Information Theory

Gurkirat,Harsha,Parth,Puneet,Rahul,Tushant

November 8, 2015

1 Introduction[1]

Information theory is a branch of applied mathematics, electrical engineering, and computer science involving the quantification of information .
Firstly, it provides a general way to deal with any information , be it language, music, binary codes or pictures by introducing a measure of information , *entropy* as the average number of bits(binary for our convenience) needed to store one symbol in a message.
This abstraction of information through entropy allows us to deal with all kinds of information mathematically and apply the same principles without worrying about it's type or form! It is remarkable to note that the concept of entropy was introduced by Ludwig Boltzmann in 1896 as a measure of disorder in a thermodynamic system , but serendipitously found it's way to become a basic concept in Information theory introduced by Shannon in 1948 .
Information theory was developed as a way to find fundamental limits on the transfer , storage and compression of data and has found it's way into many applications including Data Compression and Channel Coding.

2 Basic definitions

2.1 Intuition behind "surprise" or "information"[2]

Suppose someone tells you that the sun will come up tomorrow.This probably isn't surprising. You already knew the sun will come up tomorrow, so they didn't't really give you much information. Now suppose someone tells you that aliens have just landed on earth and have picked you to be the spokesperson for the human race. This is probably pretty surprising, and they've given you a lot of information.*The rarer something is, the more you've learned if you discover that it happened.*

2.2 Mathematical formulation of information[3]

Assume that an event E occurs with the probability p . Now , knowing that E has occurred gives us

$$I(p) = \log_2(1/p) = -\log_2 p \tag{1}$$

bits of information. We use log base 2 as this is the number of binary bits and it's convenient to work in a binary framework. We will justify our definition of information by 3 *common sense* observations about information.

1. $I(p) \geq 0$ for all $p \in (0, 1]$
   If we've learned that an event has occurred then certainly we have learned *something.*The knowledge of an event happening can't make us "lose" or "forget" information.

2. $I(p)$ is a continuous function of p.
   If we vary p slightly , then the information gained should also vary slightly. An infinitesimally small change in p would not cause a sudden jump or drop in the information gained.

3. $I(pq) = I(p) + I(q)$ for all $p, q \in (0, 1]$.
   Suppose that E and F are independent events with $Pr(E) = p, Pr(F) = q$ and $Pr(EF) = pq$. If we already know that E has occurred and we are told that F occurs, then the new information obtained is $I(q) = I(pq) - I(p)$.

Now, using these observations let us construct our original definition of information.
Let $p \in (0, 1]$.
Using property (3) , $I(p^2) = I(p) + I(p) = 2I(p)$.
Similarly, $I(p^m) = mI(p)$ for all positive integers m.
Also, $I(p) = I(p^{1/n} \dots p^{1/n}) = nI(p^{1/n})$, and hence $I(p^{1/n}) = (1/n)I(p)$ for all integers n.
These observations imply that $I(p^{m/n}) = (m/n)I(p)$. By (2), we know that the function is continuous and hence $I(p^x) = xI(p)$ for all positive real numbers x. Therefore $I(p) = I((1/2)^{-\log_2 p}) = -I(1/2)\log_2 p = -C\log_2 p$, where $C = I(1/2)$. By (1), C must be positive. For convenience, we take $C = 1$, so , $I(p) = -\log_2 p$

Shannon's seminal idea in information theory is to associate an amount of information, or entropy, with an information source.

2.3 Source

A Source **S** is a sequence of random variables $X_1, X_2, ...$ with a common range $x_1, x_2, ..., x_n$. Such a sequence is also called *discrete-time, discrete-valued stochastic sequence* , where the elements x_i are called states or symbols.
The source **S** is seen as emitting the symbols $x_1, x_2, ..., x_n$ at regular intervals of time.

If in a source the random variables X_i are independent and identically distributed, then source **S** is called *discrete-memoryless source* or *zero-memory source.*If S is a discrete-memoryless source, it is written as

$$\begin{pmatrix} x_1 & x_2 & ... & x_n \\ p_1 & p_2 & ... & p_n \end{pmatrix}$$

where each p_i is the probability of any random variable X to take value x_i.

