# The Noisy-Channel Coding Theorem

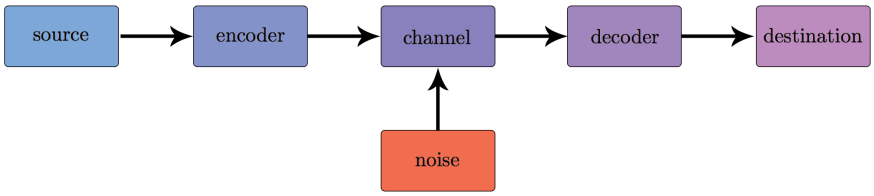
Michael W. Macon

December 18, 2015

**Abstract**

This is an exposition of two important theorems of information theory often singularly referred to as *The Noisy-Channel Coding Theorem*. Given a few assumptions about a channel and a source, the coding theorem demonstrates that information can be communicated over a noisy channel at a non-zero rate approximating the channel's capacity with an arbitrarily small probability of error. It originally appeared in C.E. Shannon's seminal paper, *A Mathematical Theory of Communication*, and was subsequently rigorously reformulated and proved by A.I. Khinchin. We first introduce the concept of entropy as a measure of information, and discuss its essential properties. We then state *McMillan's Theorem* and attempt to provide a detailed sketch of the proof of what Khinchin calls *Feinstein's Fundamental Lemma*, both crucial results used in the proof of the coding theorem. Finally, keeping in view some stringent assumptions and Khinchin's delicate arguments, we attempt to frame the proof of *The Noisy-Channel Coding Theorem*.

## 1 Introduction




C.E. Shannon wrote in 1949,

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point[5]."

At the time Shannon wrote his celebrated paper, *A Mathematical Theory of Communication*, the efforts of communications engineers were beginning to shift from analog transmission models to digital models. This meant that rather than

focusing on improving the signal-to-noise ratio by means of increasing the amplitude of the waveform signal, engineers were concerned with efficiently transmitting a discrete time sequence of symbols over a fixed bandwidth. Shannon's paper piqued the interest of engineers because it describes what can be attained by regulating the encoding system. It was Norbert Wiener in [7] who first proposed that a message should rightly be viewed as a stochastic process[3]. R.V.L. Hartley proposed the logarithmic function for a measure of *information*[5]. However, it was Shannon who formalized the theory by giving mathematical definitions of *information*, *source*, *code* and *channel*, and a way to quantify the information content of sources and channels. He established theoretical limits for the quantity of information that can be transmitted over noisy channels and also for the rate of its transmission.

At the same time, mathematicians and statisticians became interested in the new theory of information, primarily because of Shannon's paper[5] and Wiener's book[7]. As McMillan paints it, *information theory* "is a body of statistical mathematics." The model for communication became equivalent to statistical parameter estimation. If $x$ is the transmitted input message, and $y$ the received message, then the joint distribution of $x$ and $y$ completely characterizes the communication system. The problem is then reduced to accurately estimating $x$ given a joint sample $(x, y)$[3].

*The Noisy-Channel Coding Theorem* is the most consequential feature of information theory. An input message sent over a *noiseless channel* can be discerned from the output message. However, when noise is introduced to the channel, different messages at the channel input can produce the same output message. Thus noise creates a non-zero probability of error, $P_e$, in transmission. Introducing a simple coding system like a repetition code or a linear error correcting code will reduce $P_e$, but simultaneously reduce the rate, $R$, of transmission. The prevailing conventional wisdom among scientists and engineers in Shannon's day was that any code that would allow $P_e \to 0$ would also force $R \to 0$. Nevertheless, by meticulous logical deduction, Shannon showed that the output message of a source could be encoded in a way that makes $P_e$ arbitrarily small. This result is what we shall call the *Noisy Channel Coding Theorem Part 1.* Shannon further demonstrated that a code exists such that the rate of transmission can approximate the capacity of the channel, *i.e.* the maximum rate possible over a given channel. This fact we shall call *The Noisy Channel Coding Theorem Part 2.* These two results have inspired generations of engineers, and persuaded some to confer the title of "Father of the Information Age" to Claude Shannon.

As Khinchin narrates, the road to a rigorous proof of Shannon's theorems is "long and thorny." Shannon's initial proofs were considered by mathematicians to be incomplete with "artificial restrictions" that weakened and oversimplified them. In the years after the publication of Shannon's theorem, research mathematicians such as McMillan, Feinstein *et al.*, began to systematize information theory on a sound mathematical foundation[2]. Now we attempt to

follow Khinchin down the lengthy and complicated road to the proof of the *Noisy Channel Coding Theorem.*

## 2 Entropy

In his famous paper [5] Shannon proposed a measure for the amount of uncertainty or *entropy* encoded in a random variable. If a random variable $X$ is realized, then this uncertainty is resolved. So it stands to reason that the entropy is proportional to the amount of information gained by a realization of random variable $X$. In fact, we can take the amount of information gained by a realization of $X$ to be equal to the entropy of $X$. As Khinchin states,

> "...the information given us by carrying out some experiment consists in removing the uncertainty which existed before the experiment. The larger this uncertainty, the larger we consider to be the amount of information obtained by removing it[2]."

Suppose $\mathcal{A}$ is a finite set. Let an element $a$ from set $\mathcal{A}$ be called a letter, and the set $\mathcal{A}$ be called an alphabet.

**Definition 1.** *Let $X$ be a discrete random variable with alphabet $\mathcal{A}$ and probability mass function $p_X(a)$ for $a \in \mathcal{A}$. The entropy $H(X)$ of $X$ is defined by*

$$H(X) = -\sum_{a \in \mathcal{A}} p(a) \cdot \log p(a)$$

Note: The base of the logarithm is fixed, but may be arbitrarily chosen. Base 2 will be used in the calculations and examples throughout this paper. In this case, the unit of measurement is the *bit*. It is also assumed that $0 \cdot \log(0) = 0$.

For example, the entropy of the binomial random variable $X \sim B(2, 0.4)$ is given by

$$H(X) = -(0.36 \cdot \log 0.36 + 0.48 \cdot \log 0.48 + 0.16 \cdot \log 0.16) = 1.4619\, bits.$$

The entropy of a continuous random variable $X$ with probability density function $p(x)$ is given by

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx.$$

The entropy of a continuous random variable is sometimes called *differential entropy.*
As mentioned above, the entropy $H(X)$ is interpreted as a measure of the uncertainty or information encoded in a random variable, but it is also importantly viewed as the mean minimum number of binary distinctions—yes or no questions—required to describe $X$. So $H(X)$ is also called the *description length.*

The definition of entropy can be extended to two random variables.

**Definition 2.** *The joint entropy $H(X,Y)$ of a pair of discrete random variables $(X,Y)$ with a joint distribution $p(a,b)$ is defined as*

$$H(X,Y) = -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a,b)$$

**Definition 3.** *The conditional entropy $H(Y|X)$ is defined as*

$$\begin{aligned} H(Y|X) &= -\sum_{a\in\mathcal{A}} p(a)\cdot H(Y|X=a) \\ &= -\sum_{a\in\mathcal{A}} p(a)\cdot\sum_{b\in\mathcal{B}} p(b|a)\log p(b|a) \\ &= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\cdot\log p(b|a) \end{aligned}$$

The conditional entropy $H(Y|X)$ is a random variable of $X$, and is the expected value of $H(Y)$ given the realization of $X$. That is, $H(Y|X)$ is the expected value of the amount of information to be gained from the realization of $Y$ given the information gained by the realization of $X$.

