Shannon's noisy-channel theorem

Information theory

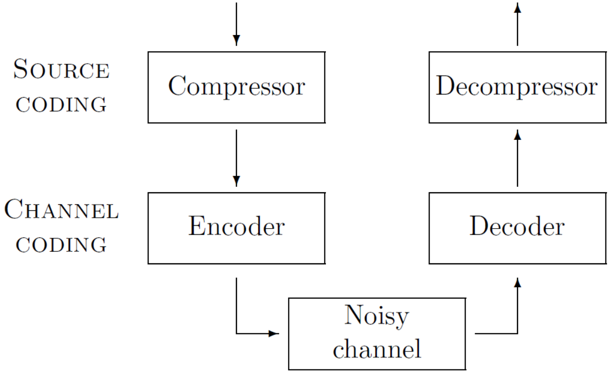
Amon Elders

Korteweg de Vries Institute for Mathematics
University of Amsterdam.

Tuesday, 26th of Januari

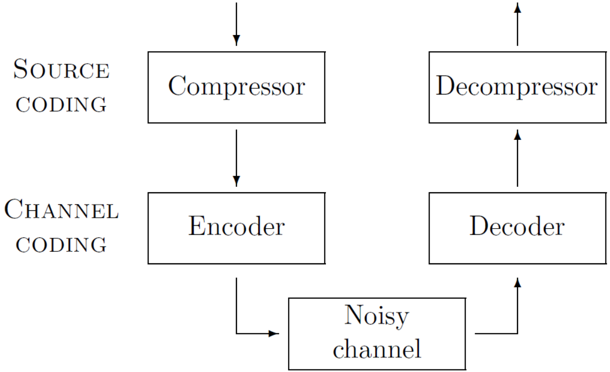
Noisy channel

- During the course we assumed that information sent over a channel was noise-free, that is information was losslessly transmitted through the channel. Real channels are noisy.



Noisy channel

- During the course we assumed that information sent over a channel was noise-free, that is information was losslessly transmitted through the channel. Real channels are noisy.



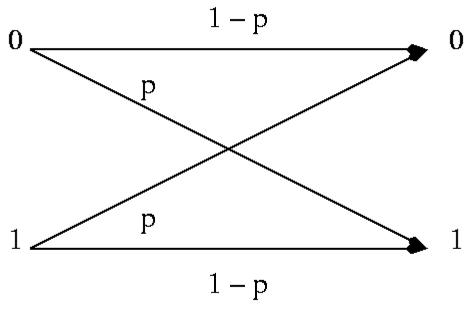
- We want to find out how to send messages through a noisy channel such that the rate of messages send is maximized, but the error is small.

Outline

1. The problem of noisy-channels
   - Example
   - Definitions
2. Shannon's noisy-channel theorem
   - Idea proof
   - Outline of proof

Binary Symmetric Channel

Example



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

Discrete memoryless channel

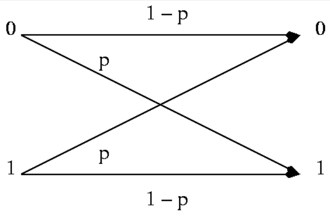
Definition

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:

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

The $n'th$ extension is the channel $(\mathcal{X}^n, P(Y^n|X^n), \mathcal{Y}^n)$ with $p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$

Binary symmetric channel



Definitions

Capacity

We define the *information channel capacity* of a discrete memoryless channel as

$$C = \max_{P_x} \mathcal{I}(X;Y)$$

Where P_x is the probability distribution of X, the random variable over the alphabet $\mathcal{X}$.

Definitions

Capacity

We define the *information channel capacity* of a discrete memoryless channel as

$$C = \max_{P_x} \mathcal{I}(X;Y)$$

Where P_x is the probability distribution of X, the random variable over the alphabet $\mathcal{X}$.

Block code

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 $y \in \mathcal{Y}^n$.

