Elements of Information Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

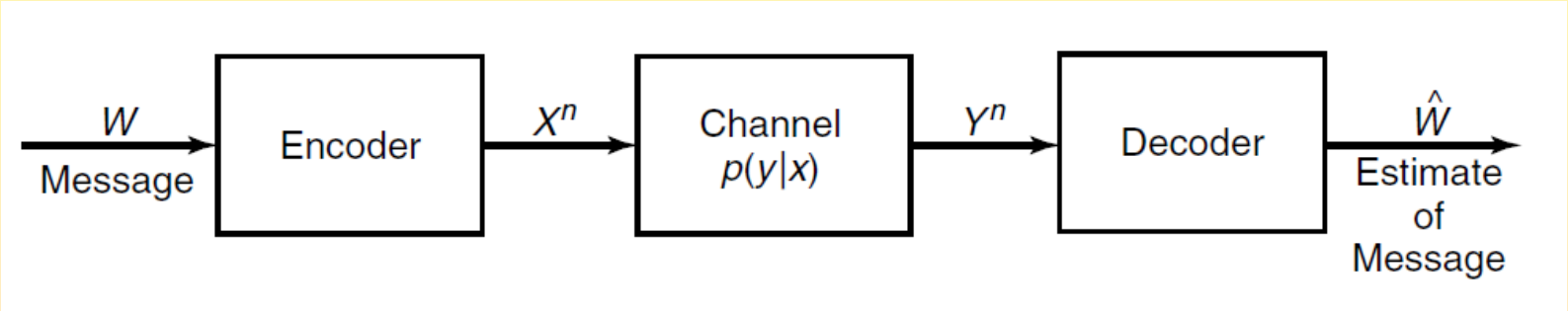
LSSU Math 500

Subsection 1

Channel Capacity

A Signaling System

- The mathematical analog of a physical signaling system:



- Source symbols from some finite alphabet are mapped into some sequence of channel symbols.
- The sequence produces the output sequence of the channel.

  The output sequence is random but has a distribution that depends on the input sequence.
- From the output sequence, we attempt to recover the transmitted message.

Discrete Channel and Channel Capacity

Definition

We define a **discrete channel** to be a system consisting of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ 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.

Definition

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

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

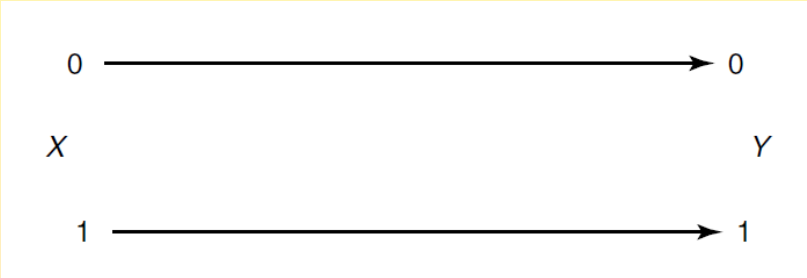
where the maximum is taken over all possible input distributions $p(x)$.

Subsection 2

Examples of Channel Capacity

Noiseless Binary Channel

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



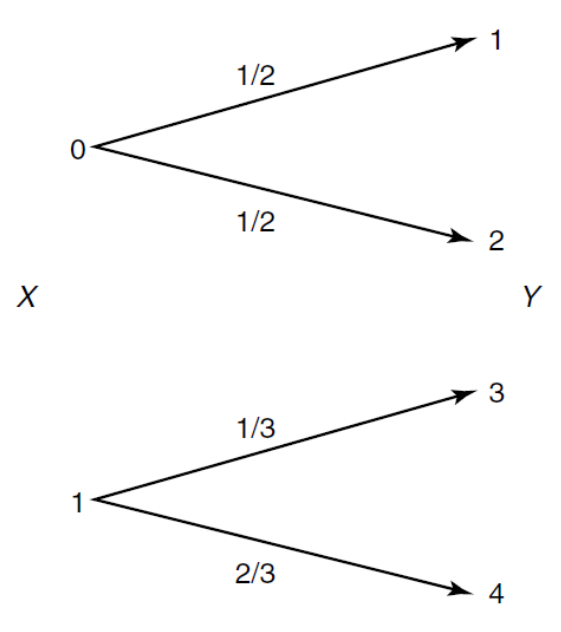
- In this case, any transmitted bit is received without error.
- Hence, one error-free bit can be transmitted per use of the channel.
- It follow that the capacity is 1 bit.
- We can also calculate the information capacity

$$\begin{aligned} C &= \max_{p}(x) I(X;Y) \\ &= \max_{p(x)} \left(p(0,0) \log \frac{p(0|0)}{p(0)} + p(1,1) \log \frac{p(1|1)}{p(1)} \right) = 1 \text{ bit}, \end{aligned}$$

achieved using $p(x) = (\frac{1}{2}, \frac{1}{2})$.

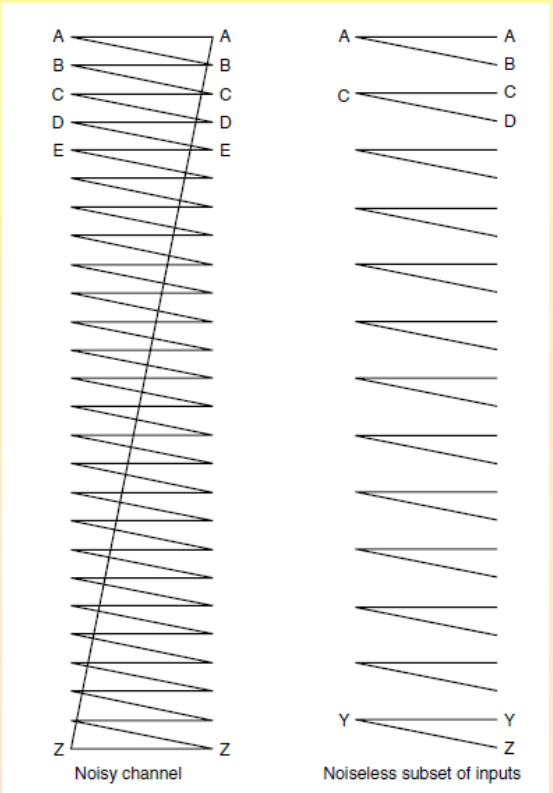
Noisy Channel with Nonoverlapping Outputs

- This channel has two possible outputs corresponding to each of the two inputs.
- The channel appears to be noisy, but really is not.
- The output of the channel is a random consequence of the input, but the input can be determined from the output.



- Hence every transmitted bit can be recovered without error.
- The capacity of this channel is also 1 bit per transmission.
- We can calculate the information capacity $C = \max_{p(x)} I(X;Y) = 1$ bit, achieved using $p(x) = (\frac{1}{2}, \frac{1}{2})$.

