j_\mu})$ where each σ_i corresponds to some sequence of length n of the s_i.

The original Markov source and its extension each have stationary distributions denoted by $P(s_i)$ and $P(\sigma_i)$ respectively.

Imagine an r-ary coding scheme for the extended source. If the code word corresponding to symbol σ_i for the extended source has a length λ_i then an average extended source code word length may be defined:

$$L_n=\sum_{i=1}^{q^n} P(\sigma_i)\lambda_i$$

The quantity L_n/n is then the average number of code symbols per single symbol from the original source S.

Shannon showed that it is possible to devise a code such that L_n/n is arbitrarily close to the entropy of the original source by making n sufficiently large:

$$\lim_{n\to\infty} \frac{L_n}{n}=\frac{H(S)}{\log r}$$

We can make the average number of r-ary code symbols per source symbol no smaller than the entropy of the source measured in r-ary units. This conclusion is also true for zero-memory sources. The cost of decreasing L_n/n is the increased coding complexity caused by the large number $(q$¿¿$n)$¿ of extended source symbols.

Efficiency, Redundancy and Compression Ratio

Efficiency, redundancy and compression ratio are various ways to quantify how close a given source code comes to the theoretical limit set by Shannon's Noiseless Coding Theorem.

The *efficiency* of a source code when the source symbols are coded one at a time is:

$$Efficiency = \frac{H(S)}{L \log r}$$

When the source symbols are coded in blocks of length n the efficiency is:

$$Efficiency = \frac{nH(S)}{L_n \log r}$$

Efficiency is a dimensionless quantity between zero and one inclusive. A source code efficiency of one occurs when the code achieves the theoretical Shannon limit.

The *redundancy* of a code is:

$$Redundancy = 1 - Efficiency$$

Redundancy is also a dimensionless quantity with possible values between zero and one. It is a measure of the "extra" source code symbols above the minimum possible.

The *compression ratio* is the ratio of the "uncoded" source code word length to the average source code word length. An uncoded or straight r-ary source code is one where the source symbols are represented by fixed length sequences of r-ary symbols. These codes are uniquely decodable and instantaneous since they are fixed length. They are also compact if and only if the number of source symbols is an exact power of r and the source symbols are equiprobable.

For source symbols coded one at a time:

$$Compression\,Ratio = \frac{\lceil \log_r M \rceil}{L}$$

For source symbols coded in blocks of length n:

$$Compression\,Ratio = \frac{n \lceil \log_r M \rceil}{L_n}$$

In both cases M is the number of source code words.

Information Channel

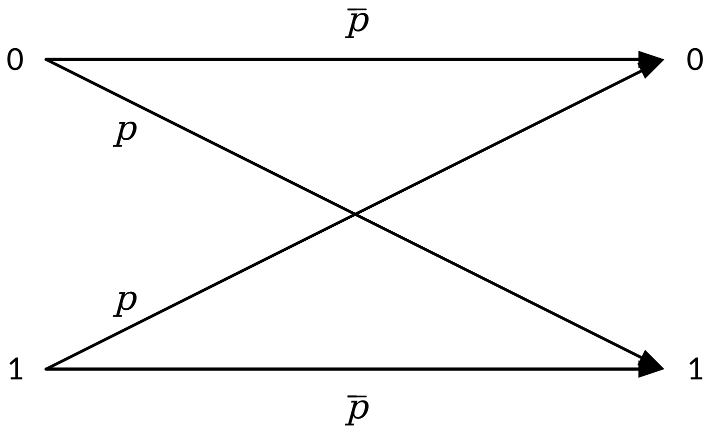
A *memoryless information channel* is defined by an input alphabet $A=\{a_1, a_2, \ldots, a_r\}$, an output alphabet $B=\{b_1, b_2, \ldots, b_s\}$ and a set of conditional probabilities $P(b_j|a_i)$ where $i=1, 2, \ldots, r$ and $j=1,2,\ldots,s$. $P(b_j|a_i)$ is the probability that output symbol b_j will be received if input symbol a_i is sent.

A *channel matrix* $[P]$ where $[P]_{i,j}=P(b_j \vee a_i)$ can be formed which completely describes the information channel.

A channel described by a channel matrix with one and only one nonzero element in each column is called a *noiseless channel*. A channel described by a channel matrix with one and only one nonzero element in each row is called a *deterministic channel*.

