Coding and Information Theory
Chapter 5
Using an Unreliable Channel

Xuejun Liang

2019 Fall

Chapter 5
Using an Unreliable Channel

1. Decision Rules
2. An Example of Improved Reliability
3. Hamming Distance
4. Statement and Outline Proof of Shannon's Theorem
5. The Converse of Shannon's Theorem
6. Comments on Shannon's Theorem

The aim of this chapter

- Shannon's Fundamental Theorem states that
  - the capacity C of Γ is the least upper bound for the rates at which one can transmit information accurately through Γ.
- We will look at a simple example of how this accurate transmission might be achieved.

5.1 Decision Rules

- A decision rule, or a decoding function $\Delta: B \rightarrow A$
  - $b_j \rightarrow \Delta(b_j) = a_{j^*}$
  - Meaning: receiver sees b_j and decides $a_i = a_{j^*}$ was sent

Example 5.1

Let Γ be the BSC, so that A = B = Z_2. If the receiver trusts this channel, then Δ should be the identity function.

The average probability Pr_C of correct decoding is

$$\mathrm{Pr}_\mathrm{C} = \sum_j q_j Q_{j^*j} = \sum_j R_{j^*j} \qquad (5.1)$$

where $\mathrm{Pr}\,(a = a_{j^*} \mid b = b_j) = Q_{j^*j}$. and $R_{ij} = q_j Q_{ij}$

Decision Rules (Cont.)

- The error probability Pr_E (the average probability of incorrect decoding) is

$$\mathrm{Pr_E} = 1 - \mathrm{Pr_C} = 1 - \sum_j R_{j^*j} = \sum_j \sum_{i \neq j^*} R_{ij} \qquad (5.2)$$

- Ideal observer rule
  - Minimizes Pr_E, or equivalently, which maximizes Pr_C
- How to maximize Pr_C
  - For each j, we choose i = j* to maximize the backward probability $\mathrm{Pr}(a_i|b_j) = Q_{ij}$. Or
  - For each j, we choose i = j* to maximize the joint probability $R_{ij} = q_j Q_{ij}$.

Decision Rules (Cont.)

- Example 5.2
  - Γ is the BSC, compute the Ideal observer rule Δ.
- A maximum likelihood rule
  - For each j, we choose i = j^* to maximize the forward probability $\Pr(b_j|a_i) = P_{ij}$.
- Among all the decision rules for a given channel, the maximum likelihood rule maximizes the integral of $\Pr_C$ over all input distributions $\boldsymbol{p} \in \boldsymbol{P}$.

$$\int_{\mathbf{p}\in\mathcal{P}} \Pr\nolimits_C \, dp_1 \ldots dp_r$$

Examples

- Example 5.3
  - Let us apply the maximum likelihood rule Δ to the BSC, where $P > 1/2$ and compute Pr_C and Pr_E. (input probabilities $p, \bar{p}$)
- Example 5.4
  - For a specific illustration, let us return to Example 4.5, where P = 0.8 and p = 0.9.
  - Compare the maximum likelihood rule and the ideal observer rule
- Example 5.5
  - Let Γ be the binary erasure channel (BEC) in Example 4.2, with P > 0. Compute the maximum likelihood rule, and compute Pr_C and Pr_E. (input probabilities $p, \bar{p}$)

5.2 An Example of Improved Reliability

- Given an unreliable channel, how can we transmit information through it with greater reliability?
  - Considering BSC with 1 > P > $1/2$.
    - Compute the maximum likelihood rule
    - Compute the mutual information I(A, B), assuming $p = 1/2$
    - Compute the error-probability Pr_E
  - Now, sending each input symbol a = 0 or 1 three times in succession. So
    - The input consists of two binary words 000 and 111.
    - the output consists of eight binary words 000, 001, 010, 100, 011, 101, 110, and 111.
    - Transmission rate is $1/3$

An Example of Improved Reliability (Cont.)

