Channel Capacity

Channel Capacity

The communication between A and B is a consequence of a physical act performed by A to induce the desired state in B. This transfer of information is subject to noise.

The communication is successful if the transmitter A and the receiver B agree on what was sent.

In this part we define the channel capacity as the logarithm of the number of distinguishable signals that can be sent through the channel.

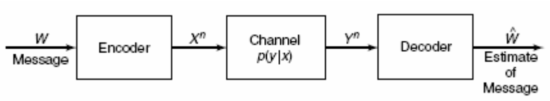
Source symbols from some finite alphabet are mapped into some sequence of channel symbols, which then produces the output sequence of the channel. The output sequence is random but has a distribution that depends on the input sequence.

Channel Capacity

Each of the possible input sequences induces a probability distribution on the output sequences.

We show that we can choose a "nonconfusable" subset of input sequences so that with high probability there is only one highly likely input that could have caused the particular output.

We can transmit a message with very low probability of error and reconstruct the source message at the output. The maximum rate at which this can be done is called the *capacity* of the channel.



Discrete Channel

Definition We define a *discrete channel* to be a system consisting of an input alphabet *X* and output alphabet *Y* and a probability transition matrix *p(y|x)* that expresses the probability of observing the output symbol *y* given that we send the symbol *x*.

The channel is said to be *memoryless* if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.

Information Channel Capacity

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

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

where the maximum is taken over all possible input distributions *p(x)*.

This means: the capacity is the maximum entropy of Y, reduced by the contribution of information given by Y.

We shall soon give an operational definition of channel capacity as the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

Shannon's second theorem establishes that the information channel capacity is equal to the operational channel capacity.

Data Compression and Transmission

There is a duality between the problems of data compression and data transmission.

During compression, we remove all the redundancy in the data to form the most compressed version possible.

During data transmission, we add redundancy in a controlled fashion to combat errors in the channel.

Examples of Channel Capacity

Noiseless binary channel. Suppose that we have a channel whose the binary input is reproduced exactly at the output.

In this case, any transmitted bit is received without error. Hence, one error-free bit can be transmitted per use of the channel, and the capacity is 1 bit.

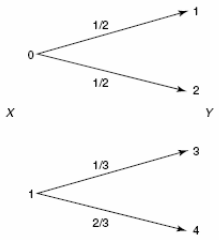
We can also calculate the information capacity $C = \max I(X; Y) = 1$ bit, which is achieved by using $p(x) = (1/2, 1/2)$.

Examples of Channel Capacity

Noisy Channel with Nonoverlapping Outputs. This channel has two possible outputs corresponding to each of the two inputs.

Even though the output of the channel is a random consequence of the input, the input can be determined from the output, and hence every transmitted bit can be recovered without error.

The capacity of this channel is also 1 bit per transmission
We can also calculate the information capacity
$C = \max I(X; Y) = 1$ bit,
which is achieved by using
$p(x) = (1/2, 1/2)$.



Example: Noisy Typewriter

Noisy typewriter. In this case the channel input is either received unchanged at the output with probability 1/2 or is transformed into the next letter with probability 1/2

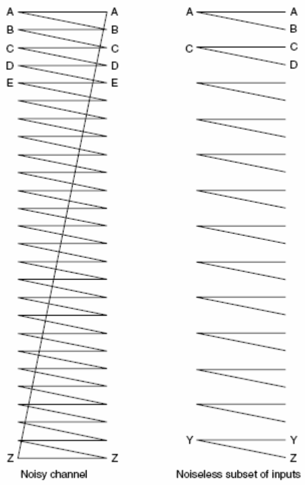
If the input has 26 symbols and we use every alternate input symbol, we can transmit one of 13 symbols without error with each transmission. Hence, the capacity of this channel is log 13 bits per transmission.

We can also calculate the information capacity:

$$C = \max I(X;Y) = \max (H(Y) - H(Y|X))$$
$$= \max H(Y) - 1 = \log 26 - 1 = \log 13,$$

achieved by using *p(x)* distributed uniformly over all the inputs. 1 bit is the uncertainty we have when we read Y, because it could be one character or the next one

Noisy Typewriter



Example: Binary Symmetric Channel

Binary symmetric channel. This is a binary channel in which the input symbols are complemented with probability *p*. This is the simplest model of a channel with errors, yet it captures most of the complexity of the general problem.