Binary Symmetric Channel

The *binary symmetric channel* (BSC) is a particular information channel of great theoretical and practical importance. This channel has two input symbols $\{0, 1\}$ and two output symbols $\{0, 1\}$. We define p as the *symbol error probability*, i.e., the probability that a 1 is received when a 0 is sent or that a 0 is received when a 1 is sent. Further we define $\acute{p}=1-p$, the probability that a given input symbol is received correctly at the output. A *channel diagram* of the BSC is shown below:



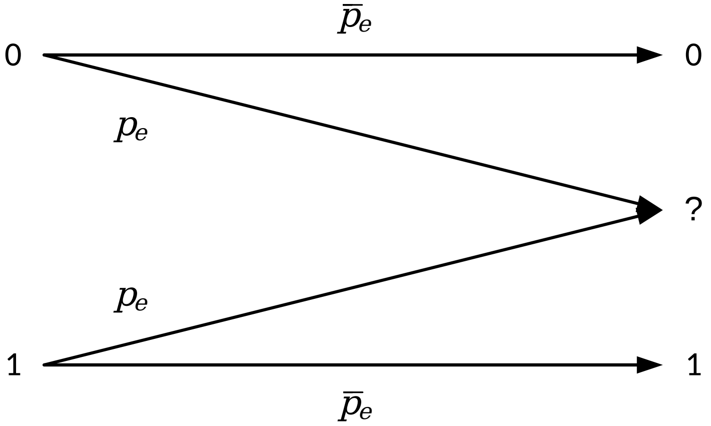
The channel matrix for the BSC is:

$$[P]=\begin{bmatrix} \acute{p} & p \\ p & \acute{p} \end{bmatrix}$$

A BSC with $p=0$ is both a noiseless and a deterministic channel.

Binary Erasure Channel

The *binary erasure channel* (BEC) is another important information channel model. This channel has two input symbols $\{0,1\}$ and three output symbols $\{0,?,1\}$. The *symbol erasure probability* p_e gives the probability that the output symbol received is indeterminate given an input symbol is sent, i.e. the symbol is “erased”. The probability $\acute{p}_e = 1 - p_e$ is the probability that an input symbol is definitively received and is correct.



The channel matrix for the BEC is:

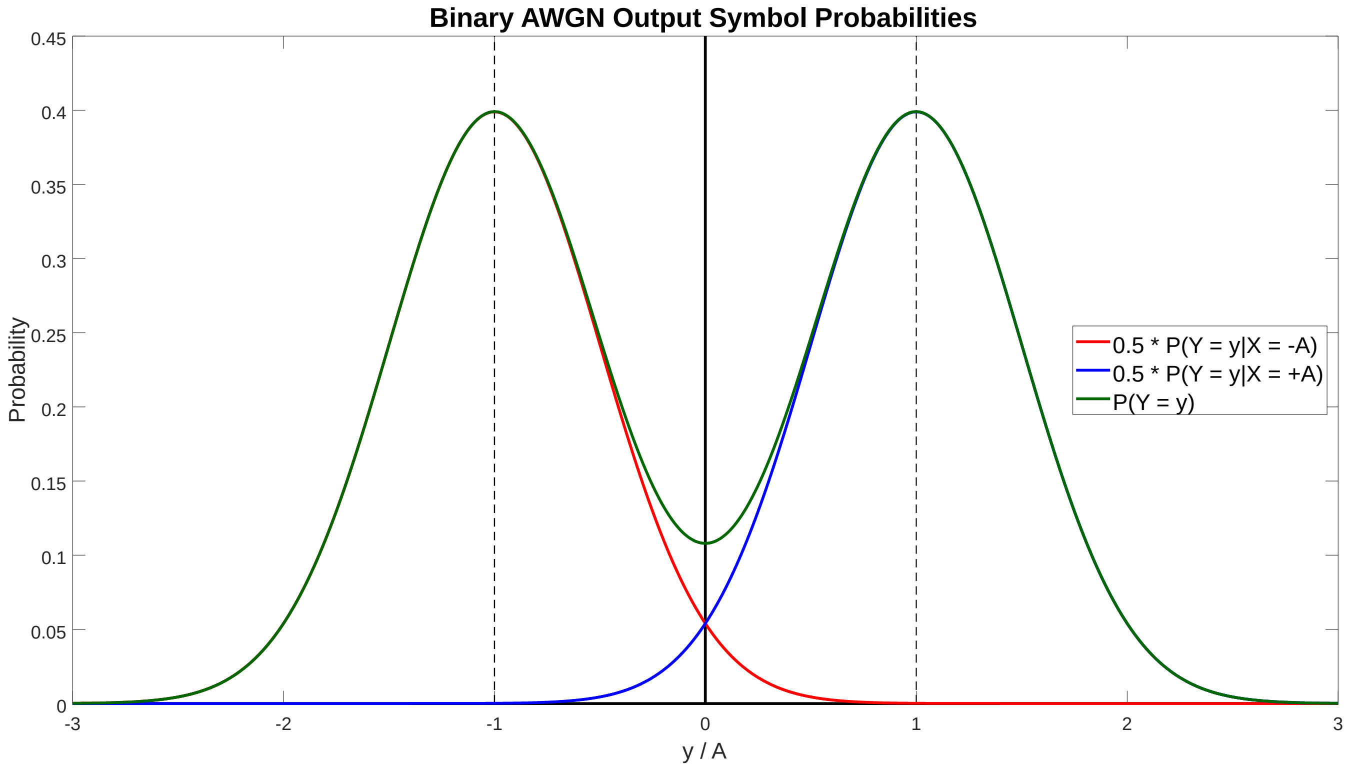
$$[P] = \begin{bmatrix} \acute{p}_e & p_e & 0 \\ 0 & p_e & \acute{p}_e \end{bmatrix}$$