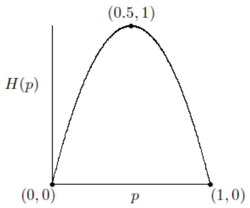


Figure 1: The Entropy Function H(p)

2.4 Entropy of a Source S

Conside a discrete-memoryless source S =

$$\begin{pmatrix} x_1 & ... & x_n \\ p_1 & ... & p_n \end{pmatrix}$$

If S emits a symbol x_i, then we say that we have received -$\log p_i$ bits of information. The average amount of information obtained per symbol is equal to **H(S)**.

$$\mathbf{H(S)} = \text{-}\sum_{i=1}^{n} p_i \log p_i bits \tag{2}$$

*$\log x$ means $\log_2 x$ in this paper.
*$\lim_{x\to 0} x \log x = 0$

2.5 Convexity of Logarithm Function

Let $p_1, p_2, ..., p_n$ and $q_1, q_2, ..., q_n$ be non-negative real numbers such that $\sum p_i$ = $\sum q_i$ = 1. Then

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \tag{3}$$

with equality *iff* $p_i = q_i$ for all i.

Proof : There is no contribution to the summations from any $p_i = 0$, hence we discard these.

$$\begin{aligned}\sum_{i=1}^{n} p_i \log q_i - \sum_{i=1}^{n} p_i \log p_i &= \sum_{i=1}^{n} p_i(\log q_i - \log p_i) \\ &= \sum_{i=1}^{n} p_i \log(\frac{q_i}{p_i}) \\ &\leq \sum_{i=1}^{n} p_i(\frac{q_i}{p_i} - 1) \\ &(using \log x \leq x - 1) \\ &= \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \\ &= 1 - 1 \\ &= 0\end{aligned}$$

2.6 Two Extreme Type of Sources

2.6.1 Uniform Source

A source S is called a Uniform Source if the probability of each state or symbol is equal, i.e.

$$\begin{pmatrix} x_1 & x_2 & ... & x_n \\ 1/n & 1/n & ... & 1/n \end{pmatrix}$$

The entropy of a uniform source = - $\sum_{i=1}^{n} (1/n) \log 1/n = \log n$.

2.6.2 Singular Source

A source S is called a Singular Source if the probability of one of the states or symbols is equal 1 and rest all are 0, i.e.

$$\begin{pmatrix} x_1 & ... & x_k & ... & x_n \\ 0 & ... & 1 & ... & 0 \end{pmatrix}$$

The entropy of a singular source = 0. Thus singular source provides no information.

Theorem 2.1 : The entropy of a source with n states satisfies the following inequality -

$$0 \leq \mathbf{H(S)} \leq \log n \tag{4}$$

proof:

$$\log x \leq 0 \textit{ iff } x \in (0, 1).$$

$$\begin{aligned}\log p_i &\leq 0 \\ p_i \log p_i &\leq 0 \\ -p_i \log p_i &\geq 0 \\ -\sum_{i=1}^{n} p_i \log p_i &\geq 0.\end{aligned}$$

This proves the lower bound for H(S) . For upper bound for H(S):

$$\begin{aligned}
-\sum_{i=1}^{n} p_i \log p_i &\leq -\sum_{i=1}^{n} p_i \log(1/n) \\
&= -\log(1/n)\sum_{i=1}^{n} p_i \\
&= \log n
\end{aligned}$$

Hence, the upper bound is $\log n$.

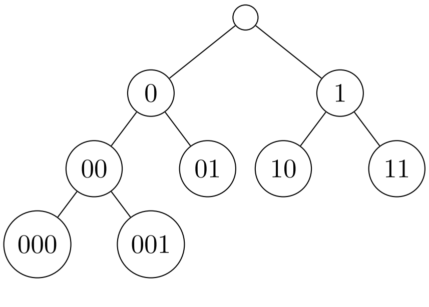
Code : A code C for a source with n states contains a sequence w1, . . . , wn of binary strings, none the prefix of another.The source is

$$\begin{pmatrix} x_1 & x_2 & ... & x_n \\ p_1 & p_2 & ... & p_n \\ w_1 & w_2 & ... & w_n \end{pmatrix}$$

3 Kraft's Inequality

Statement : A source with n states has a code with word lengths $l_1, l_2, ...l_n$ if and only if $\sum\limits_{i=1}^{n} 2^{-l_i} \leq 1$

Proof : Let us arrange the n lengths according to their size in ascending order.Without loss of generality ,we can assume that the lengths are in ascending order.Let l be the maximum among all l_i's. Consider a binary tree with 0's and 1's upto length 'l', such that descendants of a node have it as their prefixes .