Noisy Typewriter



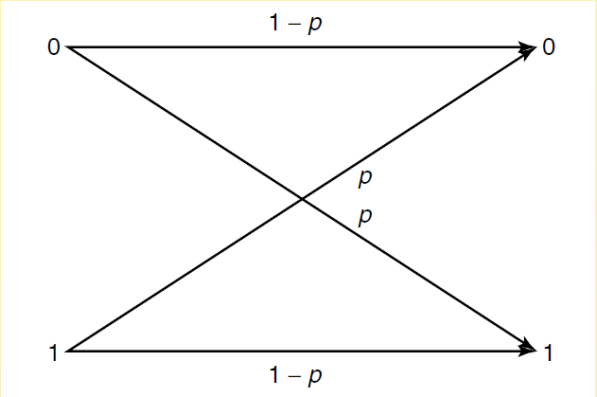
- The channel input is either received unchanged with probability $\frac{1}{2}$ or is transformed into the next letter with probability $\frac{1}{2}$.
- If the input has 26 symbols and we use every alternate input symbol, we can transmit one of 13 symbols error-free per transmission.
- Hence, the capacity of this channel is $\log 13$ bits per transmission.
- We can also calculate the information capacity

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max\left(H(Y) - H(Y|X)\right) \\ &= \max H(Y) - 1 = \log 26 - 1 = \log 13, \end{aligned}$$

achieved using $p(x)$ distributed uniformly over all inputs.

Binary Symmetric Channel

- Consider the following **binary symmetric channel** (**BSC**).
- 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.



- When an error occurs, a 0 is received as a 1, and vice versa.
- The bits received do not reveal where the errors have occurred.
- In a sense, all the bits received are unreliable.
- Later we show that we can still use such a communication channel to send information at a nonzero rate with an arbitrarily small probability of error.

Binary Symmetric Channel (Capacity)

- We bound the mutual information by

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

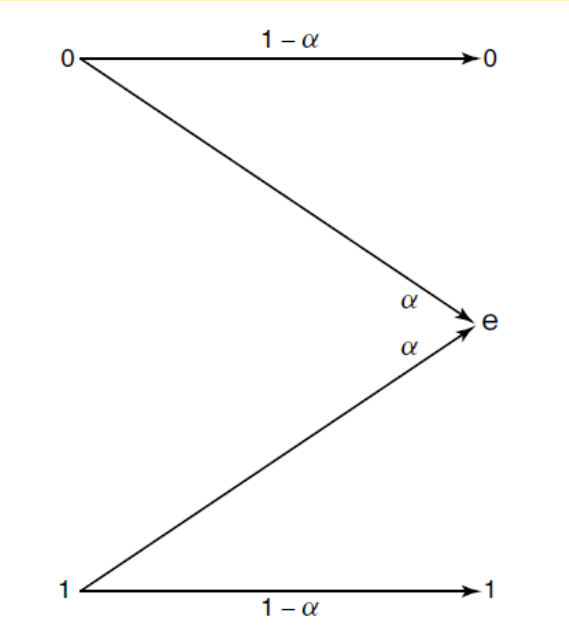
where the inequality follows because Y is a binary random variable.

Equality is achieved when the input distribution is uniform.

Hence, the information capacity of a binary symmetric channel with parameter p is $C = 1 - H(p)$ bits.

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.

Binary Erasure Channel (Capacity)

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

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} H(Y) - H(\alpha). \end{aligned}$$

The first guess for the maximum of $H(Y)$ would be $\log 3$.

But we cannot achieve this by any choice of input distribution $p(x)$.

Binary Erasure Channel (Capacity Cont'd)

- Let $\Pr(X = 1) = \pi$ and E be the event $\{Y = e\}$.
  We have
  $$H(Y) = H(Y, E) = H(E) + H(Y|E).$$
  Moreover,
  $$\begin{aligned} H(Y) &= H((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha)) \\ &= H(\alpha) + (1-\alpha)H(\pi). \end{aligned}$$
  Hence,
  $$\begin{aligned} C &= \max_{p(x)} H(Y) - H(\alpha) \\ &= \max_{\pi} (1-\alpha)H(\pi) + H(\alpha) - H(\alpha) \\ &= \max_{\pi} (1-\alpha)H(\pi) = 1-\alpha, \end{aligned}$$
  where capacity is achieved by $\pi = \frac{1}{2}$.
- The expression for the capacity has some intuitive meaning:
  Since a proportion α of the bits are lost, we can recover (at most) a proportion $1-\alpha$ of the bits. Hence the capacity is at most $1-\alpha$.

Subsection 3

Symmetric Channels

Symmetric Channels

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

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

  The entry in the x-th row and the y-th 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**.

Example

- Another example of a symmetric channel is one of the form

$$Y = X + Z \pmod{c},$$

where:

- Z has some distribution on the integers $\{0, 1, 2, \ldots, c-1\}$;
- X has the same alphabet as Z;
- Z is independent of X.

Example (Capacity)

- We can find an explicit expression for the capacity of the channel.
- Let $\boldsymbol{r}$ be a row of the transition matrix. Then we have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\boldsymbol{r}) \leq \log|\mathcal{Y}| - H(\boldsymbol{r}).$$

Equality holds if the output distribution is uniform.
But $p(x) = \frac{1}{|\mathcal{X}|}$ achieves a uniform distribution on Y.
In fact, let c be the sum of the entries in one column of the matrix.
Then

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|}\sum p(y|x) = c\frac{1}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|}.$$

- Thus, the channel has the capacity

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

C is achieved by a uniform distribution on the input.

Symmetric and Weakly Symmetric Channels

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;
- **Weakly symmetric** if every row of the transition matrix $p(\cdot|x)$ is a permutation of every other row and all the column sums $\sum_x p(y|x)$ are equal.

Example: The channel with transition matrix

$$p(y|x) = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}$$

is weakly symmetric but not symmetric.

Capacity of Weakly Symmetric Channels

- The preceding derivation for symmetric channels carries over to weakly symmetric channels as well.

Theorem

For a weakly symmetric channel,

$$C = \log |\mathcal{Y}| - H(\text{row of transition matrix}),$$

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

Subsection 4

Properties of Channel Capacity

Properties of Channel Capacity

- Recall that
$$C = \max_{p(x)} I(X;Y).$$
- The definition implies several properties of channel capacity.

1. $C \geq 0$.

   Follows from $I(X;Y) \geq 0$.

2. $C \leq \log |\mathcal{X}|$.

   We have
$$C = \max_{p(x)} I(X;Y) \leq \max H(X) = \log |\mathcal{X}|.$$

3. $C \leq \log |\mathcal{Y}|$.

   Same reason as above.

Properties of Channel Capacity (Cont'd)

4. $I(X;Y)$ is a continuous function of $p(x)$.

   Based on the form of the function.

5. $I(X;Y)$ is a concave function of $p(x)$.

   By a previous theorem.

- Since $I(X;Y)$ is a concave function over a closed convex set, a local maximum is a global maximum.

  From Properties 2 and 3, the maximum is finite.

  So we are justified in using the term maximum rather than supremum in the definition of capacity.