Binary AWGN Channel

The *binary additive white Gaussian noise channel* (BAWGNC) is a third important information channel model. This channel has two discrete input symbols $X \in \{-A, +A\}$ and a continuous output $Y = X + Z$ where Z is a zero-mean Gaussian random variable with variance σ^2. If X is an equiprobable random variable then the random variable Y will be the equally weighted combination of two conditional probability density functions:

$$P(Y = y | X = -A) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y+A)^2}{2\sigma^2}}$$

$$P(Y = y | X = +A) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-A)^2}{2\sigma^2}}$$



Suppose $X = -A$ corresponds to the transmission of a 0 and $X = +A$ corresponds to the transmission of a 1. To decide whether a 0 or a 1 has been received we set a *hard decision* threshold at $y = 0$. If $y < 0$ we conclude a 0 was received and if $y > 0$ we conclude a 1 was received.

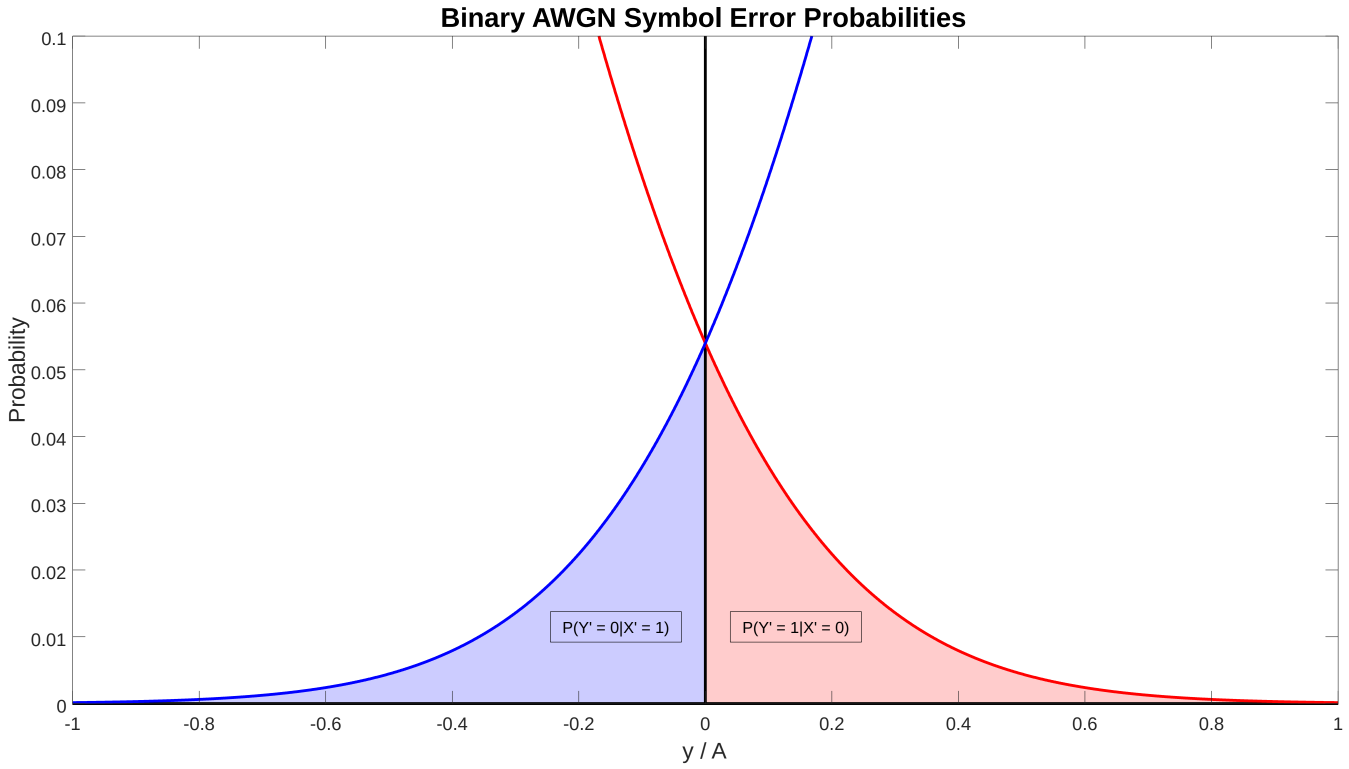
Problems (symbol errors) arise if a 0 was actually sent ($X = -A$) but $y > 0$ and we decide a 1 was received or if a 1 was sent ($X = +A$) but $y < 0$ and we decide a 0 was received. The probabilities of symbol errors occurring is related to the areas under the tails of the respective probability density functions that are on the “wrong” side of the zero point. This is depicted in the graph at the top of the next page. These areas are given by the complementary error function $\operatorname{erfc}(x)$ or the Q function $Q(x)$.