Consider a string of length l_i.Since it is not a prefix of any other word, it prevents 2^{l-l_i} strings of length l from being code words, this happens with every l_i.Moreover, no child of w_j is a child of w_i .i.e., overlapping is not possible for any two trees , where w_i refers to the node of the code word with length l_i. For each i,j, 2^{l-l_i}strings(of i) and 2^{l-l_j}strings(of j) never overlap. Summing over all code words, $\sum\limits_{i=1}^{n} 2^{l-l_i} \leq 2^l$, i.e., $\sum\limits_{i=1}^{n} 2^{-l_i} \leq 1$

Now for the converse ,we have to prove that,

If $\sum\limits_{i=1}^{n} 2^{-l_i} \leq 1$*, then there is a prefix code with the lengths* l_i*,where* $1 \leq i \leq n$.

Assume that the l_i's are in ascending order . Take length l_1 and mark any one of the binary strings of length l_1 of binary tree . As we have to form a prefix code ,we cannot use any string which is having the marked node as its root(no overlapping trees) .
Now mark l_2 and mark any one of the binary strings of length l_2 of binary tree which is not a part of the tree formed by 1^{st} marked node and continue the same for $l_3, l_4, ..., l_n$. As $\sum\limits_{i=1}^{n} 2^{l-l_i} \leq 2^l (since \sum\limits_{i=1}^{n} 2^{-l_i} \leq 1)$, we will always find a tree for each l_i such that no two trees overlap and the root string will be the code word . Here , no node marked is a prefix of the other.

Thus, a prefix code can be formed if a set of lengths $l_1, l_2, ...l_n$ are given such that $\sum\limits_{i=1}^{n} 2^{-l_i} \leq 1$ holds.

This proves the Kraft's inequality.

Average length of code :
The average length of the code C is $\bar{l} = \sum\limits_{i=1}^{n} p_i.l_i$. This quantity is the average number of code symbols per source symbol.

4 Shannon's First Theorem :

The average length $\bar{l}$ of a code is at least equal to the entropy $H(S)$ of the source. i.e.

$$\bar{l} \geq H(S) \tag{5}$$

Proof:

$$\begin{aligned}
\bar{l} &= \textstyle\sum_{i=1}^{n} P_i l_i \\
&\geq \textstyle\sum_{i=1}^{n} P_i l_i + \log\left(\sum 2^{-l_i}\right) \qquad \text{[by kraft's ineqality]} \\
&= \textstyle\sum_{i=1}^{n} P_i l_i + \sum p_i \log\left(\sum 2^{-l_i}\right) \\
&= \text{-} \textstyle\sum p_i \log\left(\frac{2^{-l_i}}{\sum 2^{-l_i}}\right) \\
&\geq \text{-} \textstyle\sum p_i \log p_i \qquad \text{[by convexity of logrithm function]} \\
&= H(S) \qquad [since \sum\limits_{i=1}^{n} \tfrac{2^{-l_i}}{\sum 2^{-l_i}} = 1]
\end{aligned}$$

4.1 Theorem 2

Given a source S, there always exists a code for S with average length $\bar{l}$ such that

$$\bar{l} < H(S) + 1 \tag{6}$$

proof:

Let source S be

$$\begin{pmatrix} x_1 & x_2 & ... & x_n \\ p_1 & p_2 & ... & p_n \end{pmatrix}$$

Define l_i to be integer such that $-\log p_i \leq l_i \leq -\log p_i + 1$. Then

$$\sum_{i=1}^{n} 2^{-l_i} \leq \sum_{i=1}^{n} 2^{\log p_i} = \sum_{i=1}^{n} p_i = 1.$$

Thus by Kraft's inequality we can encode S by with strings of length $l_1, l_2, ..., l_n$. We have selected $l_i \leq -\log p_i + 1$

$$\begin{aligned} l_i &\leq -\log p_i + 1 \\ l_i p_i &\leq (-\log p_i + 1)p_i \\ \sum_{i=1}^{n} l_i p_i &\leq \sum_{i=1}^{n} (-\log p_i + 1)p_i \\ \bar{l} &\leq -\sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} p_i \\ &= H(S) + 1 \end{aligned}$$