Formal notions of error

Given that i was sent, the probability of error:

$$\lambda_i = P(g(Y^n) \neq i | X^n = x^n(i))$$

Average :

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

Formal notions of error

Given that *i* was sent, the probability of error:

$$\lambda_i = P(g(Y^n) \neq i | X^n = x^n(i))$$

Average :

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

Rate

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

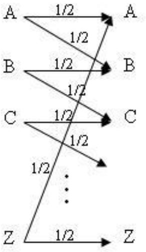
$$R = \frac{\log(M)}{n}, \text{bits per transmission}$$

Shannon's noisy-channel theorem

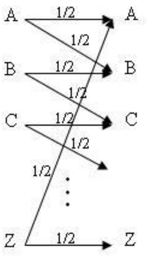
Theorem

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 $\lambda^{(n)} \to 0$

Noisy-Typewriter



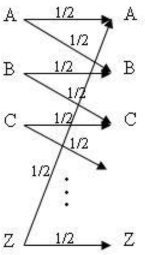
Noisy-Typewriter



Example of theorem

- We take the index set to be: $\{1, 2, \ldots, 13\}$

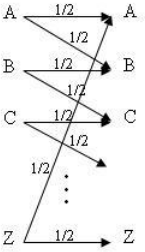
Noisy-Typewriter



Example of theorem

- We take the index set to be: $\{1, 2, \ldots, 13\}$
- The following encoding function $X(1) = a, X(2) = c, \ldots, x(M) = y$

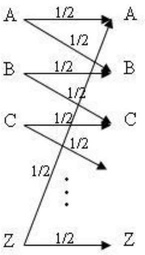
Noisy-Typewriter



Example of theorem

- We take the index set to be: $\{1, 2, \ldots, 13\}$
- The following encoding function $X(1) = a, X(2) = c, \ldots, x(M) = y$
- The decoding function maps the received letter to the nearest letter in the code.

Noisy-Typewriter



Example of theorem

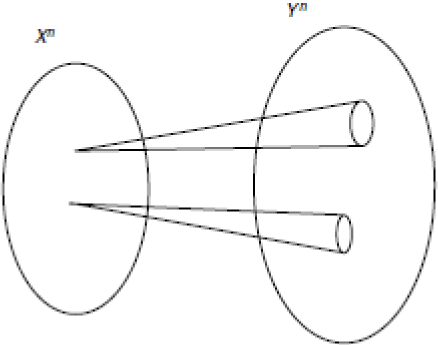
- We take the index set to be: $\{1, 2, \ldots, 13\}$
- The following encoding function $X(1) = a, X(2) = c, \ldots, x(M) = y$
- The decoding function maps the received letter to the nearest letter in the code.

Then our rate $R = \frac{log(13)}{1}$, which can be shown to be smaller than capacity and the error is always zero.

Idea

Idea

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.



Typical sequences

Definition typical sequence

Let X be a random variable over an alphabet $\mathcal{X}$. A sequence $x \in \mathcal{X}$ of length n is called typical of tolerance β if and only if

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

Typical sequences

Definition typical sequence

Let X be a random variable over an alphabet $\mathcal{X}$. A sequence $x \in \mathcal{X}$ of length n is called typical of tolerance β if and only if

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

Example

Suppose we flip a coin 10 times, then

$$x := 1111100000$$

Is typical for every $\beta \geq 0$.

Joint Typicality

Definition jointly typical sequence

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

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

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

Typicality theorems

Theorem

Typicality theorem: 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* $(X'^n, Y'^n) \sim p(x^n)p(y^n)$*, then*

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

Intuition

1. "Large messages wil always become typical"

Typicality theorems

Theorem

Typicality theorem: 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* $(X'^n, Y'^n) \sim p(x^n)p(y^n)$, *then*

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

Intuition

1. "Large messages wil always become typical"
2. "Size of set of typical messages is approximately $2^{nH(X,Y)}$"

Typicality theorems

Theorem

Typicality theorem: 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* $(X'^n, Y'^n) \sim p(x^n)p(y^n)$, *then*

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

Intuition

1. "Large messages wil always become typical"
2. "Size of set of typical messages is approximately $2^{nH(X,Y)}$"
3. "The odds that two random messages are jointly typical is small for large n and depends on the mutual information"

Decoding

Decoding by joint typicality

- The probability that any pair of typical X^n and Y^n are jointly typical is about $2^{(-n(I(X;Y))}$(part 3 of theorem),

Decoding

Decoding by joint typicality

- The probability that any pair of typical X^n and Y^n are jointly typical is about $2^{(-n(I(X;Y))}$(part 3 of theorem),
- Hence we expect that if we consider $2^{n(I(X;Y))}$ such pairs before coming across a jointly typical pair.