The two possible values for the input random variable X correspond to the amplitude A of a polar NRZ signal in real-world binary AWGN baseband channels. The variance σ^2 of the random variable Z corresponds to the noise power of the system, the electronic noise added by the channel and the receiver.

Two observations are in order:

- If the signal amplitude A is increased (more signal power) for a fixed noise power then the symbol error rate improves (less tail area on the wrong side of zero).
- If the noise power is increased (larger variance) for a fixed signal power then the symbol error rate degrades (more tail area on the wrong side of zero).

The symbol error rate therefore depends on the ratio of signal power to noise power. In modern digital communication systems, the related ratio of *symbol energy* E_S to *noise spectral density* N_0 is much more commonly used to determine the symbol error rate.



For baseband polar NRZ signals amplitude A (measured in volts) and signal power S (measured in watts) are easily related:

$$S=\frac{(\pm A)^2}{R}=\frac{A^2}{R}$$

where R is the system resistance. Often R is normalized to 1 Ω thus signal power is simply:

$$S=A^2$$

The symbol energy E_S (measured in watt-seconds or joules) is defined as the amount of energy associated with a single symbol of the symbol stream.

$$E_S=S\,T_S=\frac{S}{R_S}=A^2T_S$$

where T_S is the duration of a single symbol (measured in seconds) and $R_S=\frac{1}{T_S}$ is the symbol rate (measured in baud or symbols/second).

The variance σ^2 is proportional to the noise spectral density N_0 (measured in watts/hertz or joules) which in turn is the product of Boltzmann's constant $k=$¿ 1.38 x 10^{-23} J/K and the *equivalent system noise temperature* T_e (measured in kelvins) of the channel and receiver:

$$\sigma^2=\frac{N_0}{2}=\frac{k\,T_e}{2}$$

The symbol error rate (the probability of a symbol error) for each type of baseband or passband modulation has its own functional dependence on the $\frac{E_S}{N_0}$ ratio. For polar NRZ baseband signals this functional dependence is specifically:

$$P_S=\frac{1}{2}erfc\left(\sqrt{\frac{E_S}{N_0}}\right)=Q\left(\sqrt{\frac{2E_S}{N_0}}\right)$$

The $\frac{E_S}{N_0}$ ratio is usually expressed in decibels (dB):

$$\left(\frac{E_S}{N_0}\right)_{\text{dB}}=10\log_{10}\frac{E_S}{N_0}$$

Finally, bit is often substituted for symbol and we speak of the bit error rate or BER.

As an example consider a polar NRZ waveform with an amplitude of 5 mV and a symbol rate of 1 MBd. Determine the $\frac{E_S}{N_0}$ ratio in decibels and the symbol error rate if the system resistance is 1 Ω and the noise spectral density is 10^{-11} W/Hz.