**Theorem 1.** $H(X,Y) = H(X) + H(Y|X)$

**Proof.**

$$\begin{aligned} H(X,Y) &= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a,b) \\ &= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a)p(b|a) \\ &= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a) - \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(b|a) \\ &= -\sum_{a\in\mathcal{A}} p(a)\cdot\log p(a) - \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(b|a) \\ &= H(X) + H(Y|X) \quad \square \end{aligned}$$

**Corollary 1.** *If $X$ and $Y$ are independent, it follows immediately that*

$$H(X,Y) = H(X) + H(Y).$$

**Corollary 2.** *We also have $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.*

**Corollary 3.** *Note that $H(X) - H(X|Y) = H(Y) - H(Y|X)$.*

We now prove a useful inequality.

**Jensen's Inequality.**

If $f(x)$ is a real-valued, convex continuous function and $p_1, p_2, ..., p_n$ are positive numbers such that $\sum_{i=1}^{n} p_i = 1$ and $x_1, x_2, ..., x_n \in \mathbb{R}$, then

$$f(p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_n \cdot x_n) \leq p_1 \cdot f(x_1) + p_2 \cdot f(x_2) + ... + p_n \cdot f(x_n)$$

**Proof of Jensen's Inequality.**

Jensen's Inequality can be proved by mathematical induction. It is assumed that $f$ is convex, and for the case of $n = 2$ we have

$$f(p_1 \cdot x_1 + p_2 \cdot x_2) \leq p_1 \cdot f(x_1) + p_2 \cdot f(x_2),$$

which is the condition for a convex function. Now assume the inequality is true for $n = k$. Then we have

$$f\left(\sum_{i=1}^{k} p_i \cdot x_i\right) = f(p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_k \cdot x_k) \leq p_1 \cdot f(x_1) + p_2 \cdot f(x_2) + ... + p_k \cdot f(x_k)$$

And consider the case when $n = k + 1$:

$$f\left(\sum_{i=1}^{k+1} p_i \cdot x_i\right)$$

$$= f(\sum_{i=1}^{k} p_i \cdot x_i + p_{k+1} \cdot x_{k+1})$$

$$= f(p_{k+1} \cdot x_{k+1} + (1 - p_{k+1}) \cdot \sum_{i=1}^{k} \frac{p_i}{1 - p_{k+1}} \cdot x_i)$$

$$\leq p_{k+1} \cdot f(x_{k+1}) + (1 - p_{k+1}) \cdot f\left(\sum_{i=1}^{k} \frac{p_i}{1 - p_{k+1}} \cdot x_i\right) \text{ (since f is convex)}$$

$$= p_{k+1} \cdot f(x_{k+1}) + f\left(\sum_{i=1}^{k} p_i \cdot x_i\right)$$

$$\leq p_{k+1} \cdot f(x_{k+1}) + \sum_{i=1}^{k} p_i \cdot f(x_i) = \sum_{i=1}^{k+1} p_i \cdot f(x_i)$$

**Theorem 2.** *For any two random variables $X$ and $Y$, $H(X,Y) \leq H(X)+H(Y)$ where $H(X,Y) = H(X) + H(Y)$ if and only if $X$ and $Y$ are independent.*

**Proof.**

For the proof, bear in mind that a function is convex if its second derivative is non-negative. $f(x) = x\log_b(x)$ is convex since $f''(x) = \dfrac{1}{x\log(b)} > 0$. Also seeing that $f(x) = x\log_b(x)$ is continuous, we can apply Jensen's theorem.

$$H(X,Y) - H(X) - H(Y)$$

$$\begin{aligned}
&= -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a,b) + \sum_{a\in\mathcal{A}} p(a)\log p(a) + \sum_{b\in\mathcal{B}} p(b)\log p(b) \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log\left(\frac{1}{p(a,b)}\right) + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(a) + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log p(b) \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log\frac{p(a)p(b)}{p(a,b)} \\
&\leq \log\left(\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)p(b)\right) \text{ (by Jensen's theorem)} \\
&= \log(1) = 0
\end{aligned}$$

It is evident that if $X$ and $Y$ are independent we have $p(a)p(b) = p(a,b)$ and the equality holds. □

It can also be demonstrated that the amount of uncertainty encoded in $X$ will stay the same or diminish given the information gained from the realization of $Y$. This is demonstrated in the following theorem.

**Theorem 3.** *For any two random variables $X$ and $Y$*

$$H(X|Y) \leq H(X).$$

**Proof.** From theorems 1 and 2, it follows that

$$H(X|Y) = H(X,Y) - H(Y) \leq H(X) + H(Y) - H(Y) = H(X)$$

. □

## 2.1 Properties of Entropy

The entropy function described in definition 1 is, in fact, the most fitting measure for the uncertainty of a random variable. If $p_X(a) = 1$ for any $a \in \mathcal{A}$, then the uncertainty of the random variable will be 0. Otherwise the uncertainty will be positive. So the amount of uncertainty, *i.e.* the information encoded in a random variable, is a non-negative function. Furthermore, if all of the probabilities $p_X(a)$ for all $a \in \mathcal{A}$ are equal, then, accordingly, the uncertainty of the random variable will achieve a maximum.

Here are some necessary properties of entropy required for any sound measure of information or uncertainty:

1. $H(X) \geq 0$
2. $H(X)$ is a continuous function of $P = \{p_X(a)\}_{a \in \mathcal{A}}$
3. $H(X)$ is a symmetric function of $P$, *i.e.* the $p_X(a)$'s can be permuted.
4. $H(X)$ can be expanded. That is, if some $a$ with 0 probability is added to the alphabet, the entropy will not change.
5. If $p_X(a) = 1$ for some $a \in \mathcal{A}$, then the uncertainty vanishes and $H(X) = 0$.
6. $H(X)$ is maximum when $X$ is uniformly distributed on the alphabet $\mathcal{A}$.
7. $H(X, Y) = H(X) + H(Y|X)$

Furthermore, Khinchin shows that if a function $H(X)$ is continuous and has properties #4, #6 and #7, and $\lambda$ is a positive constant, then

$$H(X) = -\lambda \cdot \sum_{a \in \mathcal{A}} p(a) \cdot \log p(a)$$

is unique. In other words, $H(X)$ is the only reasonable measure for information with the desired properties that is possible[2].

