Shannon's Noisy-Channel Theorem

Amon Elders

February 6, 2016

Information and Communication

Begeleiding: Christian Schaffner

Korteweg-de Vries Instituut voor Wiskunde
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

Abstract

Realistic channels generally a certain amount of noise associated with them, that is, information is not losslessly transmitted. In this article we formalize the notion of communication over a noisy-channel and prove the result of Shannon's Noisy-Channel theorem. Shannon's Noisy-Channel Theorem states that for codes with less than 2^{nR} codewords, where R is the rate, it is possible to communicate over a noisy-channel with arbitrarily small error when the rate of information transmitted is smaller than capacity. Where capacity is the maximum amount of information that can be send such that the error is arbitrarily small, a formal notion justified in its intuitive sense of capacity by Shannon's noisy-channel theorem. We will also provide some thoughts on the practical applications of this theorem.

Titel: Shannon's Noisy-Channel Theorem
Auteur: Amon Elders, C.Schaffner@uva.nl, 6127901
Begeleiding: Christian Schaffner
Einddatum: February 6, 2016

Korteweg-de Vries Instituut voor Wiskunde
Universiteit van Amsterdam
Science Park 904, 1098 XH Amsterdam
`http://www.science.uva.nl/math`

Contents

1 Introduction

In practice, all channels have noise, in principle this means given a fixed input, the output of our channel will be uncertain. That is, for every input there is a certain probability, dependent on our input, that our output will differ. For example, we might send 0 over a certain channel and with probability 90% the receiver gets a 0 and with probability 10% the receiver gets a 1. Therefore this channel is *noisy*. It is clear that when we use a noisy-channel we are bound to make an error. But how do we make this mistake as small as possible? Naively we can send a 0 a hundred times over the channel and the receiver declares that either a 0 or a 1 was sent depending on which he counts the most. This would, however, not be very practical. Suppose we have a channel where the receiver gets a 0 or 1 both with probability 50%, then it is quite clear that no information can be sent. From this it seems reasonable that the amount of transmitted information depends on the mutual information between the input and the received signal, which has to depend on the particular noisy channel. The brilliance of Shannon is that he formalizes the above notions and shows that with the right scheme we can send information with arbitrarily small error while retaining a high rate of information transmitted. Where rate measures the transmission speed. We have to build our formal framework first before we can prove this result and we begin by formalizing the notion of a *noisy-channel*.

2 Discrete Memoryless Channels

Definition 2.1. *A discrete memoryless channel* $(\mathcal{X}, p(Y|X), \mathcal{Y})$ *is characterized by an input alphabet* $\mathcal{X}$ *and an output alphabet* $\mathcal{Y}$ *and a set of conditional probability distributions* $P(y|x)$*, one for each* $x \in \mathcal{X}$*. These transition probabilities may be written in matrix form:*

- $Q_{ji} := P(y = b_j|x = a_i)$

Where memoryless means that the probability distribution of the output depends only on the input at that time and is independent of previous inputs.

Example 2.1. *Let* $p \in (0, 1)$ *the* ***binary symmetric channel*** *has the following probability matrix:*

$$Q = \begin{array}{c} y|x \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array}$$

Thus,

- $P(y = 0|x = 0) = 1 - p$
- $P(y = 1|x = 0) = p$
- $P(y = 0|x = 1) = p$
- $P(y = 1|x = 1) = 1 - p.$

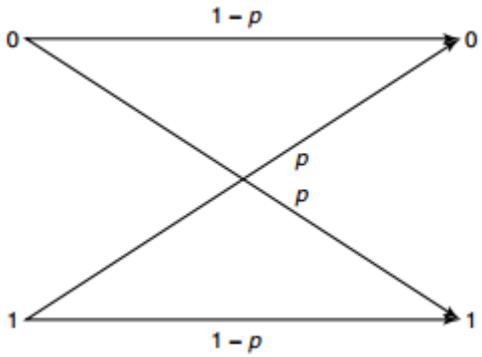

Figure 2.1: Binary symmetric channel; Elements of information theory (2006)

As noted in the introduction, our intuition is that the amount of information that can be sent through a channel depends on the mutual information. Shannon's noisy-channel theorem will give a justification for the following definition.

