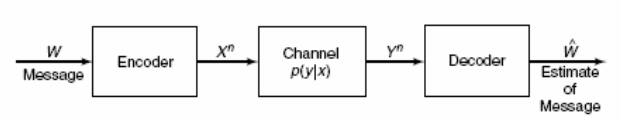
Communication System

A communication system can be represented as in Figure.

A message W, drawn from the index set $\{1, 2, \ldots, M\}$, results in the signal $X^n(W)$, which is received by the receiver as a random sequence $Y^n \sim p(y^n|x^n)$.

The receiver then guesses the index W by an appropriate decoding rule $\hat{W}=g(Y^n)$.

The receiver makes an error if $\hat{W}$ is not the same as the index W that was transmitted.



Discrete Channel and its Extension

Definition A *discrete channel*, denoted by $(\chi, p(y|x), \gamma)$, consists of two finite sets χ and γ and a collection of probability mass functions $p(y|x)$, one for each $x \in X$, such that for every x and y, $p(y|x) \geq 0$, and for every x, $\sum_y p(y|x) = 1$, with the interpretation that X is the input and Y is the output of the channel.

Definition The *nth extension* of the discrete memoryless channel (DMC) is the channel $(\chi^n, p(y^n|x^n), \gamma^n)$,

where $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$, $k = 1, 2, \ldots, n$.

Channel Without Feedback

If the channel is used *without feedback*:

i.e., if the input symbols do not depend on the past output symbols, namely,

$p(x_k | x^{k-1}, y^{k-1}) = p(x_k | x^{k-1})$,

the channel transition function for the nth extension of the discrete memoryless channel reduces to

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

(M,n) Code

An *(M, n)* code for the channel $(\chi, p(y|x), \gamma)$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \rightarrow \chi^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$. The set of codewords is called the *codebook*.
3. A decoding function $g : \gamma^n \rightarrow \{1, 2, \ldots, M\}$,

which is a deterministic rule that assigns a guess to each possible received vector.

Definitions

Definition (*Conditional probability of error*) Let

$$\lambda_i = \Pr(g(Y^n) \neq i \mid X^n = x^n(i))$$
$$= \sum_{y^n} p(y^n \mid x^n(i)) I(g(y^n) \neq i)$$

be the *conditional probability of error* given that index *i* was sent, where *I (·)* is the indicator function.

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

$$\lambda = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$

Definitions

Definition The (*arithmetic*) *average probability of error* $P_e^{(n)}$ for an *(M, n)* code is defined as

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

Note that if the index *W* is chosen according to a uniform distribution over the set {1, 2, . . . , *M*}, and $X^n = x^n(W)$, then by definition

$$P_e^{(n)} = \Pr(W \neq g(Y^n))$$

Also $P_e^{(n)} \leq \lambda^{(n)}$

Definitions

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

$$R = \frac{\log M}{n}$$

bits per transmission.

Definition A rate R is said to be *achievable* if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda(n)$ tends to 0 as $n \to \infty$.
We write $(2^{nR}, n)$ codes to mean $(\lceil 2^{nR} \rceil, n)$ codes.

Definition The *capacity* of a channel is the supremum of all achievable rates.

Thus, rates less than capacity yield arbitrarily small probability of error for sufficiently large block lengths. Note that if C is low, this means that we need large n to decode M symbols with a low probability of error

Jointly Typical Sequence

Roughly speaking, we decode a channel output Y^n as the ith index if the codeword $X^n(i)$ is "jointly typical" with the received signal Y^n.

Definition The set $A_\epsilon^{(n)}$ of *jointly typical* sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies -close to the true entropies:

$$\begin{aligned} A_\epsilon^{(n)} = \{&(x^n, y^n) \in \chi^n \times \gamma^n : \\ &\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \varepsilon, \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \varepsilon, \\ &\left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \varepsilon \} \end{aligned}$$

where:

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

Jointly Typical Sequence

Theorem (*Joint AEP*) Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to

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

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

$$\Pr\left(\tilde{X}^n, \tilde{Y}^n \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

Also, for sufficiently large n:

$$\Pr\left(\tilde{X}^n, \tilde{Y}^n \in A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\varepsilon)}$$

Jointly Typical Sequence

We see the proof of the 3rd part only, since the other two can be easilty proved using the AEP.