**Theorem 4.** *For a discrete random variable $X$ with alphabet $\mathcal{A} = \{a_1, a_2, ..., a_n\}$, $H(X) \leq \log n$. That is, the maximum value of $H(X)$ is $\log n$.*

**Proof.** For this proof keep in mind that $\ln x \leq x - 1$, which can be easily ascertained graphically. Given this, and that $\log x = \dfrac{\ln x}{\ln 2}$, it follows that

$$\ln x \leq (x-1) \iff \frac{\ln x}{\ln 2} \leq \frac{x-1}{\ln 2} \iff \log x \leq \frac{x-1}{\ln 2}$$

$$\iff \log x \leq (x-1) \cdot \frac{\ln e}{\ln 2} \iff \log x \leq (x-1) \log e.$$

So $H(X) - \log n$

$$
\begin{aligned}
&= -\sum_{a\in\mathcal{A}} p(a)\log p(a) - \log n \\
&= -\sum_{a\in\mathcal{A}} p(a)\log p(a) - \sum_{a\in\mathcal{A}} p(a)\log n \\
&= -\sum_{a\in\mathcal{A}} p(a)(\log p(a) + \log n) \\
&= \sum_{a\in\mathcal{A}} p(a)\log\left(\frac{1}{np(a)}\right) \\
&\leq \sum_{a\in\mathcal{A}} p(a)\left(\frac{1}{np(a)} - 1\right)\log e \text{ (using the identity listed above)} \\
&= \left(\sum_{a\in\mathcal{A}} \frac{1}{n} - \sum_{a\in\mathcal{A}} p(a)\right)\log e \\
&= \left(n\frac{1}{n} - 1\right)\log e \\
&= 0 \quad \square
\end{aligned}
$$

We will not prove all seven properties of entropy listed above, as these can be easily found in the literature. However, below we give an elegant demonstration of property #6.

**Proof.** Let $X$ be a random variable with alphabet $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ such that $p_X(a_i) = \frac{1}{n}$ for all $i$. Then it follows that H(X)

$$
\begin{aligned}
&= -\sum_{a\in\mathcal{A}} p_X(a)\log p_X(a) \\
&= -\sum_{a\in\mathcal{A}} \frac{1}{n}\log\frac{1}{n} \\
&= -(\frac{1}{n}\log\frac{1}{n})\sum_{a\in\mathcal{A}} 1 \\
&= -(\frac{1}{n}\log\frac{1}{n})(n) = -\log\frac{1}{n} \\
&= \log(n) \quad \square
\end{aligned}
$$

# 3 McMillan's Theorem, Sources and Channels

We now give a formal definition of *source* and *channel*, and introduce the concept of a *compound channel*. We also prove a consequential lemma, which demonstrates that sequences of trials of a Markov chain can be partitioned into two

distinct sets. Lastly, we state McMillan's theorem, a central result used in the proof of Feinstein's Fundamental Lemma and the Noisy Channel Coding Theorem. Since an information source can be represented by a sequence of random variables, and can be modeled with a Markov chain[1], we begin by defining the entropy of a Markov chain.

## 3.1 Entropy and Markov Chains

Let $\{X_l\}$ be a finite stationary Markov chain with $n$ states. Suppose the transition matrix of the chain is $Q = (q_{i,k})$ where $i, k = 1, 2, 3, ...n$. Let $P(X_l = k | X_{l-1} = i) = q_{i,k}$ denote the probability of state $l$ where $l = 1, 2, 3, ...n$. Recall that for any stationary Markov chain with finitely many states, a stationary distribution $\vec{\mu}$ exists for which $\vec{\mu}Q = \vec{\mu}$.

**Definition 4.** *Suppose $\{X_l\}$ is a stationary Markov chain with transition matrix $Q = (q_{i,k})$ and stationary distribution $\vec{\mu} = (\mu_1, \mu_2, ..., \mu_n)$ with $i, k = 1, 2, 3, ...n$. Let the distribution of $X_1$ be $\vec{\mu}$. The entropy rate of the Markov chain is given by*

$$H(\mathcal{A}) = -\sum_{i,k} \mu_i q_{i,k} \log q_{i,k}$$

The entropy rate of a Markov chain is the average transition entropy when the chain proceeds one step.

We will follow the definition and theorem stated in Khinchin.

**Definition 5.** *Let $P_k$ be the probability of state $k$ in a Markov chain and suppose that the relative frequency of state $k$ after a sufficiently large number of trials $s$ is given by $\frac{m_k}{s}$. A Markov chain is called ergodic if*

$$P\{\left|\frac{m_k}{s} - P_k\right| > \delta\} < \epsilon$$

*for arbitrarily small $\epsilon > 0$ and $\delta > 0$, and for sufficiently large $s$.*

Consider the set of all possible sequences of $s$ consecutive trials of a Markov chain. An element from this set can be represented as a sequence $k_1, k_2, ...k_s$ where $k_1, k_2, ...k_s$ are numbers from $\{1, 2, ..., n\}$. Let $C$ be an arbitrary sequence from this set. Then

$$p(C) = P_{k_1} p_{k_1 k_2} p_{k_2 k_3} ... p_{k_{s-1} k_s} = P_{k_1} \prod_{i=1}^{n} \prod_{l=1}^{n} (p_{il})^{m_{il}}$$

where $i, l = 1, 2, ...n$, $m_{il}$ is the number of pairs $k_r k_{r+1}$ with $1 \leq r \leq s-1$ and $k_r = i, k_{r+1} = l$.

## 3.2 A Partitioning Lemma

**Lemma 1.** ***A Partitioning Lemma***
*Let $\epsilon > 0$ and $\eta > 0$ be arbitrarily small numbers. For sufficiently large $n$, all sequences of the form $C$ can be divided into two sets with the following properties:*
*1) The probability of any sequence in the first group satisfies the inequality*

$$\left| \frac{\log \frac{1}{p(C)}}{s} - H \right| < \eta$$

*2) The sum of the probabilities of all sequences of the second group is less than $\epsilon$.*

The following is a proof of property 1 in the above lemma.

*Proof.* A sequence $C$ is in the first group if it satisfies the following:

1) $p(C) > 0$ and
2) $|m_{il} - sP_i p_{il}| < s\delta$
If $C$ is in the first group, we have $m_{il} = sP_i p_{il} + s\delta\theta_{il}$ where
$-1 < \theta_{il} < 1, 1 \leq i \leq n, 1 \leq l \leq n$.
Let $*$ denote the restriction of the product to non-zero factors.

Now $p(C) = P_{k_1} \prod_i \prod_l *(p_{il})^{sP_i p_{il} + s\delta\theta_{il}}$.

Thus $\log \dfrac{1}{p(C)}$
$= -\log P_{k_1} - s \sum_i \sum_l *P_i p_{il} \log p_{il} - s\delta \sum_i \sum_l *\theta_{il} \log p_{il}$
$= -\log P_{k_1} + sH - s\delta \sum_i \sum_l *\theta_{il} \log p_{il}$
Rearranging, we have
$\left| \dfrac{\log \frac{1}{p(C)}}{s} - H \right| < \dfrac{1}{s} \log \dfrac{1}{P_{k_1}} + \delta \sum_i \sum_l * \log \dfrac{1}{p_{il}}$.
Consequently, for sufficiently large $s$, sufficiently small $\delta$ and $\eta > 0$, we have that

$$\left| \frac{\log \frac{1}{p(C)}}{s} - H \right| < \eta$$

□

## 3.3 Sources

A discrete source generates messages, which are defined to be ordered sequences of symbols from a finite alphabet. Messages are considered to be a random stochastic process. Let $I$ denote the set of integers $(... -1, 0, 1, 2, ...)$ and $\mathcal{A}^I$ denote the class of infinite sequences $x = (..., x_{-1}, x_0, x_1, x_2, ...)$ where each

$x_t \in \mathcal{A}$ and $t \in I$. A set of events of the type $x$ is called a cylinder set. An information source consists of a probability measure $\mu$ defined over the $\sigma$-algebra or Borel field of subsets of $\mathcal{A}^{\mathcal{I}}$, denoted $F_{\mathcal{A}}$, along with the stochastic process $[\mathcal{A}^I, F_{\mathcal{A}}, \mu]$. We simply denote the source $[\mathcal{A}, \mu]$.