Finding Channel Capacity

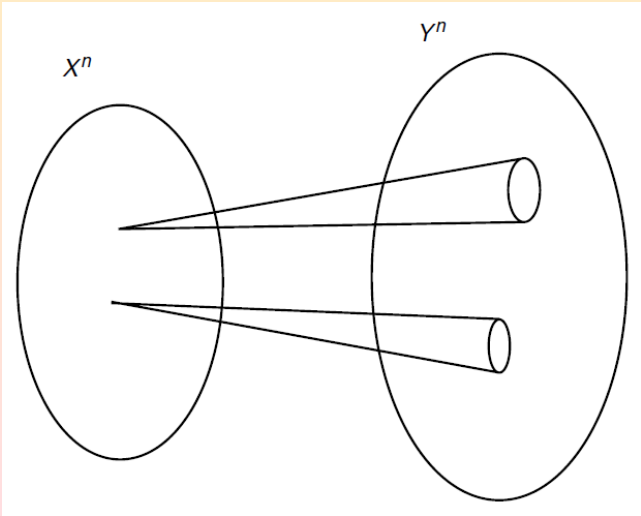
- The maximum can be found by standard nonlinear optimization techniques such as gradient search.
- Some of the methods that can be used include the following:
  - Constrained maximization using calculus and the Kuhn-Tucker conditions;
  - The Frank-Wolfe gradient search algorithm;
  - An iterative algorithm developed by Arimoto and Blahut.
- In general, there is no closed-form solution for the capacity.
- However, for many simple channels, such as the ones considered earlier, it is possible to calculate the capacity using special properties, such as symmetry.

Subsection 5

Preview of the Channel Coding Theorem

Intuition Behind Capacity

- We give an intuitive idea as to why we can transmit C bits of information over a channel.
- The basic idea is that for large block lengths, every channel looks like the noisy typewriter channel.
- Moreover, the channel has a subset of inputs that produce essentially disjoint sequences at the output.



- For each (typical) input n-sequence, there are approximately $2^{nH(Y|X)}$ possible Y sequences, 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.

Intuition Behind Capacity (Cont'd)

- The total number of possible (typical) Y sequences is $\approx 2^{nH(Y)}$.
- This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences.
- The total number of disjoint sets is less than or equal to

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

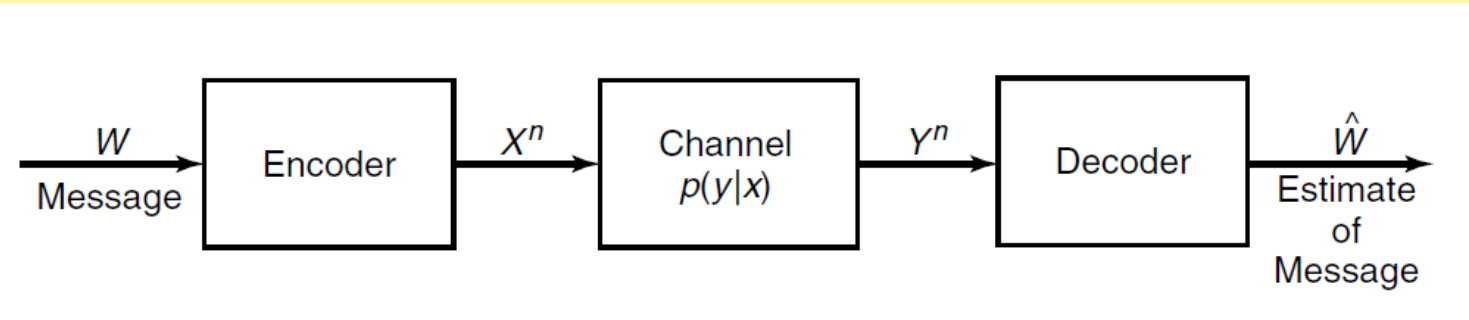
- Hence, we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n.
- This derivation outlines an upper bound on the capacity.
- A stronger version of the argument can be used to prove that this rate I is achievable with an arbitrarily low probability of error.

Subsection 6

Definitions

The Communication System

- We analyze a communication system as shown below.



- A message W, drawn from the index set $\{1, 2, \ldots, M\}$, results in the signal $X^n(W)$.
- This 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 $\widehat{W} = g(Y^n)$.
- The receiver makes an error if $\widehat{W}$ is not the same as the index W that was transmitted.

Discrete Channels and Extensions

Definition

A **discrete channel**, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$, consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability mass functions $p(y|x)$, one for each $x \in X$, such that:

- For every x and y, $p(y|x) \geq 0$;
- For every x, $\sum_y p(y|x) = 1$.

We interpret X as the input and Y as the output of the channel.

Definition

The n-**th extension of the discrete memoryless channel** (DMC) is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where

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

Discrete Memoryless Channel without Feedback

- Suppose the channel is used without feedback, i.e., if the input symbols do not depend on the past output symbols, namely,

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

- Then the channel transition function for the n-th extension of the discrete memoryless channel reduces to

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

- When we refer to the discrete memoryless channel, we mean the discrete memoryless channel without feedback unless we state explicitly otherwise.

Code for a Channel

Definition

An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. 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**.
3. A decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\},$$

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

Conditional and Maximal Probability of Error

Definition (Conditional Probability of Error)

Let

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

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

Definition

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

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

Average Probability of Error

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

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

  i.e., $P_e^{(n)}$ is the probability of error.
- Also, obviously, $P_e^{(n)} \leq \lambda^{(n)}$.

Rates, Achievable Rates and Capacity

Definition

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

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

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 $n \to \infty$.

- Later, 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.

Subsection 7

Jointly Typical Sequences

Jointly Typical Sequences

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 ϵ-close to the true entropies:

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

where

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

Joint AEP Theorem

Theorem (Joint AEP)

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

$$\Pr((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Also, for sufficiently large n,

$$\Pr((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}.$$

Proof of the Joint AEP Theorem (Part 1)

1. We begin by showing that with high probability, the sequence is in the typical set. By the Weak Law of Large Numbers,

$$-\frac{1}{n}\log p(X^n) \to -E[\log p(X)] = H(X) \text{ in probability.}$$

Hence, given $\epsilon > 0$, there exists n_1, such that for all $n > n_1$,

$$\Pr\left(\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \geq \epsilon\right) < \frac{\epsilon}{3}.$$

Similarly, by the Weak Law of Large Numbers, we get

$$-\frac{1}{n}\log p(Y^n) \to -E[\log p(Y)] = H(Y) \text{ in probability}$$

and

$$-\frac{1}{n}\log p(X^n, Y^n) \to -E[\log p(X, Y)] = H(X, Y) \text{ in probability.}$$

Proof of the Joint AEP Theorem (Part 1 Cont'd)

- So there exist n_2 and n_3, such that for all $n \geq n_2$,

$$\Pr\left(\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3}.$$

And there exist n_3, such that for all $n \geq n_3$,

$$\Pr\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3}.$$

Choosing $n > \max\{n_1, n_2, n_3\}$, the probability of the union of the sets above must be less than ϵ.

Hence for n sufficiently large, the probability of the set $A_\epsilon^{(n)}$ is greater than $1 - \epsilon$.