The symbol energy is:

$$E_S=\frac{S}{R_S}=\frac{A^2}{R_S}=\frac{.005^2}{10^6}=2.5\text{ x }10^{-11}\text{ joules}$$

The $\frac{E_S}{N_0}$ ratio is:

$$\frac{E_S}{N_0}=\frac{2.5\text{ x }10^{-11}}{10^{-11}}=2.5$$

$$\left(\frac{E_S}{N_0}\right)_{\text{dB}}=10\log_{10}\frac{E_S}{N_0}=\text{¿}10\log_{10}2.5=\text{¿}4.0\text{ dB}\text{¿¿}$$

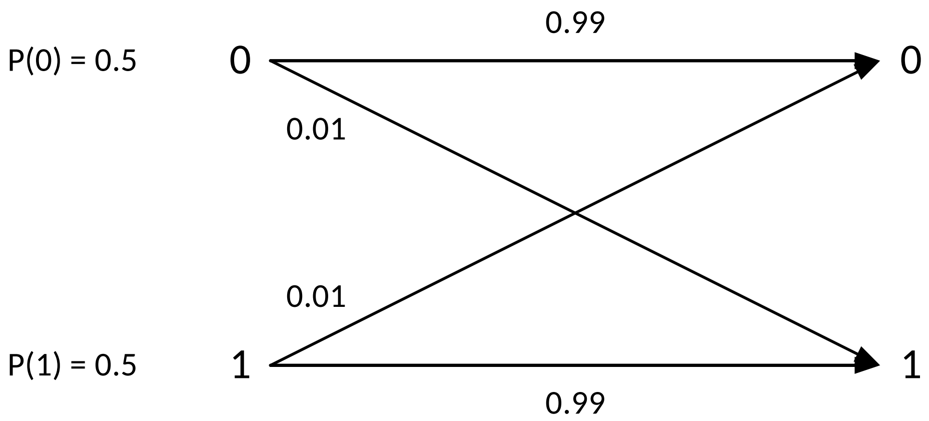
The symbol error rate is:

$$P_S=\frac{1}{2}erfc\left(\sqrt{\frac{E_S}{N_0}}\right)=\frac{1}{2}erfc\left(\sqrt{2.5}\right)=0.01267$$

Channel Capacity

The following example is taken from Shannon’s work to demonstrate the concepts associated with channel capacity.

Imagine a binary source that produces two symbols 0 and 1 that occur with equal probability $P(0)=P(1)=0.5$. The entropy of the source is $H(A)=H(0.5)=\text{¿}$ 1 bit/symbol. The output of this source is put through a BSC with $p=0.01$. This is summarized by the following channel diagram:



Suppose the source emits 1000 symbols every second. Information is therefore being produced by the source at a rate of 1000 bits/second. At the output of the BSC on average 10 in every 1000 symbols transmitted are in error. What is the rate of information transmission?

Certainly it is less than the 1000 bits/second being produced by the source. It might be tempting to simply subtract the average symbol error rate and say the rate of information transmission is 990 bits/second. This is not satisfactory since it fails to take into account the lack of knowledge of exactly where the errors occur.

The proper correction to apply to the amount of information transmitted is the amount of this information which is missing in the received message, or alternatively the uncertainty when we have received a message to what was actually sent. Since entropy is a measure of uncertainty it seems reasonable to use a source entropy conditioned on knowing the received message. This is a *conditional entropy* defined as:

$$H(A|B)=\sum_{j} P(b_j\text{¿})\,H(A\vee B=b_j)\text{¿}$$

$$\text{¿}-\sum_{j} P(b_j\text{¿})\sum_{i} P(a_i|b_j)\log P(a_i|b_j)\text{¿}$$

where A denotes the original source message and B denotes the received message.

Following this idea the rate of information transmission R is defined by subtracting the average rate of conditional entropy $H(A \vee B)$ from the average rate of information production $H(A)$:

$$R = H(A) - H(A \vee B)$$

The conditional entropy $H(A \vee B)$ is called the *equivocation*. It measures the average ambiguity of the received message.

In the example under consideration the received message B consists of two symbols 0 and 1 that are equiprobable:

$$H(A|B) = -P(0)\{P(0 \vee 0)\log P(0 \vee 0) + P(1 \vee 0)\log P(1 \vee 0)\}$$

$$-P(1)\{P(0 \vee 1)\log P(0 \vee 1) + P(1 \vee 1)\log P(1 \vee 1)\}$$

$$¿ -0.5\{0.99\log 0.99 + 0.01\log 0.01\}$$

$$-0.5\{0.01\log 0.01 + 0.99\log 0.99\}$$

$$¿ 0.081 \text{ bits/symbol}$$

The system is transmitting information at a rate $R = 1 - 0.081 = 0.919$ bits/symbol or 919 bits/second with the given symbol rate.