Definition 2.2. *We define the capacity of a discrete memoryless channel as*

$$C = \max_{p(x)} I(X;Y).$$

Where p(x) is the input distribution. As we need a way to encode and decode our input and output we have the following definition of a *block code*.

Definition 2.3. *An (M,n) code for the channel $(\mathcal{X}, P(Y|X), \mathcal{Y})$ consists of the following:*

- *An index set $\{1, 2, \ldots, M\}$.*
- *An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$, the set of codewords is called the codebook.*
- *A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$. Deterministic rule which assigns a guess to each $y \in \mathcal{Y}^n$.*

Where M can be seen as the number of messages that can be send and n is the number of times we use the channel. This gives us the following rate,

Definition 2.4. *The rate of an (M,n) code is*

$$R = \frac{\log(M)}{n} \text{ bits per transmission.}$$

We also need formal notions of error.

Definition 2.5. *Given that i was sent, the probability of error is*

$$\lambda_i = P(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} P(y^n|x^n(i)) I(g(y^n) \neq i)$$

Where $I(\cdot)$ is the indicator function.

Definition 2.6. *The maximal probability of error $\lambda^{(n)}$ for an (M, n) code is defined as*

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,m\}} \lambda_i.$$

Definition 2.7. *We define the average probability of error as*

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.$$

We can now state Shannon's noisy-channel theorem, before giving Shannon's noisy-channel theorem a formal treatment we will give some intuition as to what the main idea is behind the proof.

Theorem 2.1. *For a discrete memory-less channel, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$*

3 Noisy-Typewriter

Example 3.1. *Consider the following noisy-channel, we have the alphabet for $\mathcal{X}$ and $\mathcal{Y}$ and the input is either unchanged with probability $\frac{1}{2}$ or with probability $\frac{1}{2}$ the output received is the next letter in the alphabet.*

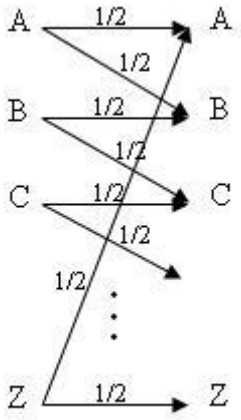

Figure 3.1: Noisy Typewriter; http://www.umsl.edu/

Note that this channel has capacity $\log(13)$ as $C = \max I(X;Y) = \max(H(Y) - H(Y|X)) = (H(Y) - 1) = \log(26) - 1 = \log(13)$. Where the maximum is attained by the uniform distribution. Now take the following scheme:

1. We take the index set to be: $\{1, 2, \ldots, 13\}$
2. The following encoding function $X(1) = a, X(2) = c, \ldots, x(13) = y$
3. The decoding function maps the received letter to the nearest letter in the code. For example; we map a and b to a and c and d to c.

It is clear that with the following scheme we achieve capacity and communicate without error. What we did is find a *non-confusable* subset of input, such that the output could be uniquely decodable.

The main idea behind Shannon's noisy channel theorem is that for large block lengths, every channel looks like the noisy typewriter; the channel has a subset of inputs that produce essentially disjoint sequences at the output.

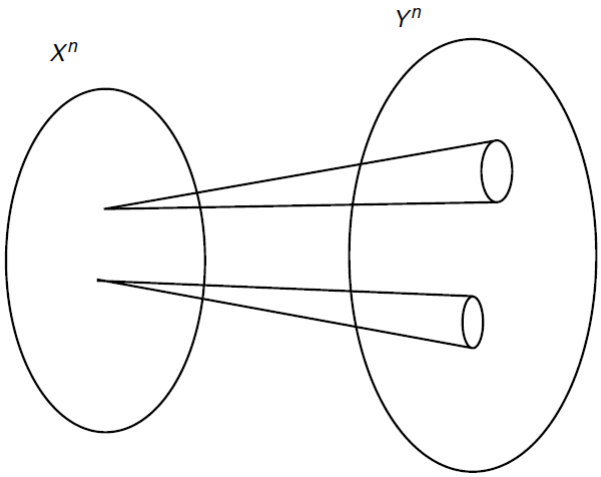

Figure 3.2: Channel after n uses; Elements of information theory(2006)

For this we need the notion of typicality.