If $\sim X^n$ and $\sim Y^n$ are independent but have the same marginals as X^n and Y^n, then:

From 2 ($|A|$)

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n)\, p(y^n)$$

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

$$\leq 2^{n(H(X,Y)+\varepsilon)} 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)}$$

$$= 2^{-n(I(X;Y)-3\varepsilon)}$$

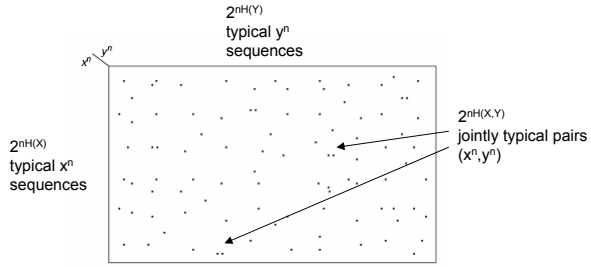
Where the inequality follows from the AEP: $p(x^n) \leq 2^{-n(H(X)-\epsilon)}$ because x^n is a typical sequence. The same applies to y^n.

With this propriety we can say that the probability that two independent sequences are jointly typical is very small for large n

Jointly Typical Set

The jointly typical set is illustrated in Figure. There are about $2^{nH(X)}$ typical X sequences and about $2^{nH(Y)}$ typical Y sequences. However, since there are only $2^{nH(X,Y)}$ jointly typical sequences, not all pairs of typical X^n and typical Y^n are also jointly typical.

The probability that any randomly chosen pair is jointly typical is about $2^{-nI(X;Y)}$. This suggests that there are about $2^{nI(X;Y)}$ distinguishable signals X^n.



Jointly Typical Set

Another way to look at this is in terms of the set of jointly typical sequences for a fixed output sequence Y^n, presumably the output sequence resulting from the true input signal X^n.

For this sequence Y^n, there are about $2^{nH(X|Y)}$ conditionally typical input signals. The probability that some randomly chosen (other) input signal X^n is jointly typical with Y^n is about $2^{nH(X|Y)} / 2^{nH(X)} = 2^{-nI(X;Y)}$.

This again suggests that we can choose about $2^{nI(X;Y)}$ codewords $X^n(W)$ before one of these codewords will get confused with the codeword that caused the output Y^n.

Channel Coding Theorem

Theorem (*Channel coding theorem*) For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda(n) \to 0$.

Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda(n) \to 0$ must have $R \leq C$.

Channel Coding Theorem

The proof makes use of the properties of typical sequences. It is based on the following decoding rule: we decode by joint typicality;

we look for a codeword that is jointly typical with the received sequence.

If we find a unique codeword satisfying this property, we declare that word to be the transmitted codeword.

Channel Coding Theorem

By the properties of joint typicality, with high probability the transmitted codeword and the received sequence are jointly typical, since they are probabilistically related.

Also, the probability that any other codeword looks jointly typical with the received sequence is 2^{-nI} . Hence, if we have fewer then 2^{nI} codewords, then with high probability there will be no other codewords that can be confused with the transmitted codeword, and the probability of error is small.

Proof

We prove that rates $R < C$ are achievable:

Fix $p(x)$. Generate a $(2^{nR}, n)$ code at random according to the distribution $p(x)$. Specifically, we generate 2^{nR} codewords independently according to the distribution

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

We exhibit the $2nR$ codewords as the rows of a matrix:

$$C = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

Each entry in the matrix is generated i.i.d. according to p(x). Thus, the probability that we generate a particular code C is

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

Proof

1. A random code C is generated according to $p(x)$.

2. The code C is then revealed to both sender and receiver. Both sender and receiver are also assumed to know the channel transition matrix $p(y|x)$ for the channel.