Proof of the Joint AEP Theorem (Part 2)

2. To prove the second part, we have

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

Hence

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}.$$

Proof of the Joint AEP Theorem (Part 3)

3. If $\widetilde{X}^n$ and $\widetilde{Y}^n$ are independent but have the same marginals as X^n and Y^n, then

$$\begin{aligned}\Pr((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \\ &= 2^{-n(I(X;Y)-3\epsilon)}.\end{aligned}$$

For sufficiently large n, $\Pr(A_\epsilon^{(n)}) \geq 1-\epsilon$. Therefore,

$$1-\epsilon \leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n,y^n) \leq |A_\epsilon^{(n)}|2^{-n(H(X,Y)-\epsilon)}.$$

So

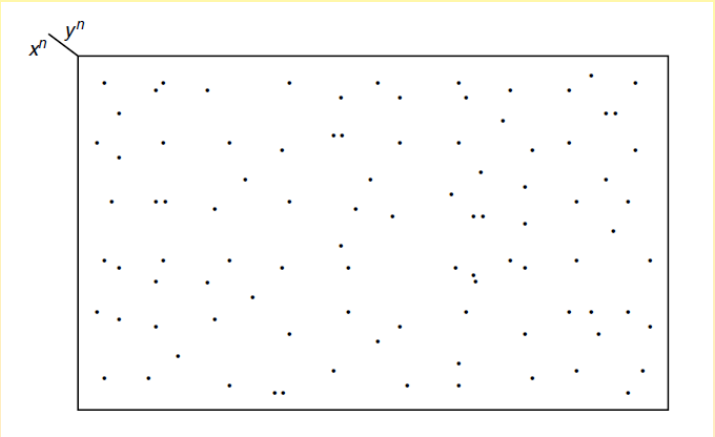
$$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}.$$

Proof of the Joint AEP Theorem (Part 3 Cont'd)

- By similar arguments to the upper bound above, we can also show that for n sufficiently large,

$$\begin{aligned} & Pr((\widetilde{X}^n, \widetilde{Y}^n) \in A_\epsilon^{(n)}) \\ & \quad = \sum_{A_\epsilon^{(n)}} p(x^n)p(y^n) \\ & \quad \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} \\ & \quad = (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}. \end{aligned}$$

Illustration and Comments



- The jointly typical set is illustrated in the figure.
- There are:
  - About $2^{nH(X)}$ typical X sequences;
  - About $2^{nH(Y)}$ typical Y sequences.
- 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 $2^{-nI(X;Y)}$.
- Hence, we can consider about $2^{nI(X;Y)}$ such pairs before we are likely to come across a jointly typical pair.
- This suggests that there are about $2^{nI(X;Y)}$ distinguishable signals X^n.

The Fixed Output Point of View

- Consider 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

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

- 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.

Subsection 8

Channel Coding Theorem

Channel Coding Theorem: Ideas of Proof

- Shannon used a number of ideas to prove that information can be sent reliably over a channel at all rates up to the channel capacity.
- These ideas include:
  - Allowing an arbitrarily small but nonzero probability of error;
  - Using the channel many times in succession, so that the law of large numbers comes into effect;
  - Calculating the average of the probability of error over a random choice of codebooks, which symmetrizes the probability, and which can then be used to show the existence of at least one good code.

Channel Coding Theorem: Comments on the Proof

- As in all the proofs, we use the same essential ideas:
  - Random code selection;
  - Calculation of the average probability of error for a random choice of codewords;
  - ⋮
- The main difference is in the 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: Comments on the Proof (Cont'd)

- 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.
- Although jointly typical decoding is suboptimal, it is simple to analyze and still achieves all rates below capacity.

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 $R \leq C$.

- For now, we show 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 $p(x) = \prod_{i=1}^{n} p(x_i)$.
  We exhibit the 2^{nR} codewords as the rows of a matrix:

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

Channel Coding Theorem (Process)

- Each entry in this matrix is generated i.i.d. according to $p(x)$.

  Thus, the probability that we generate a particular code $\mathcal{C}$ is

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

  Consider the following sequence of events:

  1. A random code $\mathcal{C}$ is generated as described above according to $p(x)$.
  2. The code $\mathcal{C}$ 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

$$\Pr(W = w) = 2^{-nR}, \quad w = 1, 2, \ldots, 2^{nR}.$$

  4. The w-th codeword $X^n(w)$, corresponding to the w-th row of $\mathcal{C}$, is sent over the channel.

Channel Coding Theorem (Process Cont'd)

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

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

6. The receiver guesses which message was sent.
   It uses jointly typical decoding, declaring that the index $\widehat{W}$ was sent, if the following conditions are satisfied:
   - $(X^n(\widehat{W}), Y^n)$ is jointly typical;
   - There is no other index $W' \neq \widehat{W}$, such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

   If no such $\widehat{W}$ exists or if there is more than one such, an error is declared.
7. There is a decoding error if $\widehat{W} \neq W$.
   Let $\mathcal{E}$ be the event $\{\widehat{W} \neq W\}$.

Analysis of the Probability of Error: Outline

- Instead of calculating the probability of error for a single code, we calculate the average over all codes generated at random according to the chosen distribution.
- By the symmetry of the code construction, the average probability of error does not depend on the particular index that was sent.
- For a typical codeword, there are two different sources of error when we use jointly typical decoding:
  - The output Y^n is not jointly typical with the transmitted codeword;
  - 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, as shown by the joint 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.

Detailed Calculation of the Probability of Error

- We let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$ and use jointly typical decoding $\widehat{W}(y^n)$.

  Let $\mathcal{E} = \{\widehat{W}(Y^n) \neq W\}$ denote the error event.

  We will calculate the average probability of error, averaged over all codewords in the codebook, and averaged over all codebooks:

$$\begin{aligned} \Pr(\mathcal{E}) &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\ &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}). \end{aligned}$$

  Here, $P_e^{(n)}(\mathcal{C})$ is defined for jointly typical decoding.

Detailed Calculation (Cont'd)

- By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index that was sent, i.e., $\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C})$ does not depend on w.

  Now we have

$$\begin{aligned} \Pr(\mathcal{E}) &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) \\ &= \sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_1(\mathcal{C}) \\ &= \Pr(\mathcal{E}|W=1). \end{aligned}$$

  Thus, without loss of generality, assume that $W = 1$ was sent.

  Y_n is the result of sending the first codeword $X^n(1)$ over the channel.

  Define the following events:

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

  E_i is the event that the i-th codeword and Y^n are jointly typical.

Detailed Calculation (Cont'd)

- An error occurs in the decoding scheme if one of the following happens.
  - E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical);
  - $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (a wrong codeword is jointly typical with the received sequence).

  Hence, letting $P(\mathcal{E})$ denote $\Pr(\mathcal{E}|W=1)$ and using the union bound,

$$\begin{aligned} \Pr(\mathcal{E}|W=1) &= P(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W=1) \\ &\leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1). \end{aligned}$$

  By the joint AEP, $P(E_1^c|W=1) \to 0$.

  Hence, $P(E_1^c|W=1) \leq \epsilon$, for n sufficiently large.

Detailed Calculation (Cont'd)

- Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, so are Y^n and $X^n(i)$.

  Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP.

  Consequently, if n is sufficiently large and $R < I(X;Y) - 3\epsilon$,

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

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

Detailed Calculation (Selections)

- We strengthen this conclusion by a series of code selections.
  1. Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieves capacity. Then the condition $R < I(X;Y)$ can be replaced by the achievability condition $R < C$.
  2. Get rid of the average over codebooks. Since the average probability of error over codebooks is small ($\leq 2\epsilon$), there exists at least one codebook $\mathcal{C}^*$ with a small average probability of error. Thus, $\Pr(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon$. Determination of $\mathcal{C}^*$ can be achieved by an exhaustive search over all $(2^{nR}, n)$ codes. Note that, since we have chosen W according to a uniform distribution,

$$\Pr(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*).$$

Detailed Calculation (Selections Cont'd)

3. Throw away the worst half of the codewords in the best codebook $\mathcal{C}^*$.
   Since the arithmetic average probability of error $P_e^{(n)}(\mathcal{C}^*)$ for this code is less then 2ϵ, we have $\Pr(\mathcal{E}|\mathcal{C}^*) \leq \frac{1}{2^{nR}} \sum \lambda_i(\mathcal{C}^*) \leq 2\epsilon$.
   Thus, at least half the indices i and their associated codewords $X^n(i)$ must have conditional probability of error λ_i less than 4ϵ (otherwise, these codewords would contribute more than 2ϵ to the sum).
   Hence, the best half of the codewords have a maximal probability of error less than 4ϵ.
   If we reindex these codewords, we have 2^{nR-1} codewords.
   Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n.

