Shannon's Noisy-Channel Coding Theorem

Lucas Slot Sebastian Zur

February 13, 2015

Outline

1. Definitions and Terminology
   - Discrete Memoryless Channels
   - Terminology
   - Jointly Typical Sets

2. Noisy-Channel Coding Theorem
   - Statement
   - Part one
   - Part two
   - Part three

Outline

Discrete Memoryless Channels

Definition

A discrete memoryless channel consist of two random variables X and Y over finite discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ that satisfy

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x) \ \forall x, y \in \mathcal{X} \times \mathcal{Y}$$

where X is the input and Y is the output of the channel.

Discrete Memoryless Channels

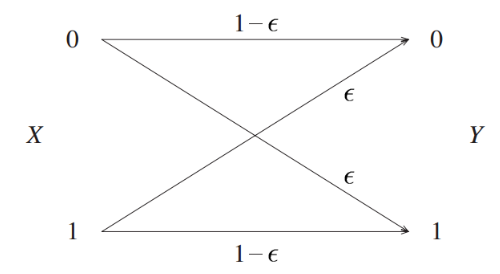
Definition

A discrete memoryless channel consist of two random variables X and Y over finite discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ that satisfy

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x) \; \forall x, y \in \mathcal{X} \times \mathcal{Y}$$

where X is the input and Y is the output of the channel.

Example



Outline

Definitions

Block Code

A Block Code converts a sequence of source bits s with length K into a sequence t of length N with $N > K$.

Probability of Block Error

The p_B of a code and decoder is:
$\sum_{s_{in}} P(s_{in})P(s_{out} \neq s_{in}|s_{in})$.

Optimal Decoder

An optimal decoder is the decoder which minimalises the probability of block error, by decoding an output y as input s, where $P(s|y)$ is maximalised.

Probability of Bit Error

The p_b of a code and decoder is the average probability that a bit is s_{out} is not equal to the correspoding bit in s_{in}.

Outline

Typical Sequences

Definition

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

$$|\frac{1}{N} \cdot \log \frac{1}{P(x)} - H(X)| < \beta$$

Example

Suppose X is the result of a coin flip. The sequence

$$x := 1000111001101100$$

is typical to any tolerance $\beta \geq 0$.

Jointly Typical Sequences

Definition

Let X, Y be random variables over alphabets $\mathcal{X}$ and $\mathcal{Y}$. Two sequences $x \in \mathcal{X}^N$ and $y \in \mathcal{Y}$ of length N are called jointly typical to tolerance β if and only if both x and y are typical and

$$|\frac{1}{N} \cdot \log \frac{1}{P(x,y)} - H(X,Y)| < \beta$$

Jointly Typical Sequences

Definition

Let X, Y be random variables over alphabets $\mathcal{X}$ and $\mathcal{Y}$. Two sequences $x \in \mathcal{X}^N$ and $y \in \mathcal{Y}$ of length N are called jointly typical to tolerance β if and only if both x and y are typical and

$$|\frac{1}{N} \cdot \log \frac{1}{P(x,y)} - H(X,Y)| < \beta$$

Example

Suppose X and Y are both random variables such that $P(x = 1) = 0.5$ and $P(y|x)$ corresponds to a channel of noise level 0.3. The sequences

$$x := 1111100000$$
$$y := 0001100000$$

are typical to any tolerance $\beta \geq 0$.

Jointly Typical Set 1

Definition

Let X, Y be random variables over alphabets $\mathcal{X}$ and $\mathcal{Y}$. The set $J_{N,\beta}$ that contains all pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ of length N jointly typical to tolerance β is called the jointly-typical set.

Jointly Typical Theorem

Theorem

Let x, y be drawn from $(XY)^N$ which is defined by

$$P(x,y) = \prod_{i=1}^{n} P(x_n, y_n)$$

1. The probability that x, y are jointly typical to tolerance β tends to 1 as $N \to \infty$

Jointly Typical Theorem

Theorem

Let x, y be drawn from $(XY)^N$ which is defined by

$$P(x, y) = \prod_{i=1}^{n} P(x_n, y_n)$$

1. The probability that x, y are jointly typical to tolerance β tends to 1 as $N \to \infty$
2. The number of jointly typical sequences $|J_{N,\beta}| \leq 2^{N(H(X,Y)+\beta)}$