Decoding

Decoding by joint typicality

- The probability that any pair of typical X^n and Y^n are jointly typical is about $2^{(-n(I(X;Y))}$(part 3 of theorem),
- Hence we expect that if we consider $2^{n(I(X;Y))}$ such pairs before coming across a jointly typical pair.
- Thus if we decode based on joint typicality, the odds that we confuse a codeword with the codeword that caused the output Y^n is small if we have $2^{n(I(X;Y))}$ codewords.

Outline of proof

Theorem

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 outline

- We select 2^{nR} random codewords.

Outline of proof

Theorem

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 outline

- We select 2^{nR} random codewords.
- Decode by joint typicality

Outline of proof

Theorem

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 outline

- We select 2^{nR} random codewords.
- Decode by joint typicality
- Analyse the error, there are 2 types:
  - The output Y^n is not jointly typical with the transmitted codeword
  - There is some other codeword jointly typical with Y^n

Outline of proof

Theorem

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 outline

- We select 2^{nR} random codewords.
- Decode by joint typicality
- Analyse the error, there are 2 types:
  - The output Y^n is not jointly typical with the transmitted codeword
  - There is some other codeword jointly typical with Y^n
- The first error type goes to zero because of (1) from previous slide

Outline of proof

Theorem

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 outline

- We select 2^{nR} random codewords.
- Decode by joint typicality
- Analyse the error, there are 2 types:
  - The output Y^n is not jointly typical with the transmitted codeword
  - There is some other codeword jointly typical with Y^n
- The first error type goes to zero because of (1) from previous slide
- The second error goes to zero if $R < C$ and because of reasoning above.

Proof

Theorem

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 $\lambda^{(n)} \to 0$

Creating a code

Fix $p(x)$. Generate 2^{nR} codewords independently at random according to the distribution

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

And assign a codeword $X(W)$ to each message W. Note that this code has rate R. Furthermore, we make this code known to both the sender and receiver.

Decoding

Theorem

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 $\lambda^{(n)} \to 0$

Decoding

The receiver declares that the message $\tilde{W}$ was sent if the following conditions are satisified:

1. $(X^n(\tilde{W}), Y^n)$ is jointly typical
2. There is no other index $W' \neq \tilde{W}$ such that $(X^n(W'), Y^n)$ are jointly typical

Thus we make a mistake when:

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

Analysing error (1)

Theorem

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 $\lambda^{(n)} \to 0$

Analysing error (1)

By the first part of the typicality theorem we know that

$$\forall \epsilon \exists N : P((X^n(\tilde{W}), Y^n) \notin A_\epsilon^{(n)}) \leq \epsilon$$

Analysing error (2)

Theorem

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 $\lambda^{(n)} \to 0$

Analysing error (2)

We know by the third part of the typicality theorem that a random $X^n(W')$ and Y^n are jointly typical with odds $\leq 2^{-n(I(X;Y)-3\epsilon)}$. There are $2^{nR} - 1$ such cases.

Overall error

Theorem

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 $\lambda^{(n)} \to 0$

Overall error

Thus with the union bound and the previous slide:
$Pr(Error) = Pr(Error1 \cup Error2) \leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}$

If n is sufficiently large and $R < I(X;Y) - 3\epsilon$ we get: $Pr(Error) \leq 2\epsilon$

Overall error

Theorem

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 $\lambda^{(n)} \to 0$

Overall error

Thus with the union bound and the previous slide:
$Pr(Error) = Pr(Error1 \cup Error2) \leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \leq \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}$

If n is sufficiently large and $R < I(X;Y) - 3\epsilon$ we get: $Pr(Error) \leq 2\epsilon$

If we take our distribution $p(x)$ to be the distribution $p^*(x)$ which achieves capacity we can replace the condition $R < I(X;Y)$ by $R < C$. This proofs our theorem.

Thanks for listening!