Information Theory

Communication Viewpoint

Part 1

Charles Widger 6/10/2025

References

Suitable for Beginners

- **The Science of Information: From Language to Black Holes**, by Benjamin Schumacher, The Great Courses, Course No. 1301.
  - Available for no charge as video streaming from kanopy.com with course notes to download.
  - Does not require a broad knowledge of mathematics and physics, but familiarity with thermodynamics is helpful as is knowledge of probability.
  - This presentation is based largely on this course material.
  - A later presentation is planned to expand on error detection and correction. Some of this uses areas of modern algebra that even graduate courses may not cover.
- **An Introduction to Information Theory**, by Fazlollah M. Reza, Dover Publications, 1994
  - Dr. Fazlollah Reza (1915 - 1919) held 2 Ph.D. degrees in electrical engineering. One from Columbia University in 1946 and the other from Polytechnic University of New York in 1950. He taught at McGill University and Massachusetts Institute of Technology. He was a fellow of the IEEE and AAAS. He also served as ambassador for Iran to Canada and UNESCO.
- The book by Reza offers much greater detail and examples than the first reference. The course by Schumacher covers more recent innovations. Wikipedia also has several articles on the topic.

Further References

Not so much for Beginners

- **Error Correcting Codes**, Second Edition, by W. Wesley Peterson (1924 - 2009) and E. J. Weldon, Jr., 1972, MIT Press
- **IEEE Transactions on Information Theory**