Definition 3.1. *Let X be a random variable over the alphabet $\mathcal{X}$. A sequences $x \in \mathcal{X}^n$ of length n is called typical of tolerance β if and only if*

$$|\frac{1}{n}\log\frac{1}{p(x^n)} - H(X)| < \beta.$$

We define $A_\beta^{(n)}$ to be the set of typical sequences.

Example 3.2. *Suppose we flip a fair coin 10 times, then*

$$x := 101100101$$

is typical for every $\beta \geq 0$ as we have five ones and five zeros.

Example 3.3. *Let X, Y be a random variable over the alphabets $\mathcal{X}, \mathcal{Y}$. Two sequences $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$ of length n are called typical of tolerance β if and only if both x and y are typical and they are jointly typical as well:*

$$|\frac{1}{n}\log\frac{1}{p(x^n, y^n)} - H(X,Y)| < \beta.$$

Note that this condition is necessary if X and Y are dependent but not if they are independent, suppose that they are independent, then the expression $|\frac{1}{n}\log\frac{1}{p(x^n,y^n)} - H(X,Y)|$ equals 0 because $H(X,Y) = H(X) + H(Y)$ and $\log\frac{1}{p(x^n,y^n)} = \log\frac{1}{p(x^n)} + \log\frac{1}{p(y^n)}$. Jointly typical sequences have the following useful properties.

Theorem 3.1. *Let (X^n, Y^n) be sequences of length n drawn i.i.d according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then:*

1. *$P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.*
2. *$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.*
3. *If $(X'^n, Y'^n) \sim p(x^n)p(y^n)$ [i.e. X'^n and Y'^n are independent with the same marginals as $p(x^n, y^n)$, then*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq P((X'^n, Y'^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

These statements have the following intuition,

1. "Large messages will always become typical."
2. "The size of set of jointly typical messages is approximately $2^{nH(X,Y)}$."
3. "The probability that any independently chosen pair of typical messages is jointly typical is about $2^{-n(I(X;Y))}$."

The first statement follows quite quickly from the weak law of large numbers applied to $\frac{1}{n}\log P(X^n)$, $\frac{1}{n}\log P(y^n)$ and $\frac{1}{n}\log P(X^n, Y^n)$. A formal proof is omitted. The second statement follows quickly from the fact that summing probabilities over the whole space equals 1 and if $|\frac{1}{n}\log\frac{1}{p(x^n,y^n)} - H(X,Y)| < \epsilon$, then $-\epsilon < \frac{1}{n}\log\frac{1}{p(x^n,y^n)} - H(X,Y) < \epsilon$ and,

$$-\epsilon < \frac{1}{n}\log\frac{1}{p(x^n,y^n)} - H(X,Y) < \epsilon \iff 2^{-n(H(X,Y)+\epsilon)} < p(x^n, y^n) < 2^{-n(H(X,Y)-\epsilon)}.$$

It follows that for the upper bound,

$$\begin{aligned} 1 &= \sum p(x^n, y^n) \\ &\geq \sum p(x^n, y^n) \\ &\geq |A_\epsilon|^n 2^{-n(H(X,Y)+\epsilon)}, \end{aligned}$$

and hence $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$. Similar reasoning applies to the lower bound. The third statement is left as an exercise to the reader as there is not enough space left on this page. (Hint: Use the second part of the theorem and the above if and only if statement). The following figure will provide some intuition as to why the third statement is true.