**Definition 6.** *If $S \subset \mathcal{A}^I$ implies $\mu(shift(S)) = \mu(S)$, a source $[\mathcal{A}, \mu]$ is called stationary, where $S$ is any set of elements $x$, $shift(S)$ is the set of all $shift(x)$, and $shift : x_k \to x_{k+1}$ is a shift operator.*

**Definition 7.** *If $shift(S) = S$, then $S$ is called an invariant set. A stationary source $[\mathcal{A}, \mu]$ is ergodic if $\mu(S) = 0$ or $\mu(S) = 1$ for every invariant set.*

Mirroring definition 1, we have that the quantity of information encoded in the set of all $n$-term sequences consisting of $a^n$ events $C$ with probabilities $\mu(C)$ is given by

$$H_n = -\sum_C \mu(C) \log \mu(C)$$

For a stationary source the expected value of the function $f_n(x) = -\frac{1}{n} \log \mu(C)$ is given by

$$\mathbb{E}[-\frac{1}{n} \log \mu(C)] = -\frac{1}{n} \sum_C \mu(C) \log \mu(C) = \frac{H_n}{n}$$

## 3.4 McMillan's Theorem

A key result can be summarized as follows: as $n \to \infty$ we have

$$\mathbb{E}(f_n(x)) \to H,$$

where $f_n(x) = -\frac{1}{n} \log \mu(C)$. Furthermore, for arbitrarily small $\epsilon > 0$, $\delta > 0$ and sufficiently large $n$, we also have

$$P\{|f_n(x) - H| > \epsilon\} < \delta.$$

So $f_n(x)$ converges in probability to $H$ as $n \to \infty$[2]. The entropy of an information source is defined to be

$$H = \lim_{n \to \infty} \frac{H_n}{n},$$

and is the average amount of information gained per symbol produced by the source. On the other hand, the source entropy also represents the average uncertainty per symbol produced by the source[3]. The limit

$$H = \lim_{n \to \infty} \frac{H_n}{n}$$

always exists—an important conclusion first published by Shannon[5].

**Definition 8.** *A source is said to have the E-property if all n-term sequences C in the output of the source can be separated into two groups such that*
*1) For every sequence C of the first group*

$$\left|\frac{\log \mu(C)}{n} + H\right| < \epsilon$$

*2) The sum of the probabilities of all the sequences of the second group is less than $\delta$, where $\epsilon > 0$ and $\delta > 0$ are arbitrarily small real numbers.*

It can be demonstrated that any ergodic source has the $E$ property. This is referred to as McMillan's Theorem after Brockway McMillan, who first published this result[3]. It parallels the *Partitioning Lemma* listed above.

**McMillan's Theorem**

*For arbitrary small $\epsilon > 0, \delta > 0$ and sufficiently large $n$, all the $a^n$ n-term sequences of the ergodic source output are divided into two groups , a high probability group, such that $|\frac{1}{n}\log \mu(C) + H| < \epsilon$, for each of its sequences, and a low probability group, such that the sum of the probabilities of its sequences is less than $\delta$.*

McMillan's proof of the existence of such high probability sequences is crucial to the proof of *Feinstein's Fundamental Lemma*, which relates the cardinality of the set of so-called *distinguishable sequences* to the ergodic capacity of the channel. This, in turn, is the key to Shannon's proof that there exists a coding system that allows information to be transmitted across any ergodic channel with an arbitrarily small probability of error. We now give a precise definition for *communication channel.*

## 3.5 Channels

A communication channel $[\mathcal{A}, \nu_x, \mathcal{B}]$ transmits a message of symbols from the input alphabet $\mathcal{A}$ of a source, and outputs the possibly corrupted message consisting of symbols from the output alphabet $\mathcal{B}$. The channel is characterized by a family of probability measures denoted by $\nu_x$, which are conditional probabilities that the signal received when a given sequence is transmitted belongs to the set $S \subset \mathcal{B}^{\mathcal{I}}$, where $\mathcal{B}^{\mathcal{I}}$ is the set of all output sequences $y$.

**Definition 9.** *A channel $[\mathcal{A}, \nu_x, \mathcal{B}]$ is stationary if for all $x \in \mathcal{A}^{\mathcal{I}}$ and $S \in \mathcal{B}^{\mathcal{I}}$*

$$\nu_{shift(x)}(shift(S)) = \nu_x(S),$$

*where $shift : x_k \to x_{k+1}$ is a shift operator.*

If for a given channel, the first received sequence is unaffected by all the sequences transmitted after the first, then the channel is called non-anticipating.

**Definition 10.** *A channel is said to be without anticipation if the distribution of message $y_n$ is independent of messages transmitted after $x_n$.*

Some channels remember the past history of the channel, and the memory may affect the distribution of error for the sent message.

**Definition 11.** *If message $y_n$ is dependent only on a limited number of preceding input signals $x_{n-m}, ..., x_{n-1}, x_n$, then a channel is said to have a finite memory $m$.*

Let $\mathcal{C}^I$ be the set of pairs $(x, y)$, where $x \in \mathcal{A}^I$ and $y \in \mathcal{B}^I$ and $\mathcal{C} = \mathcal{A} \times \mathcal{B}$ is the set of all pairs $(a, b)$ where $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Then $\mathcal{C}^I$ is the class of all sequences of the type $(x, y) = (\cdots, (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), \cdots)$. Let $S = M \times N$ where $M \in \mathcal{A}^I$ and $N \in \mathcal{B}^I$. Suppose $\omega(S)$ is the probability measure of $S \in \mathcal{C}^I$ defined by

$$\omega(S) = \omega(M \times N) = \int_M \nu_x(N) d\mu(x).$$

The connection of a channel $[\mathcal{A}, \nu_x, \mathcal{B}]$ to a source $[\mathcal{A}, \mu]$ creates a stochastic process $[\mathcal{C}^I, F_\mathcal{C}, \omega]$ that can be considered a source itself. It is referred to as a compound source denoted by $[\mathcal{C}, \omega]$. It can be demonstrated that if $[\mathcal{A}, \mu]$ and $[\mathcal{A}, \nu_x, \mathcal{B}]$ are stationary, then so is $[\mathcal{C}, \omega]$. The sources $[\mathcal{A}, \mu]$ and $[\mathcal{C}, \omega]$ correspond to the marginal distribution of the input $x$ and the joint distribution of $(x, y)$, respectively. If we define the probability measure $\eta$ on $F_\mathcal{B}$ as follows

$$\int_{\mathcal{B}^I} k(y) d\eta(y) = \int_{\mathcal{A}^I} d\mu(x) \int_{\mathcal{B}^I} k(y) d\nu_x(y),$$

then the output source $[\mathcal{B}, \eta]$ corresponds to the marginal distribution of $y$. It can be shown that if $[\mathcal{A}, \mu]$ and $[\mathcal{A}, \nu_x, \mathcal{B}]$ are stationary, $[\mathcal{B}, \eta]$ is also stationary. Additionally, if $[\mathcal{A}, \mu]$ is ergodic and if $[\mathcal{A}, \nu_x, \mathcal{B}]$ has a finite memory, then both $[\mathcal{C}, \omega]$ and $[\mathcal{B}, \eta]$ are ergodic[3] [2].