Using various entropy identities alternate expressions for the information rate R are:

$$R = H(A) - H(A \vee B)$$

$$¿ H(B) - H(B|A)$$

$$¿ H(A) + H(B) - H(A, B)$$

The first expression is the amount of information sent less the uncertainty of what is sent. The second measures the amount of information received less the part of this due to noise. The third is the sum of the information sent and the information received less the joint entropy which measures the amount of information common to the two.

The *channel capacity* C of a noisy channel should be the maximum possible rate of transmission:

$$C = max\, R$$

where the maximum is with respect to all possible information sources that could be used as an input to the channel.

In the general case solving for C is quite difficult. For the important case of a discrete channel with input symbols that all have the same set of probabilities emerging from them *and* output symbols that all have the same set of probabilities entering into them a closed form solution for C does exist:

$$C = \log M + \sum_i P_i \log P_i$$

where M is the number of output symbols and the P_i are the transition probabilities from any input symbol.

For the specific case of the BSC this formula yields:

$$C = 1 - H(p)$$

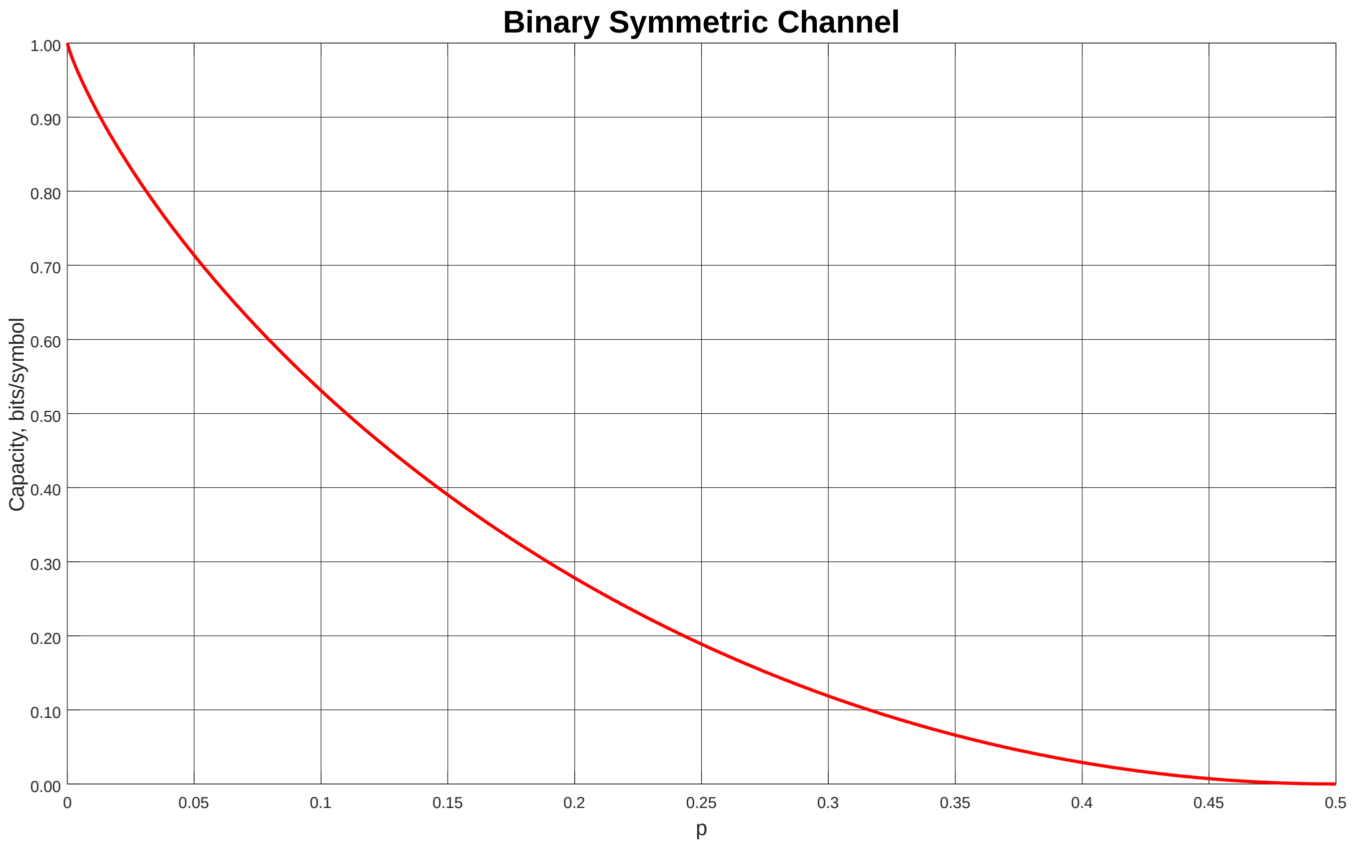
where p is the probability of a symbol error.