- the forward probabilities for this new input and output

$$\begin{pmatrix} P^3 & P^2Q & P^2Q & P^2Q & PQ^2 & PQ^2 & PQ^2 & Q^3 \\ Q^3 & PQ^2 & PQ^2 & PQ^2 & P^2Q & P^2Q & P^2Q & P^3 \end{pmatrix}$$

- the maximum likelihood rule, called majority decoding

$$\Delta : \begin{cases} 000, 001, 010, 100 \mapsto 000, \\ 011, 101, 110, 111 \mapsto 111. \end{cases}$$

- a new binary symmetric channel Γ'

$$M' = \begin{pmatrix} P^3 + 3P^2Q & 3PQ^2 + Q^3 \\ 3PQ^2 + Q^3 & P^3 + 3P^2Q \end{pmatrix} \qquad \begin{matrix} 0 \\ 1 \end{matrix} \longrightarrow \begin{matrix} 000 \\ 111 \end{matrix} \longrightarrow \Gamma \longrightarrow \begin{matrix} 000 \\ 001 \\ 010 \\ 100 \\ 011 \\ 101 \\ 110 \\ 111 \end{matrix} \longrightarrow \begin{matrix} 0 \\ 1 \end{matrix}$$

  - $\Pr_C = P^3 + 3P^2Q$
  - $\Pr_E = 3PQ^2 + Q^3 = Q^2(3 - 2Q) \approx 3Q^2$

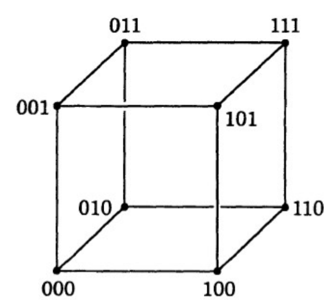
Generalized Idea

- If Γ is a channel with an input A having an alphabet A of r symbols, then any subset $C \subseteq A^n$ can be used as a set of code-words which are transmitted through Γ
  - For instance, the repetition code R^n over A consists of all the words w = $aa \ldots a$ of length n such that $a \in A$.
  - In this case, $|C| = r = r^1$. So the rate is $1/n$.
  - In general, $|C| = r^k$. So the rate is k/n.
- The transmission rate can be defined as

$$R = \frac{\log_r |C|}{n} \tag{5.3}$$

5.3 Hamming Distance

- Let $\boldsymbol{u} = u_1 \ldots u_n$ and $\boldsymbol{v} = v_1 \ldots v_n$ be words of length n in some alphabet A, so $\boldsymbol{u}, \boldsymbol{v} \in A^n$. The Hamming distance d($\boldsymbol{u}$, $\boldsymbol{v}$) between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined to be the number of subscripts i such that $u_i \neq v_i$.
- Example 5.6
  - Let $\boldsymbol{u}$ = 01101 and $\boldsymbol{v}$ = 01000 in Z_2^5. Then d($\boldsymbol{u}$, $\boldsymbol{v}$) = 2.
- Example 5.7
  - We can regard the words in Z_2^3 as the eight vertices of a cube.



Hamming Distance (Cont.)

- Lemma 5.8

  Let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in A^n$. Then

  (a) $d(\mathbf{u}, \mathbf{v}) \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{v}$;

  (b) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$;

  (c) $d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$.

- To transmit information through Γ , we choose a code $C \subseteq A^n$ for some n , and use the maximum likelihood decision rule.
  - Decode each received word as the code-word most likely to have caused it. (Using forward probability P_{ij}.)

Hamming Distance (Cont.)

- For simplicity, assume that Γ is the BSC, with P > $1/2$, so A = B = Z_2 and r = 2.
  - The **maximum likelihood** decision rule means for any output $v \in Z_2^n$, we decode v as the code-word u = $\Delta(v)$ $\in C$ which maximizes the forward probability Pr(v I u).
  - Note: a code-word u which maximizes Pr(v I u) is one which minimizes d(u, v).