- Combining all these improvements, we have constructed a code of rate $R' = R - \frac{1}{n}$, with maximal probability of error $\lambda^{(n)} \leq 4\epsilon$.

  This proves the achievability of any rate below capacity.

Subsection 9

Zero-Error Codes

Zero-Error Codes

- We prove that $P_e^{(n)} = 0$ implies that $R \leq C$.

  Assume that we have a $(2^{nR}, n)$ code with zero probability of error.

  I.e., the decoder output $g(Y^n)$ is equal to the input index W with probability 1.

  Then the input index W is determined by the output sequence, i.e.,

$$H(W|Y^n) = 0.$$

  To obtain a strong bound, we arbitrarily assume that W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

  Thus, $H(W) = nR$.

Zero-Error Codes (Cont'd)

- We can now write the string of inequalities:

$$\begin{aligned} nR &= H(W) \\ &= H(W|Y^n) + I(W;Y^n) \\ &= I(W;Y^n) \\ &\overset{W \to X^n(W) \to Y^n}{\leq} I(X^n;Y^n) \\ &\overset{*}{\leq} \sum_{i=1}^{n} I(X_i;Y_i) \\ &\overset{C}{\leq} nC. \end{aligned}$$

Inequality $*$ will be proved in the next section.

Hence, for all n, for any zero-error $(2^{nR}, n)$ code, $R \leq C$.

Subsection 10

Fano's Inequality and the Converse to the Coding Theorem

The Setup

- The index W is uniformly distributed on the set $W = \{1, 2, \ldots, 2^{nR}\}$.
- The sequence Y^n is related probabilistically to W.
- From Y^n, we estimate the index W that was sent, and the estimate is $\widehat{W} = g(Y^n)$.
- Thus, $W \to X^n(W) \to Y^n \to \widehat{W}$ forms a Markov chain.
- The probability of error is

$$\Pr(\widehat{W} \neq W) = \frac{1}{2^{nR}} \sum_i \lambda_i = P_e^{(n)}.$$

Fano's Inequality

Lemma (Fano's Inequality)

For a discrete memoryless channel with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have

$$H(W|\widehat{W}) \leq 1 + P_e^{(n)} nR.$$

- Since W is uniformly distributed, we have $P_e^{(n)} = \Pr(W \neq \widehat{W})$. We apply the weak version of Fano's inequality

$$H(X|\widehat{X}) \leq 1 + \Pr(X \neq \widehat{X}) \log |\mathcal{X}|,$$

for W in an alphabet of size 2^{nR}, to get

$$H(W|\widehat{W}) \leq 1 + P_e^{(n)} nR.$$

Capacity Per Transmission

Lemma

Let Y^n be the result of passing X^n through a discrete memoryless channel of capacity C. Then

$$I(X^n; Y^n) \leq nC, \quad \text{for all } p(x^n).$$

- By the definition of a discrete memoryless channel, Y_i depends only on X_i and is conditionally independent of everything else. Hence

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i|X_i) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) \\ &= \sum_{i=1}^n I(X_i; Y_i) \\ &\leq nC. \end{aligned}$$

Converse to the Channel Coding Theorem

- We have to show that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.
  If the maximal probability of error tends to zero, the average probability of error for the sequence of codes also goes to zero.
  That is, $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$.
  For a fixed encoding rule $X^n(\cdot)$ and a fixed decoding rule $\widehat{W} = g(Y^n)$, we have $W \to X^n(W) \to Y^n \to \widehat{W}$.
  For each n, let W be drawn uniformly over $\{1, 2, \ldots, 2^{nR}\}$.
  Then $\Pr(\widehat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \sum_i \lambda_i$. So

$$\begin{aligned} nR &\overset{\text{uniform } W}{=} H(W) \overset{\text{identity}}{=} H(W|\widehat{W}) + I(W;\widehat{W}) \\ &\overset{\text{Fano}}{\leq} 1 + P_e^{(n)} nR + I(W;\widehat{W}) \\ &\overset{\text{data-proc.}}{\leq} 1 + P_e^{(n)} nR + I(X^n;Y^n) \\ &\overset{\text{Lemma}}{\leq} 1 + P_e^{(n)} nR + nC. \end{aligned}$$

Converse to the Channel Coding Theorem (Cont'd)

- Dividing by n, we obtain $R \leq P_e^{(n)} R + \frac{1}{n} + C$.

  Letting $n \to \infty$ ($P_e^{(n)} \to 0, \frac{1}{n} \to 0$), we get $R \leq C$.

  Rewriting $P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$ we see that, if $R > C$, the probability of error is bounded away from 0 for sufficiently large n.

  The same holds for all n, since, if $P_e^{(n)} = 0$ for small n, we can construct codes for large n with $P_e^{(n)} = 0$ by concatenating these codes.

  Hence, we cannot achieve an arbitrarily low probability of error at rates above capacity.
- This converse is sometimes called the **weak converse** to the channel coding theorem.
- It is also possible to prove a **strong converse**, which states that for rates above capacity, the probability of error goes exponentially to 1.

Subsection 11

Equality in the Converse to the Channel Coding Theorem

Equality in the Channel Coding (Inequalities)

- We examine the consequences of equality in the converse, which gives some ideas as to the kinds of codes that achieve capacity.

  Repeating the steps of the converse in the case when $P_e = 0$, we have

$$\begin{aligned}
nR &= H(W) \\
&= H(W|\widehat{W}) + I(W;\widehat{W}) \\
&= I(W;\widehat{W}) \\
&\overset{(a)}{\leq} I(X^n(W);Y^n) \\
&= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \textstyle\sum_{i=1}^{n}(Y_i|X_i) \\
&\overset{(b)}{\leq} \textstyle\sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= \textstyle\sum_{i=1}^{n} I(X_i;Y_i) \\
&\overset{(c)}{\leq} nC.
\end{aligned}$$

Equality in the Channel Coding (Equalities)

- We analyze inequalities (a), (b) and (c).
  - We have equality in the data-processing inequality (a), only if

$$I(Y^n; X^n(W)|W) = 0 \quad \text{and} \quad I(X^n; Y^n|\widehat{W}) = 0.$$

    This is true if all the codewords are distinct and if $\widehat{W}$ is a sufficient statistic for decoding.
  - We have equality in inequality (b) only if the Y_i's are independent.
  - We have equality in the inequality (c) only if the distribution of X_i is $p^*(x)$, the distribution on X that achieves capacity.

We have equality in the converse only if these conditions are satisfied.