We bound the mutual information by:

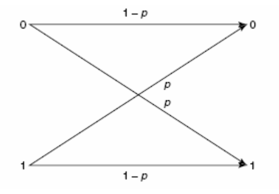
$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - \sum p(x) H(Y \mid X = x) \\ &= H(Y) - \sum p(x) H(p) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p) \end{aligned}$$

where the last inequality follows because Y is a binary random variable. Equality is achieved when the input distribution is uniform.

Binary Symmetric Channel

Hence, the information capacity of a binary symmetric channel with parameter p is

$$C = 1 - H(p)$$



Example: Binary Erasure Channel

The analog of the binary symmetric channel in which some bits are lost (rather than corrupted) is the *binary erasure channel*. In this channel, a fraction α of the bits are erased. The receiver knows which bits have been erased. The binary erasure channel has two inputs and three outputs

We calculate the capacity of the binary erasure channel as follows:

$$
\begin{aligned}
C &= \max I(X;Y) \\
&= \max(H(X) - H(X \mid Y)) \\
&= \max H(Y) - H(\alpha)
\end{aligned}
$$

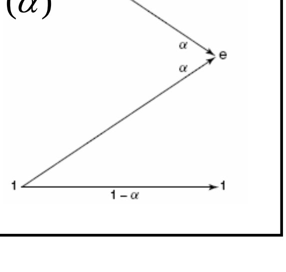
The first guess for the maximum of *H(Y)* would be log 3, but we cannot achieve this by any choice of input distribution *p(x)*. Letting *E* be the event $\{Y = e\}$, using the expansion:

Binary Erasure Channel

Expansion: H(Y)=H(Y,E)=H(E)+H(Y|E)

And letting Pr(X=1)=π we have:

$$
\begin{aligned}
H(Y) &= H((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha) \\
&= H(\alpha) + (1-\alpha)H(\pi) \\
C &= \max H(Y) - H(\alpha) \\
&= \max(1-\alpha)H(\pi) + H(\alpha) - H(\alpha) \\
&= \max(1-\alpha)H(\pi) \\
&= 1-\alpha
\end{aligned}
$$



Comments

Since a proportion α of the bits are lost in the channel, we can recover (at most) a proportion $1 - \alpha$ of the bits.

Hence the capacity is at most $1 - \alpha$. It is not immediately obvious that it is possible to achieve this rate. This will follow from Shannon's second theorem.

In many practical channels, the sender receives some feedback from the receiver. If feedback is available for the binary erasure channel, it is very clear what to do: If a bit is lost, retransmit it until it gets through.

Since the bits get through with probability $1 - \alpha$, the effective rate of transmission is $1 - \alpha$. In this way we are easily able to achieve a capacity of $1 - \alpha$ with feedback.

Symmetric Channels

The capacity of the binary symmetric channel is $C = 1 - H(p)$ bits per transmission, and the capacity of the binary erasure channel is $C = 1 - \alpha$ bits per transmission. Now consider the channel with transition matrix:

$$p(y\,|\,x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

Here the entry in the xth row and the yth column denotes the conditional probability $p(y|x)$ that y is received when x is sent.

In this channel, all the rows of the probability transition matrix are permutations of each other and so are the columns. Such a channel is said to be *symmetric*.

Symmetric Channels

Another example of a symmetric channel is one of the form: Y = X + Z (mod c).

Here *Z* has some distribution on the integers {0, 1, 2, . . . , *c* − 1}, *X* has the same alphabet as *Z*, and *Z* is independent of *X*.

In both these cases, we can easily find an explicit expression for the capacity of the channel. Letting **r** be a row of the transition matrix, we have

$$\begin{aligned} & I(X;Y) = H(Y) - H(Y \mid X) \\ & = H(Y) - H(r) \\ & \leq \log |\gamma| - H(r) \end{aligned}$$

with equality if the output distribution is uniform. But *p(x)* = 1/|*X*| achieves a uniform distribution on *Y*, because

$$p(y) = \sum_{x \in \chi} p(y \mid x) p(x) = \frac{1}{|\chi|} \sum p(y \mid x) = c \frac{1}{|\chi|} = \frac{1}{|\gamma|}$$

Symmetric Channels

where *c* is the sum of the entries in one column of the probability transition matrix. Thus, the channel in has the capacity

$$C = \max_{p(x)} I(X;Y) = \log 3 - H(0.5, 0.3, 0.2)$$

and *C* is achieved by a uniform distribution on the input.