Foundations

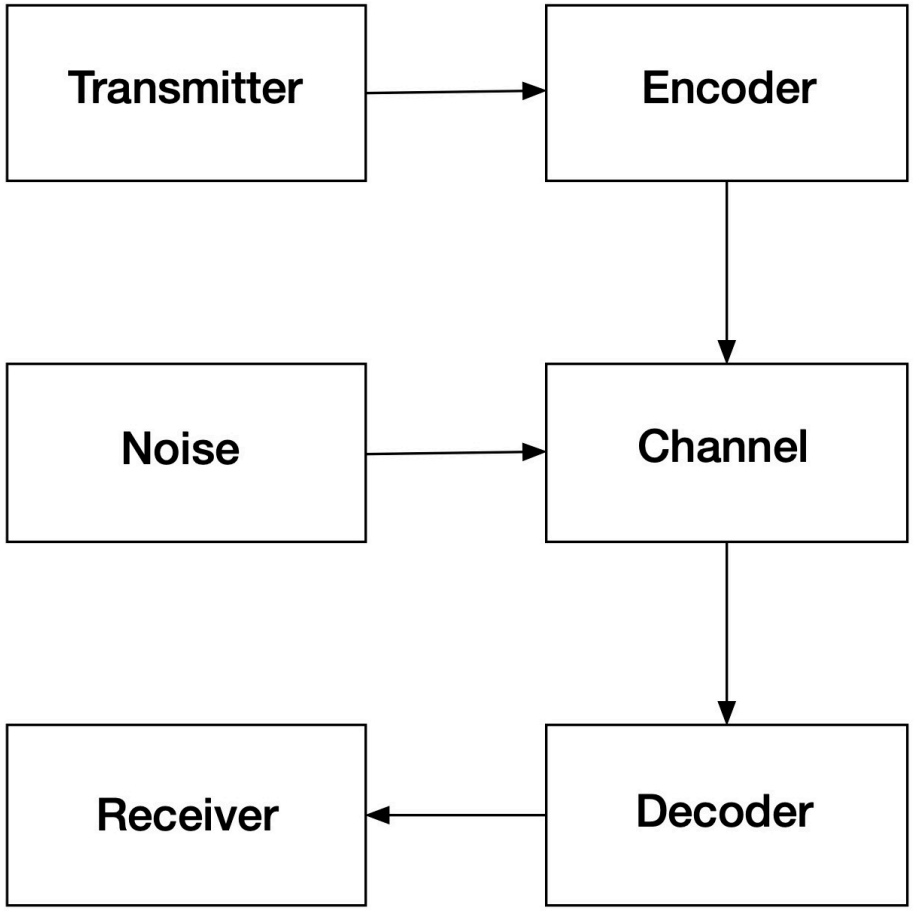
Historical Background

- The article by Claude E. Shannon *A Mathematical Theory of Communication* that appeared in 1948 in the two successive issues of the **Bell System Technical Journal** at **27** (3): pp. 379–423 and **27** (4): pp. 623–656 sets the foundation of Information Theory.
  - Shannon received a Ph.D. in mathematics from MIT in 1940 under Frank Lauren Hitchcock writing on *An Algebra for Theoretical Genetics*.
- He used some earlier results, two by Harry Nyquist, one by Ralph Hartley.
  - Harry Nyquist, *Certain factors affecting telegraph speed*, **Bell System Technical Journal**, 1924, **3** (2): 324–346.
  - Harry Nyquist, *Certain Topics in Telegraph Transmission Theory*, 1928, **Transactions of the American Institute of Electrical Engineers**, **7** (2).
  - Ralph Hartley, *Transmission of Information*, **Bell System Technical Journal**, 1924, **3** (2): 324–346.

Communication System

Major steps for Information Processing

- The Transmitter may handle a wide variety of sources: text, speech, art, music, photo, video, CT scan, are examples. Its output is called a message.
- The Encoder depends on the input source. Its input may become digitized and modulated. Its output will be considered as digits.
- The Channel may modify the message. A recording of a message on a tape may degrade with age. Speech can be become less audible with distance from the source.
- The Decoder undoes what the Encoder did. It may introduce its own modifications. For example, sight as received by the eye is always modified by the brain. The image received at the retina is at least inverted.
- The Receiver can have its own properties as well. Two different receivers may disagree. Information is the ability to distinguish reliably among possible alternatives. See here. Meaning, correctness, value, and significance are important different qualities not considered for information.



Basic Tenets

Encoding and Message Size

- Analogue signals will be digitized. To avoid aliasing, the sampling rate should meet the Nyquist-Shannon criterion, twice the signal frequency bandwidth.
- For this presentation, we will use binary data streams. Alan Turing (1912-1954) used decimals for the work done on the cryptanalysis of the German Enigma coder messages during World War II. There are differences in terminology between base 10 and base 2. Each digit in a binary message is called a “bit.”
- Coding can be chosen to reduce message size or encoding time. For some sources, lossy techniques can be tolerated.
  - Morse Code was chosen to make the most used letters take the least amount of time to encode. Prisoners in detention might use walls as a covert channel for tapping codes. These use time a distinction between letters. The tapping is an ordered pair, with a shorter time between taps than between letters. Typewriter keyboards were designed to be longer to use to type text.
  - Initial teletype used the 5-bit Baudot code. Signal transmission rate measured in *baud*. Digital data transmitted on analog telephone lines typically done by frequency shift keying. In later years, it was replaced by 6 bit Binary Coded Decimal (BCD), then by 7 bit American Standard Code for Information Interchange (ASCII). This now has been replaced by 8 bit Unicode. IBM created the 8 bit Extended Binary Coded Interchange Code (EBCDIC) for the introduction of the System/360 computers in 1964. Keypunch machines also had to be changed.
  - Indo-European languages use alphabets, the symbols of which can be expressed in Unicode (8 bits) or Braille (6 bits). Cherokee does not have an alphabet, but rather a syllabary of 86 symbols, so 7 binary bits would do for it. Native speakers can read and write almost immediately after learning the symbols.
  - I scanned a photo taken in 1892 of my maternal ancestral family that took 14 MB as a TIFF, but 111 KB as a JPEG. The name of the company that supplied their tent is easily legible in the TIFF, not in the JPEG. TIFF is loss-less, JPEG is lossy.
  - ZIP is an archive file format that provides loss-less data compression. It was released 14 Feb 1989 into public domain.

Basic Tenets

Measuring Information with Entropy

- Suppose the length of a message M is m encoded bits. Another such message would result in m^2 possibilities, although there are only 2 messages.
- This motivates the use of logarithms to capture the amount of information gained by receipt of the message M as $H = \log_2 m$. H is called the *entropy* of the message. Note $\log_2 m^2 = 2 \log_2 m$.
  - An alternate way to think of entropy is as the information not had before receipt of a message. For the Receiver, it starts at zero and increases with each received message.
  - The letter H is used to acknowledge the work of Ralph Hartley (1888-1970). This work was based on decimal digits. In Information Theory terms, each decimal digit may be called a "hartley."
  - The term entropy was coined by Claude Shannon because of its similarity to its properties as used in thermodynamics.
- Recall the ceiling function: $x \to \lceil x \rceil$, i.e, a real number x maps into the least integer greater or equal to x. Apply it to the entropy H. This places a bound on the number of bits needed to transmit the message.
  - Suppose we had 5 messages to send, say the letters A, B, C, D, E. Since $H(5) = \log_2(5) = 2.32$, so we need 3 bits to send a message one at a time. But, to send two messages together, we have $H(5^2) = 4.64$, so 5 bits are needed. Two messages are sent, so this amounts to 2.5 bits per message, closer to the entropy of the single message. This is an example of efficiency gain by "block coding." Sending 3 messages in a block does even better. $H(5^3) = 6.96$, so 7 bits will do, a mere 2.33 ... per message.