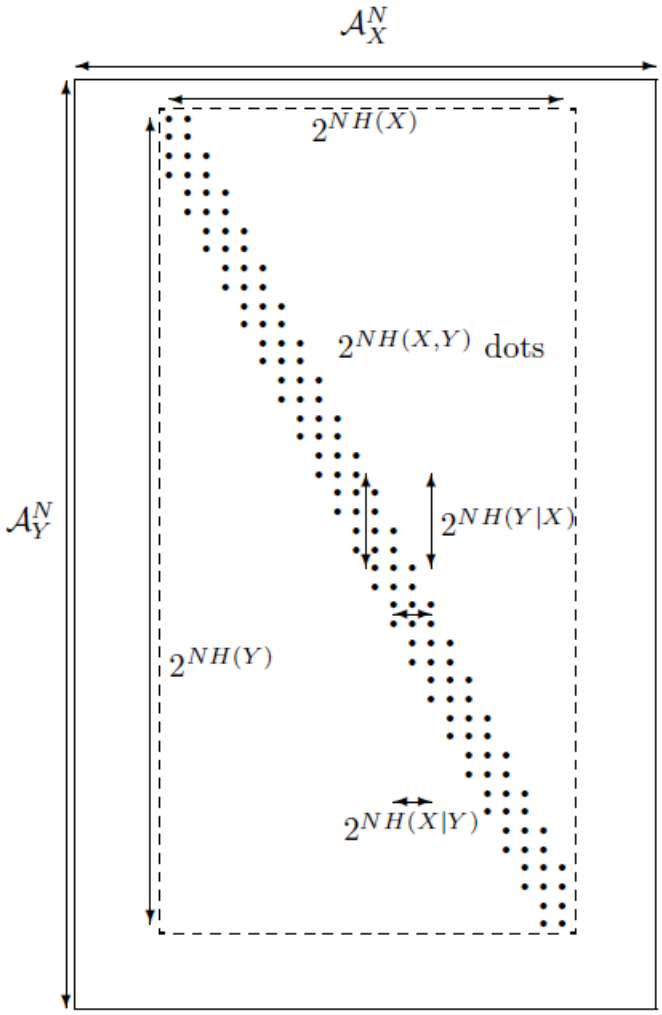

Figure 3.3: The total number of independent typical pairs is the area of the dashed rectangle; $2^{nH(X)}2^{nH(Y)} = 2^{nH(X)+nH(Y)}$, and the number of jointly-typical pairs is roughly $2^{nH(X,Y)}$, so the probability of hitting a jointly-typical pair is roughly $2^{nH(X,Y)}/2^{nH(X)+nH(Y)} = 2^{-nI(X;Y)}$; Mackay(2003)

4 Proof of Shannon's noisy-channel theorem

We can now prove Shannon's noisy channel theorem, the proof will use the notion of typicality to think of a smart encoding and decoding the scheme. The outline of the proof will be as follows:

- Generate a code randomly from a certain distribution.
- Decode by joint typicality.
- Calculate the average probability of error for a random choice of codewords and show that it becomes arbitrarily small.

Theorem 4.1. *For a discrete memory-less channel, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.*

Proof. Choose $p(x)$ to be the distribution on X that achieves capacity. We generate 2^{nR} codewords independently according to the distribution

$$p(x^n) = \prod_{i=1}^{n} p(x_i).$$

Now, our block length is n and we have to generate 2^{nR} such codewords, this gives us $n * 2^{nR}$ entries that need to be generated (independently) according to $p(x)$. Thus, the probability for a certain code is

$$Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w)).$$

Consider the following events,

1. This code $\mathcal{C}$ is revealed to both the sender and receiver.
2. A message w is chosen and the wth codeword $X^n(w)$ is sent over the channel.
3. The receiver receives a sequence according to the distribution

$$P(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))$$

4. The receiver decodes by joint typicality, that is; the receiver declares that the message $\tilde{W}$ was sent if the following conditions are satisified:
   - $(X^n(\tilde{W}), Y^n)$ is jointly typical
   - There is no other index $W' \neq \tilde{W}$ such that $(X^n(W'), Y^n)$ are jointly typical

   If no such W' exists or if there is more than one, an error is declared

Now there is a decoding eror if $W' \neq W$. We define $\mathcal{E}$ to be the event that $\{W' \neq W\}$. For a typical codeword there are two error when we use jointly typical decoding,

1. The output Y^n is not jointly typical with the transmitted codeword.
2. There is some other codeword that is jointly typical with Y^n.