Let $H(X, Y)$, $H(X)$ and $H(Y)$ denote the entropy rates of $[\mathcal{C}, \omega]$, $[\mathcal{A}, \mu]$, and $[\mathcal{B}, \eta]$, respectively. The rate of transmission attained by the source $[\mathcal{A}, \mu]$ over the channel $[\mathcal{A}, \nu_x, \mathcal{B}]$ is given by $R(X, Y) = H(X) + H(Y) - H(X, Y)$. $R$ is dependent on both the channel and information source. However, the supremum of $R(X, Y)$ over all possible ergodic sources, called the *ergodic capacity*, $C$, of the channel, depends only on the channel. As McMillan succinctly states,

> "$R$ represents that portion of the 'randomness' or average uncertainty of each output letter which is not assignable to the randomness created by the channel itself[3]."

Practically speaking, the ergodic capacity of the channel is the maximum number of *bits* of information that can be transmitted per binary digit transmitted over a channel[4].

# 4 Feinstein's Fundamental Lemma

We now state and sketch the proof of Feinstein's Fundamental Lemma in the fashion of Khinchin. This will lay the foundation for the proof of Shannon's coding theorem. We begin by presenting the following three useful inequalities, also attributed to Amiel Feinstein. Their short, ingenious proofs can be found in [2].

## 4.1 Three Inequalities

Consider two random variables $X$, $Y$ and their product $X \times Y$. Let $Z$ be a set of events $X_i \times Y_k \in X \times Y$. Let $U_0$ be some set of events $X_i \in X$ and $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$p(Z) > 1 - \delta_1$$

and

$$p(U_0) > 1 - \delta_2.$$

Denote $\Gamma_i$ for $i = 1, 2, ...n$ the set of events $Y_k \in Y$ such that $X_i \times Y_k$ does not belong to $Z$. Let $U_1$ be the set $X_i \in U_0$ for which $p_{X_i}(\Gamma_i) \leq \alpha$. Then we have

**Feinstein's Inequality #1**

$$p(U_1) > 1 - \delta_2 - \frac{\delta_1}{\alpha}.$$

Let $i_k$ denote the value of the subscript $i$ for which the probability $p(X_i \times Y_k)$ is maximal. If there is more than one maximal value, then any one will suffice. Let

$$P = \sum_{k=1}^{m} \sum_{\substack{i=1 \\ i \neq i_k}}^{n} p(X_i \times Y_k).$$

$P$ is the probability of the occurrence of $X_i \times Y_k$ such that $X_i$ is not the event in $X$ that is most probable for a given event $Y_k \subset Y$.

If for a given $\epsilon$ with $0 < \epsilon < 1$, a set $\Delta_i$ of events $Y_k$ can be associated with each $X_i$ where $i = 1, 2, ...n$ such that

$$p(\Delta_i \times \Delta_j) = 0, i \neq j$$

and

$$p_{X_i}(\Delta_i) > 1 - \epsilon, i = 1, 2, ...n$$

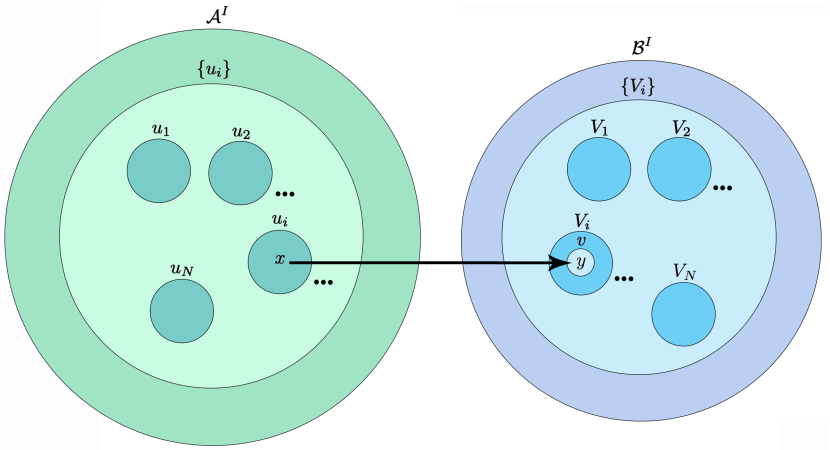
then we have **Feinstein's Inequality #2**

$$P \leq \epsilon.$$

**Feinstein's Inequality #3**
For $n > 1$

$$H(X|Y) \leq P \log(n-1) - P \log P - (1-P) \log(1-P).$$

## 4.2 Distinguishable Sequences




Suppose $m$ is the memory of a non-anticipating ergodic channel $[\mathcal{A}, \nu_x, \mathcal{B}]$. Let $u$ denote the cylinder set of sequences $\{x = (x_{-m}, ..., x_{-1}, x_0, ..., x_{n-1})\}$ with each $x_i \in \mathcal{A}$. Let $v$ denote the set of sequences of the type $\{y = (y_0, ..., y_{n-1})\}$ with all $y_i \in \mathcal{B}$.The fragment $x_{-m}, ..., x_{-1}$ is the part of the input sequence $x$ that the channel "remembers," *i.e.* it affects the probability of error in the output message. The probability $\nu_u(v)$ of receiving an output message $y \in v$ is identical for every $x \in u$, *i.e.* every $x$ with a memory block $x_{-m}, ..., x_{-1}$. Let $V$ denote the union of sequences $v$, and $\nu_u(V)$ denote the probability of output sequence $y \in V$ given the input message $x \in u$. Let $V_i \subset \mathcal{B}^I$ be a set of $v$ sequences, and $\{V_i\}$ be a class of mutually disjoint sets $V_i$ for $1 \leq i \leq N$. The set of sequences $\{u_i\}$ is *distinguishable* if $\nu_{u_i}(V_i) > 1 - \lambda$ where $0 < \lambda < \frac{1}{2}$. The diagram above illustrates the relationships between these sets of sequences.