Prepostion Given any two sources X and Y, we have

$$H(XY) = H(X) + H(Y). \tag{7}$$

Proof

$$\begin{aligned} H(XY) &= -\sum_{i=1}^{m} \sum_{i=1}^{n} p_i q_j \log p_i q_j \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p_i q_j \log p_i - \sum_{i=1}^{m} \sum_{j=1}^{n} p_i q_j \log q_j \\ &= -\sum_{i=1}^{m} p_i \log p_i \sum_{j=1}^{n} q_j - \sum_{i=1}^{m} p_i \sum_{j=1}^{n} q_j \log q_j \\ &= H(X) + H(Y) \end{aligned}$$

Channels

Channel (X, Y) consists of an input alphabet $X = \{x_1, x_2, ..., x_m\}$, an output alphabet $Y = \{y_1, y_2, ..., y_n\}$, and conditional probabilities $p_{ij} = p(y_j \mid x_i)$, for $1 \leq i \leq m$, $1 \leq j \leq n$. The conditional probability $p(y_j \mid x_i)$ is the probability that the output symbol y_j is received when the input symbol x_i is sent.

We can represent the channel by using individual probabilities for input and output alphabets or by using conditional probability matrix :

$$\begin{pmatrix} p_{11} & \cdots & p_{1n} \\ . & . & . \\ . & . & . \\ . & . & . \\ p_{m1} & \cdots & p_{mn} \end{pmatrix}$$

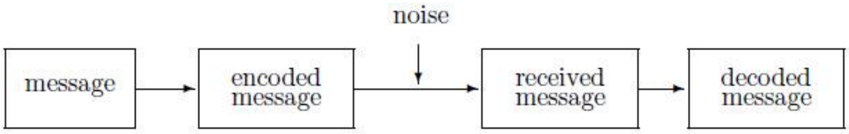


Figure 2: A Communication Model

.

We have actually defined a discrete memoryless channel(DMC) above as any instance of appearance of a symbol does not affects those that come later.

Given a channel (X, Y) following entropies are defined:

1.Input Entropy

$$H(X) = -\sum_{i=1}^{m} p_i \log p_i \tag{8}$$

2.Output Entropy

$$H(Y) = -\sum_{i=1}^{n} q_i \log q_i \tag{9}$$

3.Conditional Entropy or Equivocation

$$H(X \mid Y) = -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i \mid y_j) \tag{10}$$

$$H(Y \mid X) = -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(y_j \mid x_i) \tag{11}$$

4. Total Entropy

$$H(X, Y) = -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \tag{12}$$

5.Mutual Information

$$I(X, Y) = H(X) - H(X \mid Y) \tag{13}$$

The following figure indicates the relationships among the various entropies. By analogy with a *a priory* entropy and *a posteriori* probability, $H(X)$ and

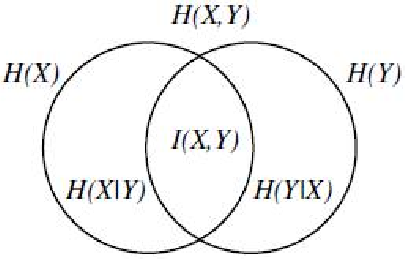


Figure 3: The entropies associated with a channel.

$H(X \mid Y)$ are called a priory entropy and a posteriori entropy, respectively.

Let's work out the above definitions. H(X) and H(Y) are defined so by definition of entropy.$H(X \mid Y)$ can be thought of as the uncertainty in X on choosing some Y and then taking sum over all possible values of Y.

$$H(X \mid Y = y_j) = -\textstyle\sum_{i=1}^{m} p(x_i \mid y_j) \log p(x_i \mid y_j)$$

Taking sum over all possible values taken by Y, we get:

$$\begin{aligned}
H(X \mid Y) &= \sum_{i=1}^{n} H(X \mid Y = y_j) q_j \\
&= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i \mid y_j) q_j \log p(x_i \mid y_j) \\
&= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) q_j \log p(x_i \mid y_j)
\end{aligned}$$

Theorem Given any channel (X, Y) , we have

$$H(X,Y) = H(Y) + H(X \mid Y) \tag{14}$$

Proof.

$$\begin{aligned}
H(X,Y) &= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j)[\log q_j + \log p(x_i \mid y_j)] \\
&= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j)[\log q_j] - \sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i \mid y_j)] \\
&= -\sum_{i=1}^{n} q_i \log q_i + H(X \mid Y) \\
&= H(Y) + H(X \mid Y)
\end{aligned}$$

Similarly we have the proof for

$$H(X, Y) = H(X) + H(Y \mid X). \tag{15}$$

Theorem Given any channel H(X, Y), we have

$$H(X, Y) \leq H(X) + H(Y)$$

Equality holds if and only if X and Y are independent.