First Fundamental Theorem

Shannon Source Coding Theorem

- The entropy of the message establishes the fewest bits needed to encode a message.
- Different encodings can be chosen to reduce the message size.

Basic Tenets

Channels, Noise, and Channel Capacity

- A channel is any means by which information is conveyed. It has an input and an output. It can also be a place where information is stored. Examples include fiber optic cable, relays through satellites, both analog and digital magnetic tape, human and animal neural networks, over-the-air broadcasts of radio, television, cell telephone relays, USB memory sticks, CD ROM, DVDs, the air between a soapbox orator and those within ear-shot, and much more. The places where information may reside is staggering and shows how much it can be transformed.
- From the communication viewpoint, it is also a node in the message path where a message can be modified, turning "violence on TV" into "violins on TV." We classify all these changes as the introduction of noise. Noise can be very momentary, or last for days, such as with weather-related power outages, or sun-spot activity. Some environments can be very harsh, such as the electromagnetic radiation in diesel-electric train locomotives and radio communication with under-the-sea submarines. Disruptions with all kinds of noise are considered in terms of probability.
- A related matter is channel capacity: how fast bits can travel through the medium.
- We will illustrate the approach to these matters with two different examples, for which the probabilities can be computed with ease: a binary symmetric channel and a binary erasure channel, each with a one bit source message. The book by Reza covers both of these on p. 114.

Basic Tenets

Communication Channel

- We have already defined entropy of a message. If the message is likely to be something, it can be defined by a probability density function, *p*. The input *x*, and output *y*, are separate, and will be defined by a joint probability, *p(x,y)*. We use *conditional probability p(x|y)* to denote the probability of getting output y upon receiving input x. Here $p(y \mid x) = \frac{p(x, y)}{p(x)}$; that is, $p(x, y) = p(x) \times p(y \mid x)$.
- At the outset, before communication, the receiver knows neither the message sent, *X*, nor the output, *Y*. The amount lacked is *H(X, Y)*. During communication, it gets the output, *Y*, and learns its information, *H(Y)*. That leaves the *conditional entropy*, *H(X|Y)*.
- A more positive measure is the amount of information gained, the *mutual information*, *I(X;Y) = H(X) - H(X|Y)*. This is illustrated in the Venn diagram.
  - The left circle represents *H(X)*, the right *H(Y)*, the two together, *H(X,Y)*. The part of *H(X)* not in *H(Y)* is *H(X|Y)*, and conversely, the part of *H(Y)* not in *H(X)* is *H(Y|X)*. The mutual information is portion shared by both *H(X)* and *H(Y)*.

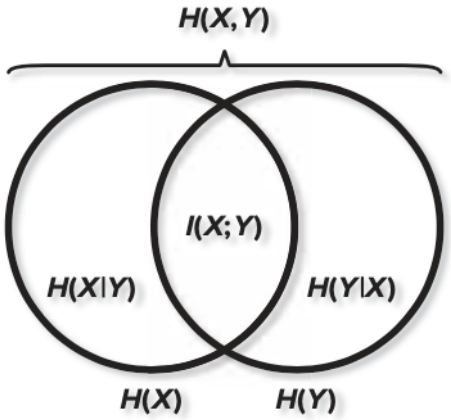
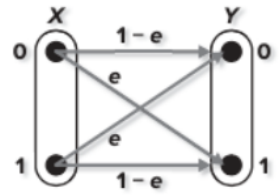