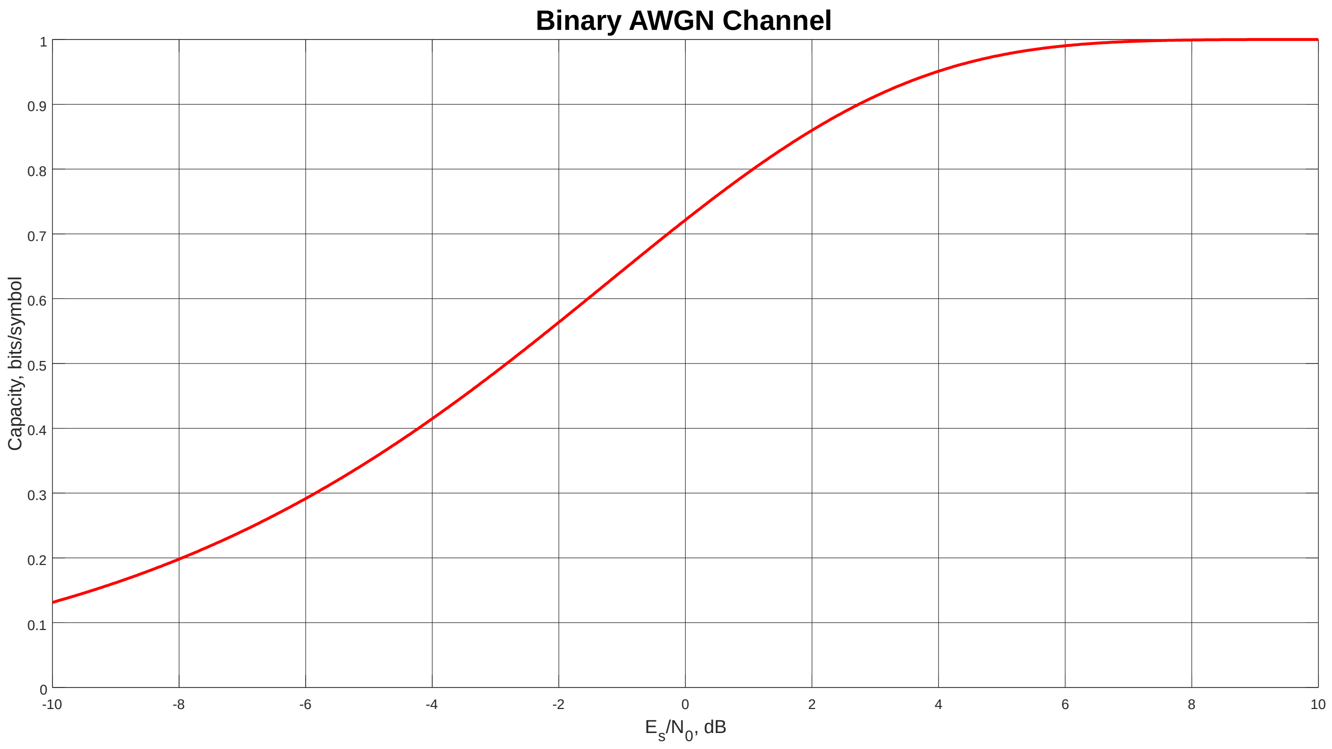
The BEC does not meet the criteria for the capacity formula above nonetheless the capacity can be found to be:

$$C = 1 - p_e$$

where p_e is the probability of a symbol erasure.

The BAWGNC is not a discrete channel since its output is a continuous value. The capacity calculation for this channel is quite complex. The summations in the equivocation formula are replaced with integrals that must be evaluated numerically. Capacity as a function of E_S/N_0 is presented in tabular or graphical form.