    If $d(u, v) = i$ then

$$\Pr(v|u) = Q^i P^{n-i} = P^n \left(\frac{Q}{P}\right)^i$$

  - So, this is also called the **nearest neighbor decoding**

5.4 Statement of Shannon's Theorem

- Informally
  - Shannon's Theorem says that if we use long enough code-words then we can send information through a channel Γ as accurately as we require, at a rate arbitrarily close to the capacity C of Γ.
- Theorem 5.9
  - Let Γ be a binary symmetric channel with P > $1/2$, so Γ has capacity C = 1 - H(P) > 0, and let $\delta, \varepsilon > 0$. Then for all sufficiently large n there is a code $C \subseteq Z_2^n$, of rate R satisfying $\mathrm{C} - \varepsilon \leq R < \mathrm{C}$, such that nearest neighbor decoding gives error-probability $\mathrm{Pr}_E < \delta$.

Outline Proof of Shannon's Theorem

- Let R < C, Randomly chose $C \subset Z_2^n$, $|C| = 2^{nR}$.
- Rate of C = $\,^{log_2 2^{nR}}/_{n} = R$
- Sending $\boldsymbol{u}$, expect to receive $\boldsymbol{v}$ such that d($\boldsymbol{u}$, $\boldsymbol{v}$) $\approx nQ$
- Receiving $\boldsymbol{v}$, decode $\Delta(\boldsymbol{v}) = \boldsymbol{u}$ such that d($\boldsymbol{u}$, $\boldsymbol{v}$) $\approx nQ$
- Using the nearest neighbor rule, if decoding is incorrect then there must be some $\boldsymbol{u}' \neq \boldsymbol{u}$ in C with d($\boldsymbol{u}'$,$\boldsymbol{v}$) $\leq$ d($\boldsymbol{u}$,$\boldsymbol{v}$).
- So

$$\mathrm{Pr_E} \leq \sum_{\mathbf{u}' \neq \mathbf{u}} \mathrm{Pr}\left(d(\mathbf{u}', \mathbf{v}) \leq nQ\right), \qquad (5.4)$$

- The upper bound on Pr_E in (5.4) is equal to

$$(|\mathcal{C}| - 1)\,\mathrm{Pr}\left(d(\mathbf{u}', \mathbf{v}) \leq nQ\right) < 2^{nR}\,\mathrm{Pr}\left(d(\mathbf{u}', \mathbf{v}) \leq nQ\right).$$

Outline Proof (Cont.)

- For any given $\boldsymbol{v}$ and i, $|\{\, \boldsymbol{u}' \in Z_2^n : \mathrm{d}(\boldsymbol{u}', \boldsymbol{v}) = i \,\}| = \binom{n}{i}$
- So, $|\{\boldsymbol{u}' \in Z_2^n : \mathrm{d}(\boldsymbol{u}', \boldsymbol{v}) \le nQ\}| = \sum_{i \le n} \binom{n}{i}$
- Therefore

$$\Pr\left(d(\mathbf{u}', \mathbf{v}) \le nQ\right) \doteq \frac{1}{2^n} \sum_{i \le nQ} \binom{n}{i}$$

- Exercise 5.7

Show that if $\lambda + \mu = 1$, where $0 \le \lambda \le \frac{1}{2}$, then

$$1 \ge \sum_{i \le \lambda n} \binom{n}{i} \lambda^i \mu^{n-i} \ge \sum_{i \le \lambda n} \binom{n}{i} \lambda^{\lambda n} \mu^{\mu n}$$

hence show that

$$\sum_{i \le \lambda n} \binom{n}{i} \le 2^{nH(\lambda)}.$$

Outline Proof (Cont.)

- Putting $\lambda = Q$ in Exercise 5.7, we have

$$\sum_{i \le nQ} \binom{n}{i} \le 2^{nH(Q)}$$

- Thus (5.4) becomes

$$\mathrm{Pr}_{\mathbf{E}} < 2^{nR}.\frac{1}{2^n}.2^{nH(Q)} = 2^{n(R-1+H(Q))} = 2^{n(R-C)}$$

  - Note: C = 1 - H(P) = 1 - H(Q).

- Now R < C, so $2^{n(R-C)} \to 0$ as $n \to \infty$, and hence $\mathrm{Pr}_E \to 0$ also.