Proof.

$$\begin{aligned} H(X, Y) &= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\ &\leq -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\ &= -\sum_{i=1}^{m} \sum_{i=1}^{n} p(x_i, y_j) \log q_j - \sum_{i=1}^{m} \sum_{i=1}^{n} \log q_j \\ &= H(X) + H(Y) \end{aligned}$$

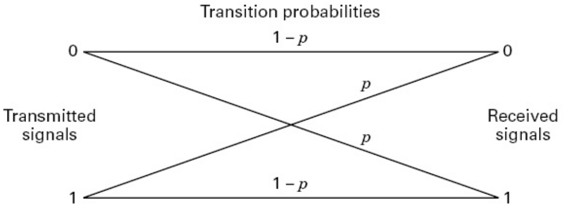
Corollary Given any channel (X, Y) , the following inequalities hold:

$$0 \leq H(X \mid Y) \leq H(X);$$

$$0 \leq H(Y \mid X) \leq H(Y);$$

$$0 \leq I(X, Y);$$

Binary Symmetric Channel (BSC) :- BSC is a channel in which both input and output (X and Y) are of the form $\{0, 1\}^1$ and accuracy of the channel is independent of the input bit.
Let p be the probability that the bit is sent inaccurately over the channel.
Then the channel can be visualised as, [1]



Lemma: Given a BSC between two sources X and Y, with probability of error p

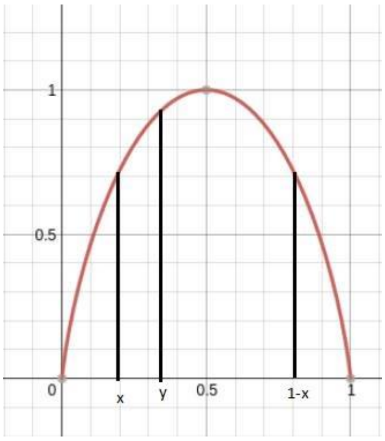
$$H(Y) \geq H(X)$$

Proof: Let x_0 and x_1 be the probabilities with which the input from X is 0 and 1 respectively and y_0 and y_1 be the probabilities with which the output from Y is 0 and 1 respectively.

[1] Image Source : http://www.ni.com/white-paper/14917/en/

Without loss of generality, assume that $x_0 \leq \frac{1}{2}$.

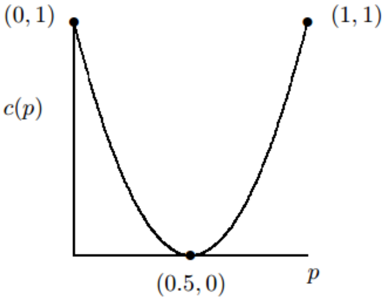
$$\begin{aligned} y_0 =& x_0 * (1-p) + x_1 * p \\ y_1 =& x_0 * (p) + x_1 * (1-p) \\ \Rightarrow & x_0 \leq y_0 \leq x_1 (x_0 \leq x_1) \\ \Rightarrow & x_0 \leq y_0 \leq 1 - x_1 \\ H(X) =& -(x_0 * log(x_0) + x_1 * log(x_1)) \\ H(X) =& -(x_0 * log(x_0) + (1-x_0) * log((1-x_0)) \\ H(Y) =& -(y_0 * log(y_0) + y_1 * log(y_1)) \\ H(Y) =& -(y_0 * log(y_0) + (1-y_0) * log((1-y_0)) \end{aligned}$$



$H(Y) \geq H(X)$

Capacity - The capacity c of a channel is defined as the maximum mutual information of the channel

$$c = \max_{p_i} I(X.Y)$$



Theorem - The capacity c(p) is given by

$$c(p) = 1 + plog(p) + qlog(q) = 1 - H(p) \tag{16}$$

Proof - From the above graph we can see that

$$H(Y) \leq 1.$$

$$\begin{aligned} I(X,Y) = H(Y) - H(Y|X) \leq 1 - H(Y|X) = 1 - H(p). \\ Therefore, c(p) = 1 - H(p) = 1 + plog(p) + qlog(q). \end{aligned}$$

The capacity gives a limit on correctly good the channel can transmit information.
If the probabilty of error p is low, H(p) is low and capacity is high and anf if it is close to 0.5 the capacity is around 0.