Jointly Typical Theorem

Theorem

Let x, y be drawn from $(XY)^N$ which is defined by

$$P(x, y) = \prod_{i=1}^{n} P(x_n, y_n)$$

1. The probability that x, y are jointly typical to tolerance β tends to 1 as $N \to \infty$
2. The number of jointly typical sequences $|J_{N,\beta}| \leq 2^{N(H(X,Y)+\beta)}$
3. For any two sequences x, y chosen *independently* from X^N and Y^N respectively that have the same marginal distribution as $P(x, y)$ we have

$$P((x, y) \in J_{N,\beta}) \leq 2^{-N(I(X,Y)-3\beta)}$$

An Illustration

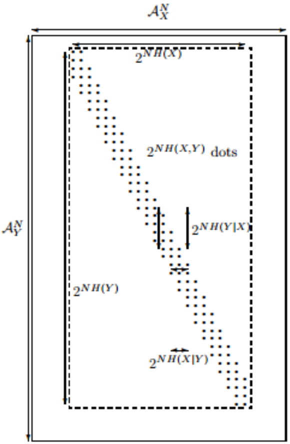


Figure: Typical Sets (from MacKay 2003)

Outline

Statement

Theorem

1. For every discrete memoryless channel. the channel capacity

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

satisfies the following property. For any $\epsilon > 0$ And rate $R < C$, for sufficiently large N, there is a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

Statement

Theorem

1. For every discrete memoryless channel. the channel capacity

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

satisfies the following property. For any $\epsilon > 0$ And rate $R < C$, for sufficiently large N, there is a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

2. If we accept bit error with probability p_b, it is possible to achieve rates up to $R(p_b)$, where

$$R(p_b) = \frac{C}{1 - h(p_b)}.$$

Statement

Theorem

1. For every discrete memoryless channel. the channel capacity

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

satisfies the following property. For any $\epsilon > 0$ And rate $R < C$, for sufficiently large N, there is a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

2. If we accept bit error with probability p_b, it is possible to achieve rates up to $R(p_b)$, where

$$R(p_b) = \frac{C}{1 - h(p_b)}.$$

3. Rates greater than $R(p_b)$ are not achievable.

Outline

Proof: An Analogy

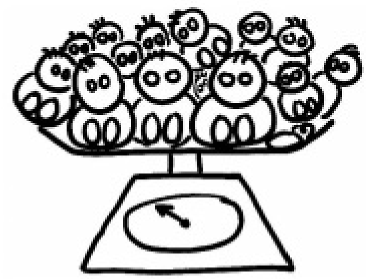


Figure: Weighing Babies (from MacKay 2003)

Coding

Creating a Code

Consider a fixed distribution $P(x)$. We will generate $S = 2^{NR'}$ codewords at random using

$$P(x) = \prod_{n=1}^{N} P(x_n)$$

and assign a codeword $x^{(s)}$ to each message s. We make this code known to both sides of the channel.

Important!

This code has a rate of R'!

Decoding

Received Signal

The signal received on the other end of the channel is y, with

$$P(y|x^{(s)}) = \prod_{n=1}^{N} P(y_n|x_n^{(s)})$$

Decoding

We will decode using *typical-set decoding*. We will decode y as s if $(x^{(s)}, y)$ are jointly typical and there is no other message s' such that $(x^{(s')}, y)$ are jointly typical.

Mistakes

We will make a mistake when

1. There is no jointly typical $x^{(s)}$.
2. There are multiple jointly typical $x^{(s)}$.

Three Types of Errors

Block Error

$p_B(\mathcal{C}) \equiv P(\hat{s} \neq s | \mathcal{C})$

Three Types of Errors

Block Error

$p_B(\mathcal{C}) \equiv P(\hat{s} \neq s|\mathcal{C})$

Average Block Error

$< p_B > \equiv \sum_{\mathcal{C}} P(\hat{s} \neq s|\mathcal{C})P(\mathcal{C})$

Three Types of Errors

Block Error

$p_B(\mathcal{C}) \equiv P(\hat{s} \neq s|\mathcal{C})$

Average Block Error

$< p_B > \equiv \sum_{\mathcal{C}} P(\hat{s} \neq s|\mathcal{C})P(\mathcal{C})$

Maximal Block Error

$p_{BM}(\mathcal{C}) \equiv \max_s P(\hat{s} \neq s|s, \mathcal{C})$

Bounding the Errors (1)

No Jointly Typical $x^{(s)}$

By the first part of the jointly typical theorem

$$\forall \delta \; \exists N(\delta) : P(x^{(1)}, y) \notin J_{N,\beta}) < 2\delta$$

Too Many Jointly Typical $x^{(s)}$

The chance for a random $x^{(s')}$ and y to be jointly typical $\leq 2^{-N(I(X;Y)-3\beta}$. There are $2^{(NR')} - 1$ candidates.