## 4.3 Feinstein's Fundamental Lemma.

*If a given channel is stationary, without anticipation, and with finite memory, $m$, then, for sufficiently small $\lambda > 0$ and sufficiently large $n$, there exists a distinguishable group $\{u_i\}$ $1 \leq i \leq N$ of $u$-sequences with*

$$N > 2^{n(C-\lambda)}$$

*members, where $C$ is the ergodic capacity of the channel.*

**Proof.** Let $[\mathcal{A}, \mu]$ be an ergodic source and $[\mathcal{A}, \nu_x, \mathcal{B}]$ an ergodic, stationary, non-anticipating channel with finite memory. Since $C$ is the least upper bound of $R(X, Y)$ over all ergodic sources, we have

$$R(X, Y) = H(X) - H(X, Y) > C - \frac{\lambda}{4}.$$

By McMillan's theorem, $[\mathcal{A}, \mu]$ has the $E$ property. Let $w$ be a cylinder set in $\mathcal{A}^I$ and $W_0$ be the set of all cylinders such that

$$\left| \frac{\log \mu(w)}{n} + H(X) \right| \leq \frac{\lambda}{4}.$$

Likewise, the sources $[\mathcal{C}, \omega]$ and $[\mathcal{B}, \eta]$ both have the $E$ Property . Let $v \subset \mathcal{B}^I$, $(w, v) \in \mathcal{C}^I$ and $Z$ denote the set of all cylinders such that

$$\left| \frac{\log \omega(w, v)}{n} + H(X, Y) \right| \leq \frac{\lambda}{4}$$

and

$$\left| \frac{\eta(v)}{n} + H(Y) \right| \leq \frac{\lambda}{4}.$$

The sets $W_0$, $Z$ and $v$ are "high probability" sets. That is, for sufficiently large $n$ and arbitrarily small $\lambda > 0$, $\delta > 0$, $\epsilon > 0$,

$$\mu(W_0) > 1 - \frac{\lambda}{4}$$

$$\omega(Z) > 1 - \delta,$$

and

$$\eta(v) > 1 - \epsilon.$$

Khinchin now finds a set of sequences $W_1 \subset \mathcal{A}^I$ that contains every sequence $w \subset W_0$ such that

$$\frac{\omega(w, A_w)}{\mu(w)} = \frac{\omega(w, A_w)}{\omega(w, \mathcal{B}^I)} > 1 - \frac{\lambda}{2}.$$

Using Feinstein's Inequality #1, Khinichin estimates the probability of $W_1$ to be

$$\mu(W_1) = 1 - 2\lambda.$$

Let $w \subset W_1$ and $v \subset A_w$. If follows that $(w, v) \subset X$. Thus from the inequalities listed above, we have

$$\frac{\log \mu(w)}{n} + H(X) \leq \frac{\lambda}{4},$$

$$\frac{\log \eta(v)}{n} + H(Y) \leq \frac{\lambda}{4},$$

and

$$\frac{\log \omega(w, v)}{n} + H(X, Y) \geq -\frac{\lambda}{4}.$$

Then

$$\log\left(\frac{\omega(w,v)}{\mu(w)\eta(v)}\right)+n[H(X,Y)-H(X)-H(Y)]\geq-\frac{3}{4}n\lambda.$$

Since $R(X,Y)=H(X)+H(Y)-H(X,Y)$, we have

$$\log\frac{\omega(w,v)}{\mu(w)\eta(v)}\geq n[R(X,Y)-\frac{3}{4}\lambda]\iff\frac{\omega(w,v)}{\mu(w)}\geq 2^{n[R(X,Y)-\frac{3}{4}\lambda]}\eta(v).$$

Recalling that $R(X,Y)>C-\dfrac{\lambda}{4}$, it follows that

$$\frac{\omega(w,v)}{\mu(w)}\geq 2^{n[C-\lambda]}\eta(v).$$

Denote $A_w\subset\mathcal{B}^I$ the set of sequences such that $(w,v)\in X$. Then summing over all $v\subset A_w$, we have

$$\frac{\omega(w,A_w)}{\mu(w)}\geq 2^{n[C-\lambda]}\eta(A_w).$$

Since the ratio $\dfrac{\omega(w,A_w)}{\mu(w)}\leq 1$, we can state that

$$\eta(A_w)\leq 2^{-n(C-\lambda)}.$$

Khinchin now constructs what he calls *special groups of w-sequences*. A group $\{w_i\}_{i=1}^N$ is special if a set $B_i$ of $v$-sequences can be associated with each sequence $w_i$ such that

1) $B_i\cap B_j=\varnothing$

2) $\dfrac{\omega(w_i,B_i)}{\mu(w_i)}>1-\lambda$

3) $\eta(B_i)<2^{-n(C-\lambda)}$.

Note how this description of *special groups* parallels the definition of *distinguishable sequences*. It can be demonstrated that every sequence $w\subset W_1$ is a *special group*. A special group is called *maximal* if the addition of any sequence to it results in the group no longer being *special*. It can be further shown that, for sufficiently large $n$, the cardinality, $N$, of the set of *maximal special groups of w- sequences* is given by

$$N>2^{n(C-2\lambda)}.$$

Now let $\{w_i\}_{i=1}^n$ be an arbitrary maximal special group. Take each sequence $w_i$ and concatenate $m$ letters on the left, obtaining a new sequence $u_i$ of length

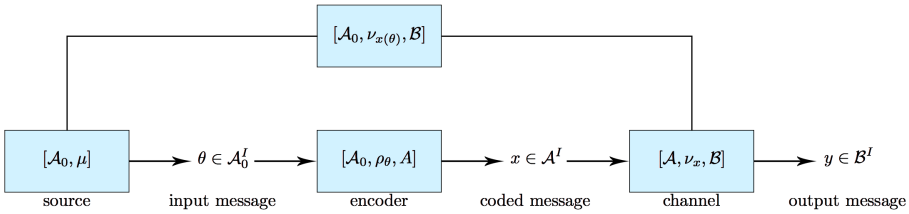
$n+m$, which is an extension of $w_i$. By choosing the appropriate extension it can be shown that for $i = 1, 2, ...N$

$$\frac{\omega(u_i, B_i)}{\mu(u_i)} > 1 - \lambda \iff \nu_{u_i}(B_i) > 1 - \lambda.$$

Therefore, the sequences $\{u_i\}$ form a *distinguishable group* for which $N > 2^{n(C-2\lambda)}$. Since $\lambda$ can be made arbitrarily small, we have $N > 2^{n(C-\lambda)}$, as desired. □

# 5 Shannon's Theorem Part 1

## 5.1 Coding




The source and the channel need not share the same alphabet. Let $[\mathcal{A}_0, \mu]$ be an information source with alphabet $\mathcal{A}_0$ transmitting information through channel $[\mathcal{A}, \nu_x, \mathcal{B}]$. The source transmits a message $\theta = (...\theta_{-1}, \theta_0, \theta_1, ...)$, where $\theta \in \mathcal{A}_0^I$. Each $\theta$ is encoded into an $x \in \mathcal{A}^I$ by a mapping $x(\theta) = x$. The mapping is called a *code*, and it can be regarded as a channel itself. The code channel is denoted $[\mathcal{A}_0, \rho_\theta, \mathcal{A}]$, where for $M \in \mathcal{A}^I$

$$\rho_\theta(M) = \begin{cases} 1 & : x(\theta) \in M \\ 0 & : x(\theta) \notin M. \end{cases}$$