Illustration 1

Binary Symmetric Channel

- Alice and Bob are at a noisy party, and are separated by a some space. Alice sends Bob a single bit message, either *yes*, or *no*. Bob thinks she said yes, but he couldn't hear it well. He thinks she said yes because her lips seemed less rounded than they would be if she had said no. So he is 90% sure it is a yes, and so he thinks his error rate is 10%. Call this number p.
- The channel is binary because there are only 2 possible messages, and symmetric because the probability of error is the same in both cases.
- Reza gives the formula $H(Y \mid X) = -\left(e \times \log_2 e + (1-e) \times \log_2(1-e)\right)$. There is also an Appendix providing a table for these values. He also provides formulas for mutual information: $I(X;Y) = H(Y) + e \times \log_2 e + (1-e) \times \log_2(1-e)$ and channel capacity: $C = 1 + e \times \log_2 e + (1-e) \times \log_2(1-e)$.
- Despite Bob's error rate, which seems pretty low, *H(Y|X)* is 0.469 according to Schumacher and 0.468 according to Reza, about half a bit. Bob's still lacking a lot of information.



Binary Symmetric Channel (BSC)

U W Physics Exams

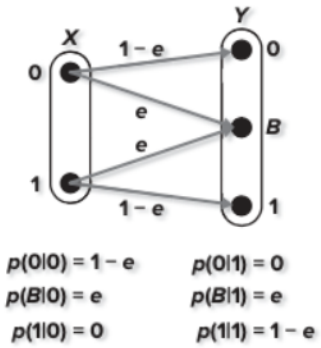
Scenario for Illustration 2

- When I started as an undergraduate in 1961, I enrolled Physics 121, 122, 123, first year physics for science majors. For each of the terms of instruction, the professor gave final exams consisting to 100 true/false questions.
- Questions could be challenging: is the result 1.4 or is it $\sqrt{2}$? The correct answers were posed in a quasi-random manner. One technique might be work the problems one knew, count the numbers of confident true answers and false answers, and guess the rest to balance out the two. This models binary symmetric channel.
- The rub was in the test grading. If the student got the answer right, 1 was added to the cumulative score. If wrong, 1 point was deducted. The final score was the absolute value of the total. The expected score for just guessing everything was about 0. If one were smart enough to get them all wrong, you were regarded to be knowing enough to get them all right, so you got 100%.
- The advice was, if you didn't know, leave the question blank. This models the binary erasure channel. For each input message, there are three outputs, yes, no, and no repsonse.
- I think the professor used the results as a means of identifying areas to improve his lectures.

Illustration 2

Binary Erasure Channel

- The channel is binary because there are only 2 possible messages input messages, but a possible output may erased.
- The input entropy $H(X)$ and joint entropy $H(X, Y)$ remain the same as with BSC. But now, the 3-way output entropy $H(X \mid Y) = e$ bits per question, and the mutual information is $I(X; Y) = (1 - e)$ bits per question. Knowing what questions are not guessed is a big help.
- So, if Charlie takes my professor's test, and guesses instead of leaving blank, with a 0.1 error rate, he has a entropy of $H(X \mid Y) = 0.468$. If Diane leaves the 0.1 of the questions blank where she has to guess, she has entropy $H(X \mid Y) = 0.9$. It is a lot easier to help Diane because the professor knows what she had trouble with.



Binary Erasure Channel (BEC)

Second Fundamental Theorem

Noisy Channel Coding Theorem

- Channel capacity establishes the maximum communication rate.
- Exceeding the channel capacity drives the probability of error to dead certain.
- The theorem also guarantees the existence of error correcting codes, but does not say how to create them.
  - One simple kind is use of Phonetic Letters Spelled Out (PLSO): Alpha, Bravo, Charlie, Delta, ... Errors might be detected by misspellings. "..pha" might reasonably turn into Alpha.

Error Dectection and Correcting Codes

Dealing with noisy channels

- Natural languages have redundancies which help to reduce errors in understanding.
  - The title of Hemingway's book is **For Whom the Bell Tolls**, the declension of "who" to "whom" reinforces its grammatical use. It is not subject (what does the tolling - the bell), nor the direct object (not mentioned - assumed to be a service for death), but rather the indirect object (those acknowledged by the tolling being made).
  - Jewish scribes in the 7th - 10th centuries made counts of words in a line, book, section, and other data in the Torah. Information was kept in the Numerical Masorah to help insure accuracy in copies. Discovery of the Dead Sea Scrolls (1947 - 1956) helped confirm the effectiveness of these data over the centuries.
- An example of a coding redundancy is the introduction of a parity bit in a binary data stream. This can accomplish single error detection, but not always. The parity bit may itself have been flipped. That might be confirmed by examining the block the count applies to, but it still leaves open whether more than one data bit is in error.
- An earlier contributor to binary error dectection and correction is Richard Hamming (1915 - 1998). In 1950, he published his family of codes. An explanation of why they work depends on an understanding of vector spaces over Galois Fields modulo 2. We will speak to the Hamming(7,4) code as an illustration, next. It works on 4-bit blocks of code, to which it adds 3 parity bits. Each parity bit covers 3 data points. This code can detect a single bit error and correct it, and can detect a 2 bit error, but cannot correct either of them. It is useful when errors are infrequent. This code is used in error-correcting memory in both conventional and quantum computers.