5.5 The Converse of Shannon's Theorem

- Informally
  - The converse of Shannon’s Theorem says that one can not do better than what the Shannon’s Theorem says.
- The converse of Shannon’s Theorem
  - If C' > C then it is not true that for every $\varepsilon > 0$ there is a sequence of codes C, of lengths $n \to \infty$, and of rates R satisfying C' - $\varepsilon \leq$ R < C', such that $\mathrm{Pr}_E \to 0$ as $n \to \infty$.
- The Fano bound
  - gives a lower bound on the error-probability. (See Theorem 5.10 on the next slide.)

The Fano Bound

- Theorem 5.10
  - Let Γ be a channel with input A and output B. Then the error-probability $\mathrm{Pr_E}$ corresponding to any decision rule Δ for Γ satisfies

$$H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathrm{Pr_E}) + \mathrm{Pr_E} \log(r-1) \qquad (5.5)$$

    where r is the number of symbols in A
- Meaning of inequality (5.5)
  - Given b_j, the receiver decodes $a_{j*} = \Delta(b_j)$, which may or may not be the actual symbol a_i transmitted.
  - The left-hand side of (5.5) is the extra information the receiver needs (on average) in order to know a_i

The Fano Bound (Cont.)

- Meaning of inequality (5.5)
  - This extra information can be divided into two parts:
    a) Whether or not decoding is correct, that is, whether or not $a_{j*} = a_i$;
    b) If decoding is incorrect, then which $a_i (i \neq j^*)$ out of r-1 symbols was transmitted.
  - The information in (a) has value $H(\mathrm{Pr}_E)$
  - The information in (b) has value at most $\mathrm{Pr}_E \log(r-1)$
- Note: we have

$$\mathrm{Pr_C} = \sum_j R_{j^*j} \quad \text{and} \quad \mathrm{Pr_E} = \sum_j \sum_{i \neq j^*} R_{ij},$$

Examples

- Example 5.11
  - Let Γ be the BSC , and as a rather extreme example of a code let us take C = A^n, so R = 1.
  - If 0 < P < 1 we have C = 1 - H(P) < 1, so R > C.
  - Using the identity function $\Delta(u)$ = u as a decision rule, we see that decoding is correct if and only if there are no errors, so $\mathrm{Pr}_E = 1 - P^n \to 1$ as $n \to \infty$.

Examples (Cont.)

- Example 5.12
  - The Hamming codes of length n of the form $2^c - 1$ and rate R = (n - c)/n, so R → 1 as n → ∞.
  - If we use a BSC with 0 < P < 1, then C = 1 - H(P) < 1 and hence R > C for all sufficiently large n.
  - The nearest neighbor decoding is correct if and only if there is at most one error (shall see this in §7.4), so $\mathrm{Pr}_E = 1 - P^n - nP^{n-1}Q \rightarrow 1$ as $n \rightarrow \infty$.

5.6 Comments on Shannon's Theorem

- Theorem 5.13 (The general form of Shannon's Theorem)
  - Let Γ be an information channel with capacity C > 0, and let $\delta, \varepsilon > 0$. For all sufficiently large n there is a code C of length n, of rate R satisfying $\mathrm{C} - \varepsilon \leq R < \mathrm{C}$, together with a decision rule which has error-probability $\mathrm{Pr}_E < \delta$.
- Comment 5.14
  - In order to achieve values of R close to C and Pr_E close to 0, one may have to use a very large value of n.
  - This means that code-words are very long, so encoding and decoding may become difficult and time-consuming.

Comments on Shannon's Theorem

- Comment 5.14
  - Moreover, if n is large then the receiver experiences delays while waiting for complete codewords to come through; when a received word is decoded, there is a sudden burst of information, which may be difficult to handle.
- Comment 5.15
  - Shannon's Theorem tells us that good codes exist, but neither the statement nor the proof give one much help in finding them.

Comment 5.15 (Cont.)

- The proof shows that the "average" code is good, but there is no guarantee that any specific code is good: this has to be proved by examining that code in detail.
- One might choose a code at random, as in the proof of the Theorem, and there is a reasonable chance that it will be good.
- However, random codes are very difficult to use: ideally, one wants a code to have plenty of structure, which can then be used to design effective algorithms for encoding and decoding.
- We will see examples of this in Chapters 6 and 7, when we construct specific codes with good transmission rates or error-probabilities.