3. A message W is chosen according to a uniform distribution: $\Pr(W = w) = 2^{-nR}$, $w = 1, 2, \ldots, 2^{nR}$.

4. The wth codeword $X^n(w)$, corresponding to the wth row of C, is sent over the channel.

5. The receiver receives a sequence Y^n according to the distribution

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

Proof

6. The receiver guesses which message was sent. We will use *jointly typical decoding*, In jointly typical decoding, the receiver declares that the index $\hat{W}$ was sent if the following conditions are satisfied:

- $(X^n(\hat{W}), Y^n)$ is jointly typical.
- There is no other index $W' \neq \hat{W}$ such that $(X^n(W'), Y^n) \in A_\varepsilon^{(n)}$.

If no such $\hat{W}$ exists or if there is more than one such, an error is declared. (We may assume that the receiver outputs a dummy index such as 0 in this case.)

7. There is a decoding error if $\hat{W} \neq W$. Let E be the event $\{\hat{W} \neq W\}$.

Proof: Probability of Error

- We calculate the average probability of error
- It does not depends on the index (because of the symmetry of code construction)
- Two possible sources of error:
  - Y^n is not jointly typical with the transmitted X^n
  - There is some other codeword that is jointly typical with Y^n
- The probability that the transmitted codeword and the received sequence are jointly typical goes to 1 (AEP).
- For any rival codeword, the probability that it is jointly typical with the received sequence is approximately 2^{-nI}
  - hence we can use about 2^{nI} codewords and still have a low probability of error.

Proof: Probability of Error

- We let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$ and use jointly typical decoding $\hat{W}(y^n)$ as described in step 6
- Average probability of error:

$$\Pr(E) = \sum_C \Pr(C) P_e^{(n)}(C)$$

$$= \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C \Pr(C) \lambda_w(C)$$

- By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index, then we can assume W=1 was sent. Then we have:

$$\Pr(E) = \sum_C \Pr(C) \lambda_1(C)$$

$$= \Pr(E \mid W = 1)$$

- Define the following events: $E_i = \{ (X^n(i), Y^n)$ is in $A_\varepsilon^{(n)} \}$, $i \in \{1, 2, \ldots, 2^{nR}\}$, where E_i is the event that the ith codeword and Y^n are jointly typical.
  - Recall that Y^n is the result of sending the first codeword $X^n(1)$ over the channel.
- Error occurs if
  - E_1^c occurs (transmitted codeword and the received sequence are not jointly typical)
  - $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (when a wrong codeword is jointly typical with the received sequence)
- Hence, letting $P(E)$ denote $\Pr(E \mid W = 1)$, we have

$$\Pr(E \mid W = 1) = P(E_1^C \cup E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}} \mid W = 1)$$

$$\leq P(E_1^C \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1)$$

by the union of events bound for probabilities.

- By the joint AEP, $P(E_1^c \mid W = 1) \to 0$, and hence $P(E_1^c \mid W = 1) \leq \varepsilon$ for n sufficiently large.
- $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, then Y^n and $X^n(i)$ are independent.
  - the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\varepsilon)}$ by the joint AEP:

$$\Pr(E) = \Pr(E \mid W = 1) \leq P(E_1^C \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1)$$

$$\leq \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\varepsilon)} = \varepsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\varepsilon)}$$

$$\leq \varepsilon + 2^{3n\varepsilon} 2^{-n(I(X;Y)-R)} \leq 2\varepsilon$$

if n is sufficiently large and $R < I(X; Y) - 3\varepsilon$. Hence, if $R < I(X; Y)$, we can choose and n so that the average probability of error, averaged over codebooks and codewords, is less than 2ε.

Comments

Although the theorem shows that there exist good codes with arbitrarily small probability of error for long block lengths, it does not provide a way of constructing the best codes.

If we used the scheme suggested by the proof and generate a code at random with the appropriate distribution, the code constructed is likely to be good for long block lengths.

We discuss Hamming codes, the simplest of a class of algebraic error correcting codes that can correct one error in a block of bits.