The transition matrix of the symmetric channel defined above is doubly stochastic. In the computation of the capacity, we used the facts that the rows were permutations of one another and that all the column sums were equal.

A Condition for Simmetry

Definition A channel is said to be *symmetric* if the rows of the channel transition matrix *p(y|x)* are permutations of each other and the columns are permutations of each other. A channel is said to be *weakly symmetric* if every row of the transition matrix *p(·|x)* is a permutation of every other row and all the column sums Σ_x *p(y|x)* are equal.

For example, the channel with transition matrix is weakly symmetric but not symmetric.

$$p(y\,|\,x) = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

The above derivation for symmetric channels carries over to weakly symmetric channels as well. We have the following theorem for weakly symmetric channels.

Weakly Symmetric Channel

Theorem For a weakly symmetric channel,

C = log |*Y*| − *H(row of transition matrix),*

and this is achieved by a uniform distribution on the input alphabet.

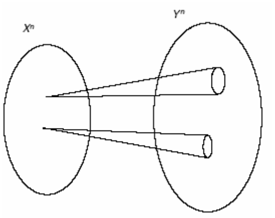
Properties of Channel Capacity

1. $C \geq 0$ since $I(X; Y) \geq 0$.
2. $C \leq \log |X|$ since $C = \max I(X; Y) \leq \max H(X) = \log |X|$.
3. $C \leq \log |Y|$ for the same reason.
4. $I(X; Y)$ is a continuous function of $p(x)$.
5. $I(X; Y)$ is a concave function of $p(x)$

Why C is the Capacity?

If we consider an input sequence of n symbols that we want to transmit over the channel, there are approximately $2^{nH(Y|X)}$ possible Y sequences for each typical n-sequence, all of them equally likely.

We wish to ensure that no two X sequences produce the same Y output sequence. Otherwise, we will not be able to decide which X sequence was sent.



Why C is the Capacity?

The total number of possible (typical) Y sequences is ≈ $2nH(Y)$.

This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences that are the producers.

The total number of disjoint sets is less than or equal to

$2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$

Hence, we can send at most ≈ $2^{nI(X;Y)}$ distinguishable sequences of length n.

In fact, the maximum number of disjoint sets is the maximum number of "independent" output sets.

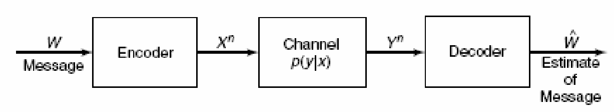
Communication System

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.



Some Definitions

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.

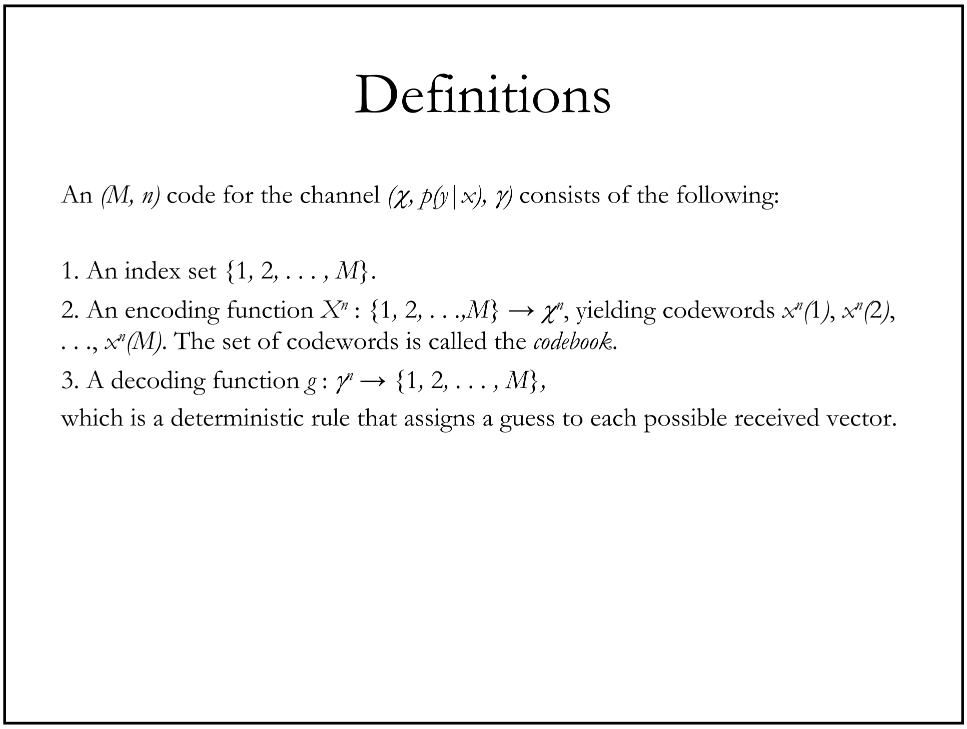
Definition The *nth extension* of the discrete memoryless channel (DMC) is the channel $(\chi^n, p(y^n|x^n), \gamma^n)$,