Binary Symmetric Channel
Capacity, bits/symbol
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
p


Binary AWGN Channel
Capacity, bits/symbol
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10
-8
-6
-4
-2
0
2
4
6
8
10
E_s/N_0, dB

Probability Relations in a Channel

Consider a discrete information channel with r input symbols and s output symbols. We define the channel matrix $[P]$ with the conditional probabilities $P(b_j \vee a_i)$ as above:

$$[P]_{i,j} = P(b_j \vee a_i)$$

If the input symbol probabilities $P(a_i)$ define a row vector $a]$ where:

$$a]_i = P(a_i)$$

Then the output symbol probabilities $P(b_j)$ are given by the matrix equation:

$$b] = a][P]$$

where:

$$b]_j = P(b_j)$$

The *backward probabilities* $P(a_i \vee b_j)$ needed for the channel information rate calculation can be found from the *forward probabilities* $P(b_j \vee a_i)$ using Bayes' theorem:

$$P(a_i | b_j) = \frac{P(b_j | a_i) P(a_i)}{P(b_j)}$$

The probabilities $P(a_i)$ are also called the *a priori* probabilities of the input symbols – the input symbol probabilities *before* the reception of a given output symbol. The conditional probabilities $P(a_i \vee b_j)$ are also called the *a posteriori* probabilities of the input symbols – the input symbol probabilities *after* the reception of a given output symbol.

Shannon’s Noisy Coding Theorem

Suppose an information source is transmitting information at a rate R through a memoryless channel with capacity C. Shannon showed there is a coding system to transmit this information through the channel with an arbitrarily small error rate $\epsilon > 0$ when $R \leq C$. If $R > C$ it is not possible to transmit the information with an arbitrarily small error rate $\epsilon > 0$.

Granted not all real-world channels are well modeled as memoryless. Further, the channel capacity can be difficult to determine in the general case. Finally, Shannon’s theorem does not show how to find a specific code that satisfies his conclusion. That said the theorem is remarkable in that it proves the information rate R need not approach zero to achieve an arbitrarily small desired error rate. Instead the code block size n may need to increase along with the complexity of the coding and decoding algorithms. Today after seventy years of work there are viable codes which closely approach the limit imposed by Shannon’s theorem.

Summary

- All information transmission systems have the same five-element model consisting of a source, transmitter, channel, receiver and destination.

- The amount of information associated with an event is related to the probability of the occurrence of that event:

$$I(E) = -\log P(E)$$

If the logarithm is base-2 then the unit of information is the bit.

- A source produces a sequences of symbols. We consider two types of sources: zero-memory sources where each successive symbol is independent of all preceding symbols and Markov sources where successive symbols depend on some number of preceding symbols.

- The entropy of a source is the average amount of information per source symbol. For a zero-memory source the entropy is:

$$H(S) = -\sum_{S} P(s_i) \log P(s_i)$$

For a Markov source the entropy is:

$$H(S) = -\sum_{i=1}^{q^m} P(S\text{¿¿}i) \sum_{j=1}^{q} P(s_j \vee S_i) \log P(s_j \vee S_i)\text{¿}$$

The unit of entropy is bits/symbol or if a symbol rate is specified bits/second.

- An extension of a source is a block of *n* symbols from that source. For a zero-memory source the probability of a given sequence of symbols is the product of the probabilities of the individual symbols in the sequence since they are independent. For a Markov source the probability of a given sequence is found by using the conditional probabilities and stationary distribution of the source.

- The transmitter transforms information source symbols into waveforms suitable for transmission over the channel. Steps in this process may include formatting, source coding, encryption, channel coding and modulation. We consider only baseband systems where the information source symbols are converted into sequences of binary symbols (binits) and ultimately low-frequency electrical signals (voltage waveforms).

- Source coding attempts to more efficiently represent formatted information source symbols. The Huffman coding process generates a source code that is both compact and instantaneous. Its average length has a lower bound related to the source entropy and the size of the code alphabet:

$$L \geq \frac{H(S)}{\log r}$$

- Shannon’s Noiseless Coding Theorem states that by source coding extensions of an information source the average length of the source code equals the lower bound described above in the limit as the extension length approaches infinity:

  $$\lim_{n \to \infty} \frac{L_n}{n} = \frac{H(S)}{\log r}$$

- Efficiency, redundancy and compression ratio are various ways to quantify how close a given source code comes to the theoretical limit set by Shannon’s Noiseless Coding Theorem.

- The binary symmetric channel (BSC), binary erasure channel (BEC) and binary additive Gaussian white noise channel (BAWGNC) are three important models used for noisy, memoryless channels.

- The rate of information transmission through a noisy channel is given by the difference of the entropy of the source and the equivocation due to the channel:

$$R = H(A) - H(A \vee B)$$

- The capacity of an information channel is the maximum rate of information transmission:

$$C = max\, R$$

  where the maximum is with respect to all possible information sources that could be used as an input to the channel.

- Shannon’s Noisy Coding Theorem states that it is possible to transmit information at a rate R through a noisy channel at an arbitrarily small but nonzero error rate provided R is less than or equal to the channel capacity C. If R is greater than C then an arbitrarily small error rate is not possible.