The linking of the code to the channel can be viewed as a new channel $[\mathcal{A}_0, \lambda_\theta, \mathcal{B}]$, where for $Q \in F_\mathcal{B}$,

$$\lambda_\theta(Q) = \nu_{x(\theta)}(Q).$$

Now $[\mathcal{A}_0, \lambda_\theta, \mathcal{B}]$ is equivalent to $[\mathcal{A}_0, \nu_{x(\theta)}, \mathcal{B}]$, and this again creates a compound source $[\mathcal{C}, \omega]$ where $\mathcal{C} = \mathcal{A} \times \mathcal{B}$, and for $M \in F_{\mathcal{A}_0}$ and $N \in F_\mathcal{B}$

$$\omega(M \times N) = \int_M \lambda_\theta(N) d\mu(\theta) = \int_M \nu_{x(\theta)}(N) d\mu(\theta).$$

**Lemma 2.** ***A Useful Inequality***

*Suppose $A$ and $B$ are finite random variables and $p_{A_i}(B_k) = \dfrac{p(A_i, B_k)}{p(A_i)}$, where*

*$A_i$ and $B_k$ are events in $A$ and $B$, respectively.* $\sum_k *$ *denotes the summation over certain values of $k$. Then,*

$$\sum_{i=1}^{n} p(A_i) \sum_k *p_{A_i}(B_k) \log p_{A_i}(B_k) \geq \sum_k *p(B_k) \log p(B_k)$$

The above inequality does not depend on whether or not the events $B_k$ form a complete probability distribution.

**The Noisy Channel Coding Theorem Part 1.**
*Given 1) a stationary, non-anticipating channel $[\mathcal{A}, \nu_x, \mathcal{B}]$ with an ergodic capacity $C$ and finite memory $m$ and 2) an ergodic source $[\mathcal{A}_0, \mu]$ with entropy $H_0 < C$. Let $\epsilon > 0$. Then for sufficiently large $n$, the output of the source $[\mathcal{A}_0, \mu]$ can be encoded into the alphabet $\mathcal{A}$ in such a way that each sequence $\alpha_i$ of $n$ letters from the alphabet $\mathcal{A}_0$ is mapped into a sequence $u_i$ of $n+m$ letters from the alphabet $\mathcal{A}$, and such that if the sequence $u_i$ is transmitted through the given channel, we can determine the transmitted sequence $\alpha_i$ with a probability greater than $1-\epsilon$ from the sequence received at the channel output.*

We will show that a code exists such that $P_e$, the probability of transmission error, can be made arbitrarily small.

**Proof of the Noisy Channel Coding Theorem Part 1.**
Assume $H_0 < C$. Let $HPS = \{\alpha_1, \alpha_2, ....\}$ denote the set of sequences that have a high probability, and $\alpha_0$ denote the set of all sequences in the low probability set, as defined by McMillan. Then there exists a $\lambda$ such that $\lambda < \frac{1}{2}(C - H_0)$. By McMillan's Theorem, $[\mathcal{A}_0, \mu]$ has the $E$ property. Then for an $\alpha \in \mathcal{A}_0$ in the "high probability" group

$$\frac{\log \mu(\alpha)}{n} + H_0 > -\lambda \iff \mu(\alpha) > 2^{-n(H_0+\lambda)}.$$

Since $H_0 < C$, the cardinality of the $HPS$ is less than

$$2^{n(H_0+\lambda)} < 2^{n(C-\lambda)}.$$

By Feinstein's Fundamental Lemma, there exists a distinguishable set of sequences $\{u_i\}$ with $N > 2^{n(C-\lambda)}$ elements. Observe that $N$ is larger than $|HPS|$. So we can construct a mapping that takes each $\alpha_i$ to a unique $u_i$. This implies that there is at least one $u_i$ that is not mapped to some $\alpha_i$. We can map the remaining $u_i$ to the set $\alpha_0$. Since this set $\{u_i\}$ forms a distinguishable set of sequences, there is a group $\{B_i\}_{i=1}^n$ such that $\nu_{u_i}(B_i) > 1-\lambda$ and $B_i \cap B_j = \varnothing$ for $i \neq j$. Now divide the sequence $\theta \in \mathcal{A}_0^I$ into subsequences of length $n$ and divide $x \in \mathcal{A}^I$ into subsequences of length $n+m$. Let the $k^{th}$ subsequence $\alpha$ in

the message $\theta$ correspond to the $k^{th}$ subsequence $u_i$ in $x$. This correspondence is the unique coding map, $x = x(\theta)$. Let $\beta_k$ denote the distinct sequences of length $n$ from $\mathcal{B}^I$, where $k = 1, 2, ..., n$. Then

$$\omega(\alpha_i \times \beta_k) = \int_{\alpha_i} \lambda_\theta(\beta_k) d\mu(\theta) = \int_{\alpha_i} \nu_{x(\theta)}(\beta_k) d\mu(\theta)$$

and

$$\omega(\alpha_i \times \beta_k) = \mu(\alpha_i)\nu_{u_i}(\beta_k).$$

Given a sequence of length $n$ transmitted from the source $[\mathcal{A}_0, \mu]$, $\omega(\alpha_i \times \beta_k)$ is then the joint probability that this sequence corresponds to an $\alpha_i$, and yields a sequence of length $n + m$ with letters from $\mathcal{B}$ after transmission through the channel $[\mathcal{A}_0, \lambda_\theta, \mathcal{B}]$ such that the last $n$ letters comprise the sequence $\beta_k$. Thus we have

$$\sum_{\beta_k \subset B_i} \omega(\alpha_i \times \beta_k) = \mu(\alpha_i)\nu_{u_i}(B_i) > (1 - \lambda)\mu(\alpha_i).$$

Now we invoke Feinstein's Inequality #2. Let $i_k$ denote the value of the index so that $\omega(\alpha_i, \beta_k)$ has its maximum value. Again, if there are more than one maximum, then any one will suffice. Let

$$P = \sum_k \sum_{i \neq i_k} \omega(\alpha_i, B_k).$$

$P$ is the probability that $\alpha_i$ is not the most probable input sequence given the output sequence $\beta_k$. The probability of output $\beta_k$ given input $\alpha_i$ is

$$P_{\alpha_i}(B_i) = \frac{\omega(\alpha_i, B_i)}{\mu(\alpha_i)} > 1 - \lambda.$$

It follows from Feinstein's Inequality #2 that

$$P \leq \lambda \iff 1 - P \geq 1 - \lambda.$$

This means that the probability that $\alpha_i$ is the most probable input sequence, given the receipt of output sequence $\beta_k$, is greater than $1 - \lambda$, and so proves Part 1 of the Noisy Channel Coding Theorem. □

# 6 Shannon's Theorem Part 2

**The Noisy Channel Coding Theorem Part 2.**
*Under the conditions of the Noisy Channel Coding Theorem Part 1, there exists a code $x = x(\theta)$ such that the rate of transmission can be made arbitrarily close to $H_0$.*

**Proof of the Noisy Channel Coding Theorem Part 2.**
Our goal is to show that the rate of transmission can be made arbitrarily close

to the entropy of the source $[\mathcal{A}_0, \mu]$, denoted $H_0$, assuming $H_0 < C$, where $C$ is the ergodic capacity of the channel. To this end, we will find an estimate for the rate. We can regard the output of a source as a stream of symbols from the alphabet $\mathcal{A}_0$ that we can divide into sequences $\alpha_i$ of length $n$, which, when passed through the channel $[\mathcal{A}_0, \lambda_\theta, \mathcal{B}]$, yield sequences of length $n+m$ composed of letters from alphabet $\mathcal{A}$. The last $n$ letters of the sequences from the channel output are the delivered sequences, $\beta_k$.