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

Definitions

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 *n*th extension of the discrete memoryless channel reduces to

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

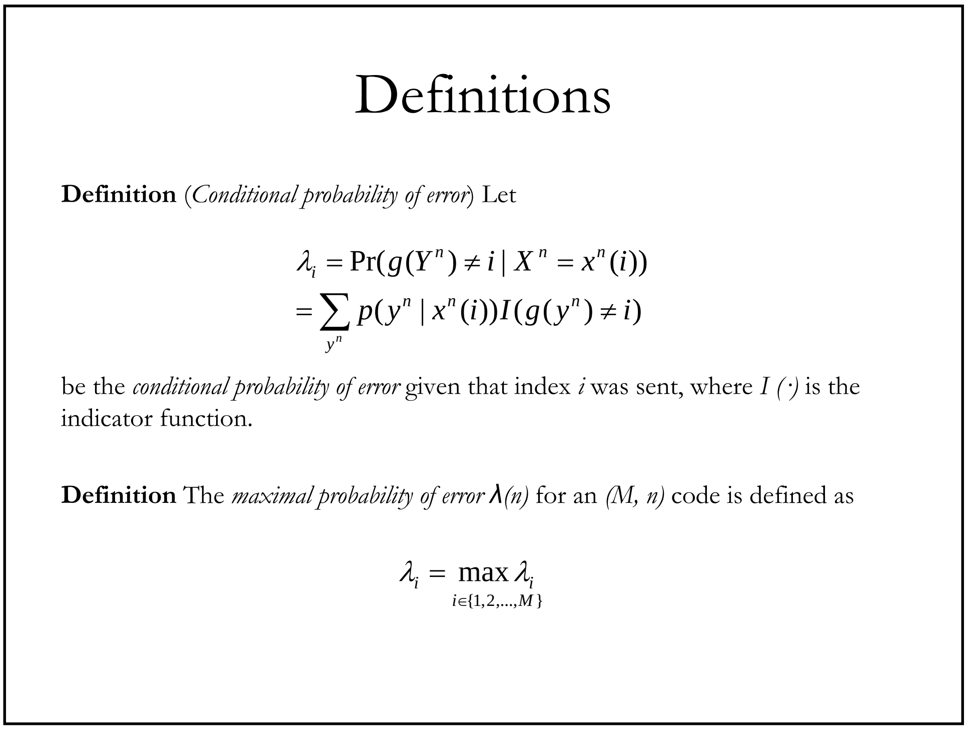


Definitions

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

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

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

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 $\{1, 2, \ldots, M\}$, 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.

Jointly Typical Sequence

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

Definition The set $A_\in^{(n)}$ 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:

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

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. $\Pr((X^n, Y^n) \in A_\in^{(n)}) \to 1$ as $n \to \infty$.
2. $|A_\in^{(n)}| \le 2^{n(H(X,Y)+\in)}$.
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

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

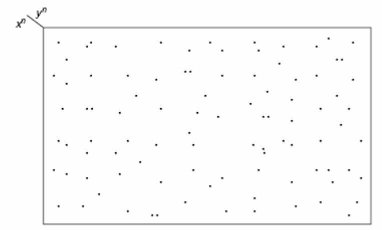
Also, for sufficiently large n:

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

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 $\lambda(n) \rightarrow 0$.

Conversely, any sequence of (2nR, n) codes with $\lambda(n) \rightarrow 0$ 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.