Conclusions

- This indicates that a capacity-achieving zero-error code has distinct codewords and the distribution of the Y_i's must be i.i.d. with

$$p^*(y) = \sum_x p^*(x)p(y|x),$$

  the distribution on Y induced by the optimum distribution on X.
- The distribution referred to in the converse is the empirical distribution on X and Y induced by a uniform distribution over codewords, that is,

$$p(x_i, y_i) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} I(X_i(w) = x_i)p(y_i|x_i).$$

Subsection 12

Hamming Codes

Redundancy

- The object of coding is to introduce redundancy so that, even if some of the information is lost or corrupted, it will still be possible to recover the message at the receiver.
- The most obvious coding scheme is to repeat information.
  Example: To send a 1, we send 11111, and to send a 0, we send 00000.
  - This scheme uses five symbols to send 1 bit, and therefore has a rate of $\frac{1}{5}$ bit per symbol.
  - If this code is used on a binary symmetric channel, the optimum decoding scheme is to take the majority vote of each block of five received bits.
  - An error occurs if and only if more than three of the bits are changed.
- By using longer repetition codes, we can achieve an arbitrarily low probability of error.
- But the rate of the code also goes to zero with block length.
- So, even though the code is "simple", it is not a very useful code.

Error-Detection

- Instead of simply repeating the bits, we can combine the bits in some intelligent fashion so that each extra bit checks whether there is an error in some subset of the information bits.

  Example: In a parity check code, starting with a block of $n-1$ information bits, we choose the n-th bit so that the parity of the entire block is 0 (the number of 1's in the block is even).

  Then if there is an odd number of errors during the transmission, the receiver will notice that the parity has changed and detect the error.

- This is the simplest example of an **error-detecting code**.

  Example (Cont'd): The preceding code has limitations.

  - It does not detect an even number of errors;
  - It does not give any information about how to correct the errors that occur.

Example

- Consider a binary code of block length 7.

  All operations will be done modulo 2.

  Consider the set of all nonzero binary vectors of length 3.

  Arrange them in columns to form a matrix:

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

  Consider the set of vectors of length 7 in the null space of H.

  These are the vectors which when multiplied by H give 000.

  From the theory of linear spaces, since H has rank 3, we expect the null space of H to have dimension 4.

Example (Cont'd)

- The 2^4 codewords in the null space of H are

0000000	0100101	1000011	1100110
0001111	0101010	1001100	1101001
0010110	0110011	1010101	1110000
0011001	0111100	1011010	1111111

Since the set of codewords is the null space of a matrix, it is linear in the sense that the sum of any two codewords is also a codeword.

The set of codewords therefore forms a linear subspace of dimension 4 in the vector space of dimension 7.

Example (Minimum Weight)

- Looking at the codewords, we notice that other than the all-0 codeword, the minimum number of 1's in any codeword is 3.

  This is called the **minimum weight** of the code.

  The following reasoning shows that the minimum weight of this code has to be exactly 3.
  - Since all columns of H are different, no two columns can add to 000. So the minimum weight of a code has to be at least 3.
  - The sum of any two columns must be one of the columns of the matrix. So the minimum distance is exactly 3.

Example (Minimum Distance)

- The difference between any two codewords is also a codeword.

  Hence, any two codewords differ in at least three places.
- The minimum number of places in which two codewords differ is called the **minimum distance** of the code.
- The minimum distance of the code is a measure of how far apart the codewords are and will determine how distinguishable the codewords will be at the output of the channel.
- The minimum distance is equal to the minimum weight for a linear code.
- We aim to develop codes that have a large minimum distance.

Example (Decoding)

- We saw that the minimum distance is 3.
- So, if a codeword $\boldsymbol{c}$ is corrupted in only one place, it will differ from any other codeword in at least two places.
- It follows that it will be closer to $\boldsymbol{c}$ than to any other codeword.
- We can discover which is the closest codeword without searching over all the codewords by using the structure of the matrix H for decoding.

Example (Decoding Using H)

- The matrix H, called the **parity check matrix**, has the property that for every codeword $\boldsymbol{c}$,
$$H\boldsymbol{c} = 0.$$
- Let $\boldsymbol{e}_i$ be a vector with a 1 in the i-th position and 0's elsewhere.
- If the codeword is corrupted at position i, the received vector
$$\boldsymbol{r} = \boldsymbol{c} + \boldsymbol{e}_i.$$
- If we multiply this vector by the matrix H, we obtain
$$H\boldsymbol{r} = H(\boldsymbol{c} + \boldsymbol{e}_i) = H\boldsymbol{c} + H\boldsymbol{e}_i = H\boldsymbol{e}_i.$$
- This is the vector corresponding to the i-th column of H.
- So the product $H\boldsymbol{r}$ reveals which position of the vector was corrupted.
- Reversing this bit will give us a codeword.
- This codebook with 16 codewords of block length 7, which can correct up to one error, is called a **Hamming code**.

Example (Encoding Procedure)

- For encoding, we could use any mapping from a set of 16 messages into the codewords.
- But if we examine the first 4 bits of the codewords in the table, we observe that they cycle through all 2^4 combinations of 4 bits.
- Thus, we could use these 4 bits to be the 4 bits of the message we want to send.
- The remaining 3 bits are then determined by the code.
- In general, it is possible to modify a linear code so that the mapping is explicit:
  - The first k bits in each codeword represent the message;
  - The last $n - k$ bits are parity check bits.
- Such a code is called a **systematic code**.

Identifying Features

- The code is often identified by:
  - Its block length n;
  - The number of information bits k;
  - The minimum distance d.

Example: We revisit the code in the example.

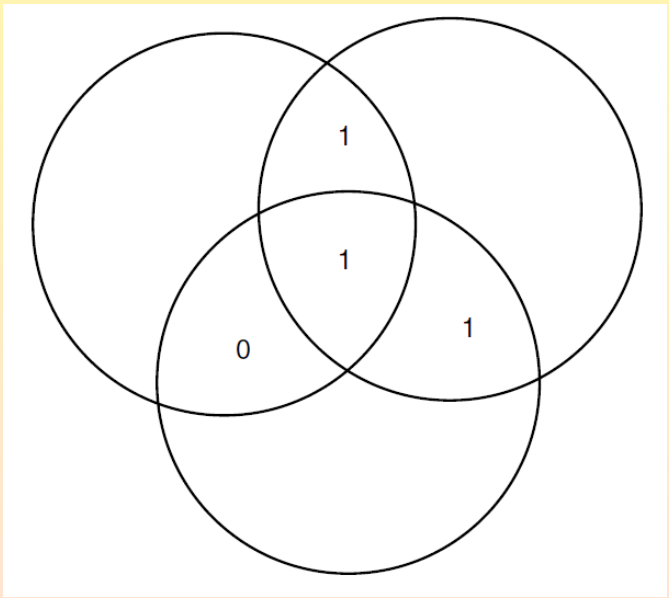
It is called a $(7, 4, 3)$ Hamming code.