Now take a sequence of length $s = nt + r$ output from source $[\mathcal{A}_0, \mu]$. Let $X$ denote a sequence of this type, and $\{X\}$ denote the set of all such sequences. Similarly, let $Y$ denote an output sequence of length $s$ from the channel $[\mathcal{A}_0, \lambda_\theta, \mathcal{B}]$, composed of letters from alphabet $\mathcal{B}$, and let $\{Y\}$ denote the set of all sequences $Y$. Let $H(X|Y)$ denote the entropy of space $\{X\}$ conditioned over $Y$.

Each sequence $X$ of length $s = nt + r$ can be subdivided into $t$ consecutive sequences $\alpha^{(1)}, \alpha^{(2)}, ..., \alpha^{(t)}$ and a remainder sequence of length $r < n$, denoted $\alpha^*$. Now

$$\{X\} = \{\alpha^{(j)}\} \times \{\alpha^*\}.$$

So we have

$$H(X|Y_0) \leq \sum_{j=1}^{t} H(\{\alpha^{(j)}\}|Y_0) + H(\alpha^*|Y_0).$$

Averaging over $Y$, it follows that

$$H(X|Y) \leq \sum_{j=1}^{t} H(\{\alpha^{(j)}\}|Y) + H(\alpha^*|Y).$$

As with $X$, we can also subdivide each sequence $Y$ into $t$ sequences $\{\beta^{(j)}\}_{j=1}^{t}$ of length $n$ with a remainder sequence $\{\beta^*\}$ of length $r < n$ so that $\{Y\} = \{\beta^{(j)}\} \times \{\beta^*\}$. Note that the spaces $\{\alpha^{(j)}, \beta^{(j)}\}$ and $\{\beta_k^{(j)}\}$ have distributions $\omega(\alpha_i, \beta_k)$ and $\eta(\beta_k)$, respectively. Let $B^{(j)}$ denote the set of sequences $\{\beta^{(l)}\}_{l=1}^{t}$ and $\beta^*$ that comprise $Y$ except $\beta^{(j)}$. In other words, for $j = 1, 2, ..., t$, $\{Y\} = \{\beta^{(j)}\} \times B^{(j)}$.
By Lemma #2 it follows that

$$H(\{\alpha^{(j)}\}|Y) = H(\{\alpha^{(j)}\}|\beta^{(j)}B^{(j)}) \leq H(\{\alpha^{(j)}\}|\beta^{(j)}) = H(\alpha|\beta).$$

Since $a$ is the number of letters in alphabet $\mathcal{A}$, there are $a^r$ sequences in $\{\alpha^*\}$. It follows from Theorem # 4 that

$$H(\alpha^*|Y) < \log a^n \iff H(\alpha^*|Y) < n \log a$$

and for sufficiently large $n$ and $\lambda > 0$,

$$H(X|Y) = tH(\alpha|\beta) + n \log a \iff H(X|Y) < \lambda tn + n \log a \leq \lambda s + n \log a.$$

The quantity of information encoded in $X$ before transmission is $sH_0$, and since $H(X|Y)$ is the amount of information lost after transmission, we have that the amount of information retained is $sH_0 - H(X|Y)$. Also, when $X$ is transmitted through the channel, the total number of symbols transmitted is $(t+1)(n+m)$. Thus the rate in *bits per symbol* is given by

$$\frac{sH_0 - H(X|Y)}{(t+1)(n+m)}.$$

Recall that $s = nt + r$. So $s + n = nt + r + n$, and consequently, $s + n > nt + n$. Since $H(X|Y) \leq \lambda s + n \log a$, we have

$$\frac{sH_0 - H(X|Y)}{(t+1)(n+m)} \geq \frac{sH_0 - \lambda s - n\log a}{(nt+n)(1+\frac{m}{n})} \geq \frac{sH_0 - \lambda s - n\log a}{(s+n)(1+\frac{m}{n})} = \frac{H_0 - \lambda - \frac{n\log a}{s}}{(1+\frac{n}{s})(1+\frac{m}{n})}.$$

So by choosing $n$ and $t$ sufficiently large such that $\dfrac{m}{n} < \epsilon$ and $\dfrac{n}{s} \leq \dfrac{1}{t} < \epsilon$, for some $\epsilon > 0$, we have

$$\frac{H_0 - \lambda - \epsilon \log a}{(1+\epsilon)^2} < H_0 - 2\lambda$$

as $\epsilon \to 0$.

The last inequality implies the existence a code for which rate of transmission can be made arbitrarily close to $H_0 < C$. □

# 7 Conclusion

We cannot generally reconstruct a message sent over a noisy channel. However, if we restrict ourselves to sending *distinguishable sequences* such that, with a high probability, each is mapped into a particular set at the channel output within a class of mutually disjoint sets, then we can determine the original message virtually without error, so long as the number of different sequences from the output of the given source does not exceed the number of *distinguishable sequences* at the channel input. The proof of the Noisy Channel Coding Theorem hinges on this powerful idea, which is a synthesis of the results of Shannon, Feinstein and McMillan. McMillan showed that the number of high probability sequences approximates $2^{nH_0}$, and Feinstein proved the existence of $N > 2^{n(C-\epsilon)}$ distinguishable groups. If $H_0 < C$, the claims of the Noisy Channel Coding Theorem follow.

Although Shannon's theorem is an existence theorem, its implications have motivated scientists to persist in their search for efficient coding systems. This is evident in the modern-day forward error correction coding systems, some of which closely approximate Shannon's theoretical maximum transmission rates.

Additionally, Shannon's ideas can be found in research fields as varied as linguistics, genetics, neuroscience, computer science and digital communications engineering. The advancement of information theory has provoked a radical shift in perspective within the scientific community. Indeed, Shannon arguably "single-handedly accelerated the rate of scientific progress"[6].

# References


[1] Robert B. Ash. *Information Theory*. Dover Publications, New York, 1990. Unabridged and corrected republication of the work originally published by Interscience Publishers ... New York, 1965–T.p. verso.

[2] A.I. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, Inc., New York, New York, 1957.

[3] Brockway McMillan. The basic theorems of information theory. *The Annals of Mathematical Statistics*, 24(2):196–219, 1953.

[4] John Robinson Pierce. *An Introduction to Information Theory : Symbols, Signals & Noise*. Dover Publications, New York, second, revised edition.

[5] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal, The*, 27(4):623–656, Oct 1948.

[6] James V. Stone. *Information Theory: A Tutorial Introduction*. James V Stone, 2013.

[7] Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.