Hamming (7,4) Code

Data Coded in 4 bit Blocks with 3 parity bits added

- Define a *code generator matrix* **G** and a *parity check matrix* **H**:

  $G^T := \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $H := \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$. Then, given a 4 block of data vector $q := \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix}$,

  take the matrix product $G^T q$ modulo 2 and transmit it as a codeword x for p. Note the transformation $G^T : GF(2)^4 \mapsto GF(2)^7$ and sends a basis in the domain vector space to a basis in the range. H strips off the parity bits in the received response, also known as the *syndrome*.

- Codewords are transmitted in the following order: $p_1, p_2, d_1, p_3, d_2, d_3, d_4$.

- If no error is detected, the received data r will be mapped into the kernel of H; i.e. Hr will be the null vector. If it is not the null vector, the product Hr will be the position of the bit that was flipped. So $Hr = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ means a error in the 5th bit transmitted, which happens to be d_2.

BCH Error Correcting Codes

BR-CH might be a more accurate name

- This set of codes was first proposed by Alexis Hocquenghem in 1959. They were also published in 1960 by Raj Chandra Bose and his Ph.D. student Dwijendra Kumar Ray-Chaudhuri from his dissertation, awarded also in 1959. They were at the University of North Carolina at Chapel Hill.
- A cyclic code is a block code where every right (or left) circular shift of a codeword is also a codeword.
  - Theorem: In the algebra of polynomials modulo $X^n - 1$, a subspace is cyclic if and only if it is an ideal. This ideal is principle, hence generated by an unique polynomial, called the *generator*.
  - The Hamming(7,4) is a cyclic code with generator $g(X) = X^3 + X^2 + 1$.
- The distance between distinct codewords is defined as the number of positions by which the two differ. Every code has a distance of at least 1. The greater the distance a code has, the more errors it can detect and correct.
- A BCH code allows specification of a minimal distance for the code. It also has an easy way to decode the syndrome.
  - Let α be an element of $GF(q^m)$, q a prime, For any specified d such that $d \leq q^m - 1$, the code generated by $g(X)$ is a BCH code if and only if g(X) is a polynomial of lowest degree over $GF(q)$ for which $\alpha^m, \alpha^{m+1}, \ldots, \alpha^{m+d-2}$ are roots.
  - The BCH code with generator $g(x) = x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1$ has 5 data bits and 10 checksum bits. It is known as a (15,5) BCH code. It is used for the QR code. It has Hamming distance at least 7 and corrects up to 3 errors.

Reed-Solomon Codes

These have wide application

- Irving S. Reed (1923-2012)and Gustave Solomon (1930-1996) defined these codes in 1960 while on the staff of the MIT Lincoln Laboratory. Their work was published in the Journal of the Society of Industrial and Applied Mathematics (SIAM).
- Consider messages in $GF(q^m)$ where q is prime. Let k be the input message length and n be the total output message length and set $t = n - k$, the number of checksums to be added to the input code. A Reed-Solomon code can detect up to t errors or detect and correct up to $\lfloor t/2 \rfloor$ errors at unknown locations. As an erasure code, it can correct up to t erasures at known locations. It can also correct combinations of errors and erasures. These codes are also suitable as a multiple burst bit error correcting code.
- As originally defined, finding a suitably efficient decoder, remained a issue. This was initially addressed by using a BCH-like scheme. In 1986, the Berlenkamp-Welch algorithm took care of the matter. Ten years on, 1996 a listing of decoders was published.
- Reed Solomon codes may been interleaved to increase performance.
- Applications using Reed-Solomon codes include data storage, bar code, data transmission, and space transmission systems.

Information Theory

End Part 1

Charles Widger 6/10/2025