Bounding the Errors (2)

Average Block Error

Now, using the union bound we find

$$
\begin{aligned}
< p_{BM} > &\leq \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta} \\
&\leq \delta + 2^{-N(I(X;Y)-R'-3\beta)}
\end{aligned}
$$

If $R' < I(X;Y) - 3\beta$ we can make this error very small (smaller than 2δ).

Three modifications

Pick the best $P(x)$

We choose the best $P(x)$, so $R' < I(X;Y) - 3\beta$ becomes $R' < C - 3\beta$.

Pick the best $\mathcal{C}$

Using our baby-argument, there must be a code with $p_B(\mathcal{C}) < 2\delta$

Perform a trick

From this $\mathcal{C}$ we now toss the worst half of the codewords. Those that remain must have probability of error $< 4\delta$. We define a new code with these $(2^{NR'-1})$ codewords.

Conclusion

We have proven the existence of a code $\mathcal{C}$ with rate $R' < C - 3\beta$ with a maximal probability of error $< 4\delta$. The theorem can now be proven be setting

$$R' = \frac{R + C}{2}$$
$$\delta = \frac{\epsilon}{4}$$
$$\beta < \frac{(C - R')}{3}$$
$$N \text{ big enough}$$

Outline

Overview

1. What happens when we try to communicate with a rate, greater than the capacity?
2. We could just send $1/R$ of the source bits and guess the rest of the source. This will give us an average p_b of $\frac{1}{2}(1 - 1/R)$.
3. However, it turns out we can minimalise this error so that we get: $H_2(p_b) = 1 - 1/R$.

Method

1. We take an excellent (N, K) code with rate $R' = K/N$.
2. This code is capable of correcting errors in our channel with transition probability q.
3. Asymptotically we may assume that $R \simeq 1 - H_2(p_b)$.
4. We know chop our source our source up in blocks of length N and pass it through our decoder, which gives us blocks of length K, which then get communicated over the noiseless channel.
5. When we pass this new message to our encoder, we will receiver a message which we will differ at an average of qN bits from the original message, so $p_b = q$.
6. Attaching this compressor to our capacity-C error-free communicator we get a rate of $R = \frac{NC}{K} = \frac{C}{R'} = \frac{C}{1-H_2(p_b)}$.

Outline

Proof

1. A Markov chain is defined by the source, encoder, noisy channel and the decoder:

$$P(s, x, y, \hat{s}) = P(s)P(x|s)P(y|x)P(\hat{s}|y)$$

Proof

1. A Markov chain is defined by the source, encoder, noisy channel and the decoder:

$$P(s, x, y, \hat{s}) = P(s)P(x|s)P(y|x)P(\hat{s}|y)$$

2. $I(s; \hat{s}) \leq I(x; y)$ because of the data processing inequality.

Proof

1. A Markov chain is defined by the source, encoder, noisy channel and the decoder:

$$P(s, x, y, \hat{s}) = P(s)P(x|s)P(y|x)P(\hat{s}|y)$$

2. $I(s; \hat{s}) \leq I(x; y)$ because of the data processing inequality.
3. By the definition of channel capacity we know that $I(x; y) \leq NC$, so $I(s; \hat{s}) \leq NC$.

Proof

1. A Markov chain is defined by the source, encoder, noisy channel and the decoder:

$$P(s,x,y,\hat{s}) = P(s)P(x|s)P(y|x)P(\hat{s}|y)$$

2. $I(s;\hat{s}) \leq I(x;y)$ because of the data processing inequality.
3. By the definition of channel capacity we know that $I(x;y) \leq NC$, so $I(s;\hat{s}) \leq NC$.
4. Assuming that our system has a rate R and bit error probability p_b, then $I(s;\hat{s}) \geq NR(1 - H_2(p_b))$.

Proof

1. A Markov chain is defined by the source, encoder, noisy channel and the decoder:

$$P(s, x, y, \hat{s}) = P(s)P(x|s)P(y|x)P(\hat{s}|y)$$

2. $I(s; \hat{s}) \leq I(x; y)$ because of the data processing inequality.
3. By the definition of channel capacity we know that $I(x; y) \leq NC$, so $I(s; \hat{s}) \leq NC$.
4. Assuming that our system has a rate R and bit error probability p_b, then $I(s; \hat{s}) \geq NR(1 - H_2(p_b))$.
5. If $R > R(p_b) = \frac{C}{1-H_2(p_b)}$, then $I(s; \hat{s}) \geq NC$. This gives a contradiction, so for any p_b, there is no larger rate possible than $R(p_b)$.

References

David J.C. MacKay - *Information Theory, Inference, and Learning Algorithms* - Cambridge University Press 2003 - Accessed via http://www.inference.phy.cam.ac.uk/itprnn/book.pdf