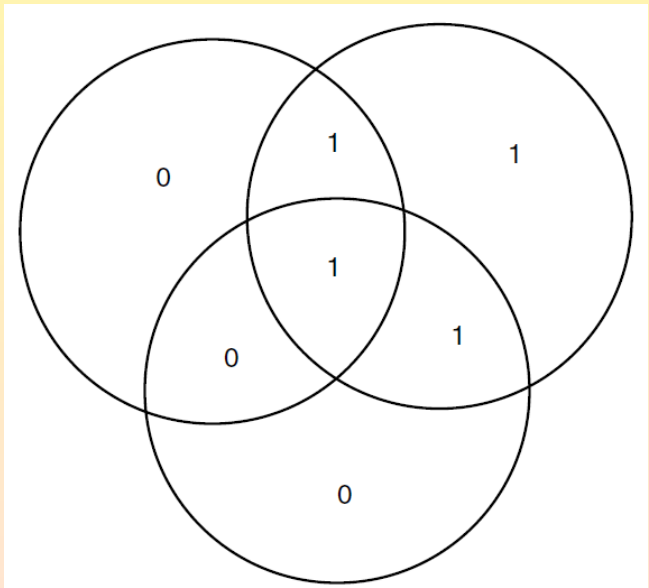
This terminology refers to the parameters

$$n = 7, \quad k = 4, \quad d = 3.$$

Hamming Codes using Venn Diagrams

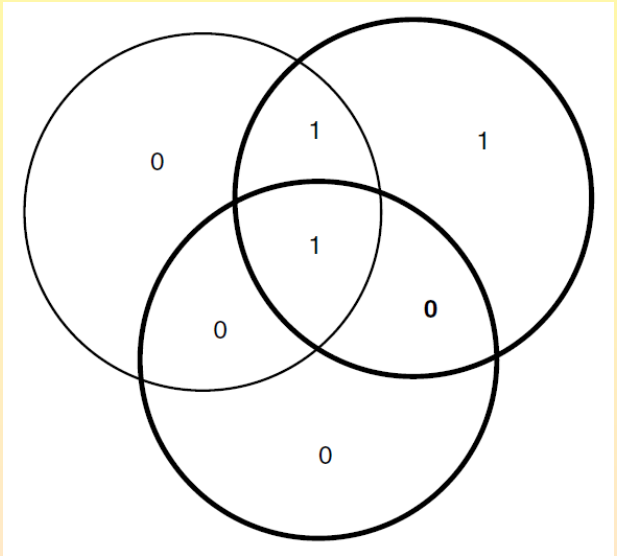
- Consider the Venn diagram on the left with three circles and with four intersection regions.





- To send the information sequence 1101, we place the 4 information bits in the four intersection regions as shown.
- We then place a parity bit in each of the three remaining regions so that the parity of each circle is even, i.e., there are an even number of 1's in each circle.

Hamming Codes using Venn Diagrams (Cont'd)



- Now assume that one of the bits is changed, say from 1 to 0 as shown in the figure.
- Then the parity constraints are violated for two of the circles (highlighted in the figure).
- It is not hard to see that given these violations, the only single bit error that could have caused it is at the intersection of the two circles (i.e., the bit that was changed).
- Similarly working through the other error cases, it is not hard to see that this code can detect and correct any single bit error in the received codeword.

Generalization

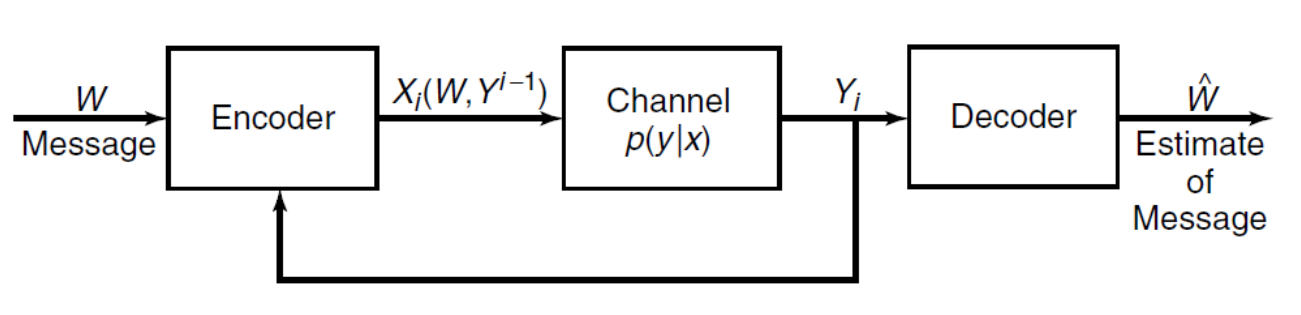
- We can generalize this procedure to construct larger matrices H.
- In general, if we use ℓ rows in H, the code that we obtain will have:
  - Block length $n = 2^\ell - 1$;
  - $k = 2^\ell - \ell - 1$;
  - Minimum distance 3.
- All these codes are called Hamming codes.
- They can correct one error.

Subsection 13

Feedback Capacity

Channel with Feedback

- A channel with feedback is illustrated below.



- We assume that all the received symbols are sent back immediately and noiselessly to the transmitter.
- The transmitter can use them to decide which symbol to send next.
- We define a $(2^{nR}, n)$ **feedback code** as:
  - A sequence of mappings $x_i(W, Y^{i-1})$, where each x_i is a function only of the message $W \in 2^{nR}$ and the previously received $Y_1, Y_2, \ldots, Y_{i-1}$;
  - A sequence of decoding functions $g : \mathcal{Y}^n \to \{1, 2, \ldots, 2^{nR}\}$.
- We have $P_e^{(n)} = \Pr\{g(Y^n) \neq W\}$, when W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

Feedback Capacity

Definition

The **capacity with feedback**, C_{FB}, of a discrete memoryless channel is the supremum of all rates achievable by feedback codes.

Theorem (Feedback Capacity)

We have

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

- A nonfeedback code is a special case of a feedback code.

  So any rate that can be achieved without feedback can certainly be achieved with feedback.

  Therefore, $C_{\mathrm{FB}} \geq C$.

Feedback Capacity (Converse)

- Let W be uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

  Then

  $$\Pr(W \neq \widehat{W}) = P_e^{(n)}.$$

  Moreover,

  $$\begin{aligned} nR &= H(W) \\ &= H(W|\widehat{W}) + I(W;\widehat{W}) \\ &\overset{\text{Fano}}{\leq} 1 + P_e^{(n)} nR + I(W;\widehat{W}) \\ &\overset{\text{data-proc}}{\leq} 1 + P_e^{(n)} nR + I(W;Y^n). \end{aligned}$$

  Now it suffices to bound $I(W;Y^n)$.

Feedback Capacity (Bounding $I(W; Y^n)$)

$$
\begin{aligned}
I(W; Y^n) &= H(Y^n) - H(Y^n|W) \\
&= H(Y^n) - \textstyle\sum_{i=1}^n H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W) \\
&= H(Y^n) - \textstyle\sum_{i=1}^n H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W, X_i) \\
&\quad (X_i \text{ is a function of } Y_1, \ldots, Y_{i-1} \text{ and } W) \\
&= H(Y^n) - \textstyle\sum_{i=1}^n H(Y_i|X_i) \\
&\quad (X_i \text{ conditional on } X_i, Y_i \text{ is independent} \\
&\quad \text{of } W \text{ and past samples of } Y) \\
&\leq \textstyle\sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) \\
&= \textstyle\sum_{i=1}^n I(X_i; Y_i) \\
&\leq nC \ (C \text{ for a discrete memoryless channel}).
\end{aligned}
$$

Putting these together, we obtain $nR \leq P_e^{(n)} nR + 1 + nC$.