Rate r(C) - The rate of a code C(with all code words of length n) is the ratio of the size of C to the length n.

$$r(C) = \frac{log_2(|C|)}{n}$$

It follows from the definition that r(C) $\leq$ 1.
The rate r captures essentially the density of the coding. A highly dense coding is more susceptible to errors and this idea is formalised by the Shannon's Second Theorem.

5 Shannon's Second Theorem

Consider a BSC(Binary Symmetric Channel) with probabilty of error $p < \frac{1}{2}$ and capacity $c = 1 - H(p)$.

Let $R < c$ and $\varepsilon > 0$. For sufficiently large n,there exists a subset of $M \geq 2^{Rn}$ code words from the set of 2^n possible inputs such that probability of error(per word) is less than ε.
The code guaranteed by Shannon's second theorem has rate

$$\frac{\log M}{n} \geq \frac{\log 2^{Rn}}{n} = R \tag{17}$$

Proof of theorem: Thus it is possible, by choosing n sufficiently large, to reduce the maximum probability of error to an amount as low as desired while at the same time maintaining the transmission rate near the channel capacity. We establish some technical preliminaries. Choose R_0 with $R < R_0 < c$. Let $\delta = \epsilon/2$. Choose Δ so that $R_0 < 1 - H(p\delta) = c(p\delta) < c$,
The Hamming distance between two code words, named after Richard Hamming (1915–1998), is the number of coordinates in which the words differ. The Hamming distance of a code is the minimum Hamming distance between code words. Assume that the channel is a BSC^n with probability of error p, with n to be determined. Suppose that α is transmitted and β is received. The expected Hamming distance between α and β is np. Consider a sphere T of radius $np\Delta$ about β. Our decision procedure is as follows: if there is a unique word in T, then we accept it. If there is no code word in T, or more than one code word, then we concede an error.

Lemma If n is a positive integer and $0 < x < 1/2$, then

$$\sum_{i=1}^{[nx]} \binom{n}{k} < 2^{nH(x)} \tag{18}$$

The proof uses the probabilistic method, in which the existence of a desired object (a good code) is established by showing that it exists with positive probability. The probability of error is

$$\begin{aligned} Pr(error) &= Pr(\alpha \notin T) + Pr(\alpha \in T)Pr(\alpha\prime \in T : \alpha\prime \neq \alpha) \\ &\leq Pr(\alpha \notin T) + Pr(\alpha\prime \in T : \alpha\prime \neq \alpha) \\ &\leq Pr(\alpha \notin T) + \sum_{\alpha\prime \neq \alpha} Pr(\alpha\prime \in T) \end{aligned}$$

It follows from the law of large numbers that, given Δ and δ, there exists n_o such that

$$Pr(\frac{|X - np|}{n} > \Delta) < \delta$$

for n $\geq n_o$. Hence, the probability that the number of errors, X, exceeds the expected number of errors, np, by more than $n\Delta$ is less than δ. Therefore, we may make the first term arbitrarily small (less https://preview.overleaf.com/public/nmtdnvpprfsw/images/1856a7f
δ).
Choose M with $2_{nR} \leq$ M $\leq 2_{nR\prime}$. Suppose that M words are selected randomly from the 2^n possible words. There are 2^{nM} possible codes, each selected with probability 2^{-nM}. Thus

$$\begin{aligned} \bar{P}r(error) &< delta + (M-1)\bar{P}r(\alpha\prime \in T) \\ &\leq \delta + M\bar{P}r(\alpha\prime \in T), \end{aligned}$$

where $\bar{P}r$ denotes an average probability over all 2^{nM} codes. Now $Pr(\alpha\prime \in T) = \frac{|T|}{2^n}, where\, |T| = \sum_{k=0}^{[np_\Delta]} \binom{n}{k}$. Therefore, we have

$$\begin{aligned} \bar{P}r(error) &< \delta + M2^{-n}2^{nH(p\Delta)} \\ &\leq \delta + 2^{nR\prime - n + nH(p\Delta)} \\ &= \delta + 2^{n(R\prime - 1 + H(p_\Delta))} \\ &= \delta + 2^{n(R\prime - c(p_\Delta))}. \end{aligned}$$

References

[1] Information theory
https://en.wikipedia.org/wiki/Information_theory

[2] Taken from Ryan Moulton's *A Short, Simple Introduction to Information Theory*
https://moultano.wordpress.com/2010/10/23/a-short-simple-introduction-to-3kbzhsxyg4467-7/

[3] Martin Erikson. *Pearls of Discrete Mathematics.*