We know from (1) of Theorem 3.1 that the probability that the transmitted codeword and the received sequence are jointly typical goes to 1 for large n. And for any other codeword the probability that it is jointly typical with the received sequence is approximately 2^{-nC} by (3) of Theorem 3.1. As we choose the distribution that achieved capacity, more formally we will calculate the average probability of error, averaged over all codewords in the codebook, *and* averaged over all codewords; that is, we calculate

$$Pr(\mathcal{E}) = \sum_{\mathcal{C}} Pr(\mathcal{C}) P_\epsilon^n(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^n R} \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_w(\mathcal{C}).$$

As every codeword is generated independently according to the same distribution the *average probability* of error averaged over all codes does not depend on the particular index that was sent thus we can assume without loss of generality that the message $W = 1$ was sent, thus $Pr(\mathcal{E}) = Pr(\mathcal{E}|W = 1)$. Now define the event E_i to be the event that the ith codeword and Y^n are jointly typical, that is

$$E_i = \{X^n(i), Y^n) \in A_\epsilon^n\} \text{ for } i \in \{1, 2, \ldots, 2^{nR}\}.$$

Hence, noting that we assumed $W = 1$ was sent, an error happens when E_1^C or $E_2 \cup E_3 \cup \ldots \cup E_{2^n R}$ occurs. Which respectively corresponds to an error of type 1 and an error of type 2. Now,

$$\begin{aligned} Pr(\mathcal{E}) = Pr(\mathcal{E}|W = 1) &= P(E_1^C \cup E_2 \cup \ldots \cup E_{2^n R}|W = 1) \\ &\leq P(E_1^C|W = 1) + \sum_{i=2}^{2^{nR}} P(E_i|W = 1) \end{aligned}$$

By the union bound. We know by (1) from Theorem 3.1 that $P(E_1^C|W = 1)$ will become arbitrarily small, and since by the code generation process, X^n and $X^n(i)$ are independent for $i \neq 1$ and so are Y^n and $X^n(i)$, hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(C-3\epsilon)}$ by (3) from Theorem 3.1. And thus,

$$
\begin{aligned}
Pr(\mathcal{E}) &\leq P(E_1^C|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1) \\
&\leq \epsilon_1 + \sum_{i=2}^{2^{nR}} 2^{-n(C-3\epsilon)} \\
&= \epsilon_1 + (2^{nR}-1)2^{-n(C-3\epsilon)} \\
&\leq \epsilon_1 + 2^{-n(C-3\epsilon-R)} \\
&\leq 2\epsilon
\end{aligned}
$$

Where we choose n to be sufficiently large and $R < C - 3\epsilon$. Hence, if $R < C$ we can choose ϵ and n such that the average probability of error, averaged over codebooks and codewords, is less than 2ϵ. This proves Shannon's noisy channel theorem. $\square$

5 Conclusion

In conclusion, we have constructed a formal framework of noisy channels and shown that with the right scheme we can send information with arbitrarily small error while retaining a high rate of information transmitted. An interesting question is the impact of Shannon's noisy channel theorem in practice. The theorem is an asymptotic statement, that is, we can send messages with arbitrarily small error for large n, and in practice the value of n might be too large to use the above random code scheme. An open question is thus to find theoretically how large n must be for a particular error(ϵ). Shannon's noisy channel theorem does however give us confidence that it is possible to communicate with a high rate and small error and forces us to look for smarter schemes that can send information with a rate below capacity. Also, the power of the formal framework allows us to ask new questions and gain deeper insights in noisy channels.

Bibliography

[1] David J.C. MacKay - *Information theory, inference and learning* - Cambridge: Cambridge univeristy press, 2003.

[2] Thomas M. Cover, Joy A - *Elements of information theory* - New york: Wiley-Interscience, 2006