Dividing by n and letting $n \to \infty$, we conclude that $R \leq C$.

So we cannot achieve any higher rates with feedback than we can without feedback, and $C_{\text{FB}} = C$.

Subsection 14

Source-Channel Separation Theorem

Setup

- We have a source V that generates symbols from an alphabet $\mathcal{V}$.
- The process V is assumed to be of a finite alphabet and to satisfy the AEP.
- We want to send the sequence of symbols $V^n = V_1, V_2, \ldots, V_n$ over the channel so that the receiver can reconstruct the sequence.
- To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel.
- The receiver looks at his received sequence Y^n and makes an estimate $\widehat{V}^n$ of the sequence V^n that was sent.
- The receiver makes an error if $V^n \neq \widehat{V}^n$, with probability

$$\Pr(V^n \neq \widehat{V}^n) = \sum_{y^n} \sum_{v^n} p(v^n) p(y^n | x^n(v^n)) I(g(y^n) \neq v^n),$$

where I is the indicator function and $g(y^n)$ is the decoding function.

Source-Channel Coding Theorem

Theorem (Source-Channel Coding Theorem)

If $V_1, V_2, \ldots, V_n$ is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source-channel code with probability of error $\Pr(\widehat{V}^n \neq V^n) \to 0$.
Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.

- **Achievability** We rely on a two-stage encoding.
  By hypothesis, the stochastic process satisfies the AEP.
  So there exists a typical set $A_\epsilon^{(n)}$ of size $\leq 2^{n(H(\mathcal{V})+\epsilon)}$ which contains most of the probability.
  We will encode only the source sequences belonging to the typical set.
  All other sequences will result in an error.
  This will contribute at most ϵ to the probability of error.

Source-Channel Coding Theorem (Achievability)

- We index all the sequences belonging to $A_\epsilon^{(n)}$.
  There are at most $2^{n(H+\epsilon)}$ such sequences.
  Thus, $n(H+\epsilon)$ bits suffice to index them.
  We can transmit the desired index to the receiver with probability of error less than ϵ if $H(\mathcal{V}) + \epsilon = R < C$.
  The receiver can reconstruct V^n by enumerating the typical set $A_\epsilon^{(n)}$ and choosing the sequence corresponding to the estimated index.
  It will agree with the transmitted sequence with high probability.
  To be precise, for n sufficiently large,

$$\begin{aligned} P(V^n \neq \widehat{V}^n) &\leq P(V^n \notin A_\epsilon^{(n)}) + P(g(Y^n) \neq V^n | V^n \in A_\epsilon^{(n)}) \\ &\leq \epsilon + \epsilon \\ &= 2\epsilon. \end{aligned}$$

  Hence, we can reconstruct the sequence with low probability of error for n sufficiently large, if $H(\mathcal{V}) < C$.

Source-Channel Coding Theorem (Converse)

- **Converse** We wish to show that $\Pr(\widehat{V}^n \neq V^n) \to 0$ implies that $H(\mathcal{V}) \leq C$, for any sequence of source-channel codes

$$X^n(V^n) : \mathcal{V}^n \to \mathcal{X}^n, \quad g_n(Y^n) : \mathcal{Y}^n \to \mathcal{V}^n.$$

Thus $X^n(\cdot)$ is an arbitrary (perhaps random) assignment of codewords to data sequences V^n, and $g_n(\cdot)$ is any decoding function.

By Fano's inequality, we must have

$$\begin{aligned} H(V^n|\widehat{V}^n) &\leq 1 + \Pr(\widehat{V}^n \neq V^n) \log |\mathcal{V}^n| \\ &= 1 + \Pr(\widehat{V}^n \neq V^n) n \log |\mathcal{V}|. \end{aligned}$$

Source-Channel Coding Theorem (Converse Cont'd)

- Hence for the code,

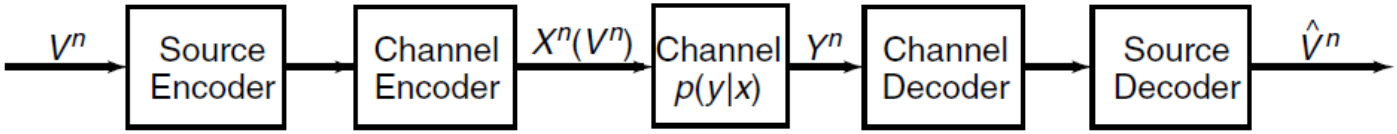
$$
\begin{aligned}
H(\mathcal{V}) &\overset{\text{entropy}}{\leq} \frac{H(V_1, V_2, \ldots, V_n)}{n} \\
&= \frac{H(V^n)}{n} \\
&= \frac{1}{n} H(V^n|\widehat{V}^n) + \frac{1}{n} I(V^n; \widehat{V}^n) \\
&\overset{\text{Fano}}{\leq} \frac{1}{n}(1 + \Pr(\widehat{V}^n \neq V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(V^n; \widehat{V}^n) \\
&\overset{\text{data-proc}}{\leq} \frac{1}{n}(1 + \Pr(\widehat{V}^n \neq V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(X^n; Y^n) \\
&\overset{\text{memoryless}}{\leq} \frac{1}{n} + \Pr(\widehat{V}^n \neq V^n) \log |\mathcal{V}| + C.
\end{aligned}
$$

Letting $n \to \infty$, we have $\Pr(\widehat{V}^n \neq V^n) \to 0$.

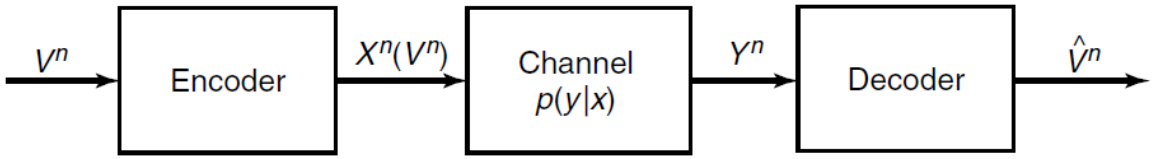
Hence, $H(\mathcal{V}) \leq C$.

Problems of Source and Channel Coding

- Hence, we can transmit a stationary ergodic source over a channel if and only if its entropy rate is less than the capacity of the channel.
- The joint source-channel separation theorem enables us to consider the problems of source and channel coding separately.
  - The source coder tries to find the most efficient representation of the source;
  - The channel coder encodes the message to combat the noise and errors introduced by the channel.
- The separation theorem says that the separate encoders



can achieve the same rates as the joint encoder



Data Compression and Data Transmission

- This result ties together the two basic theorems of Information Theory, Data Compression and Data Transmission.
  - The Data Compression Theorem is a consequence of the AEP, which shows that there exists a "small" subset (of size 2^{nH}) of all possible source sequences that contain most of the probability.
    Consequently, we can represent the source, with a small probability of error, using H bits per symbol.
  - The Data Transmission theorem is based on the joint AEP.
    It uses the fact that for long block lengths:
    - The output sequence of the channel is very likely to be jointly typical with the input codeword;
    - Any other codeword is jointly typical with probability $\approx 2^{-nI}$.

    Consequently, we can use about 2^{nI} codewords and still have negligible probability of error.
- The Source-Channel Separation Theorem shows that we can design the source code and the channel code separately and combine the results to achieve optimal performance.