Information Theory - Communication Through Noisy Channels

Eric Filiol

ESIEA - Laval
Laboratoire de cryptologie et de virologie opérationnelles
$(C + V)^O$
filiol@esiea.fr

2013 - 2014

Introduction : Communication Channel (DMC)

- A *communication channel* C is a black box accepting string of symbols from its input alphabet Σ_1 and emits string of symbols from its output alphabet Σ_2.
- We restrict (wlog) first to *discrete memoryless channels* with $\Sigma_1 = \{a_1, a_2, \ldots, a_m\}$ and $\Sigma_2 = \{b_1, b_2, \ldots, b_n\}$.
- It is fully defined by its *channel matrix* $P = (p_{ij}, 1 \leq i \leq m, 1 \leq j \leq n)$
  - The channel mode of operation is that, if any sequence $(u_1, u_2, \ldots, u_N)$ of symbols from Σ_1 is input, the output sequence is a string $(v_1, v_2, \ldots, v_N)$ (of the same length) of symbols from Σ_2.
  - We have

$$P(v_k = b_j | u_k = a_i) = p_{ij} \qquad (1 \leq i \leq m, 1 \leq j \leq n)$$

  - For each i, we have $\sum_j p_{ij} = 1$.

Introduction : Communication Channel (DMC) (2)

- P is a *stochastic matrix* (matrix whose entries are non-negative, and with row sums equal to 1). P is in fact the transition matrix of a Markov chain.

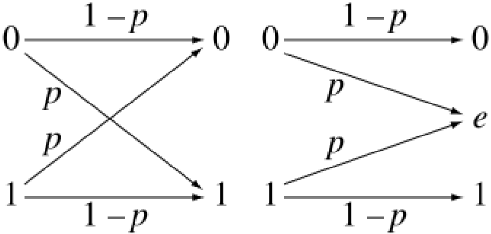


FIGURE: Binary Symmetric Channel Diagram (left) & Binary Erasure Channel Diagram (right)

- Exercice : give the matrix P for these two communication channels.

Introduction : Extentions of DMC

- The r *extension* of a discrete memoryless channel with input alphabet $\Sigma_1^{(r)}$, output alphabet $\Sigma_2^{(r)}$ and channel matrix $P^{(r)}$ is defined as follows
  - The (i, j) entry of $P^{(r)}$ correspond to an input $\sigma_i = \alpha_1\alpha_2 \ldots \alpha_r$ with $\alpha_k \in \Sigma_1$ and an output $\tau_j = \beta_1\beta_2 \ldots \beta_r$ with $\beta_k \in \Sigma_2$
  - We have
    $$P^{(r)}(i, j) = p(\beta_1|\alpha_1).p(\beta_2|\alpha_2) \ldots p(\beta_r|\alpha_r)$$
  - $p(\beta_k|\alpha_k)$ is the probability that β_k is received when α_k is emitted.
- Exercice : give the matrix $P^{(r)}$ for $r = 2$ as well its diagram (hint : think of an r-th extension of C as r parallel and independent copies of C).

Agenda

1. Introduction : The Discrete Memoryless Channel
2. The Noisy Channel Issue
3. Codes and Decoding Rules
4. Channel Capacity
5. Shannon's Second Theorem on Noisy Coding
6. Conclusion
7. Bibliography

Agenda

1 Introduction : The Discrete Memoryless Channel

2 The Noisy Channel Issue

3 Codes and Decoding Rules

4 Channel Capacity

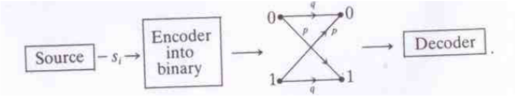
5 Shannon's Second Theorem on Noisy Coding

6 Conclusion

7 Bibliography

Introduction : Connecting the Source to the Channel

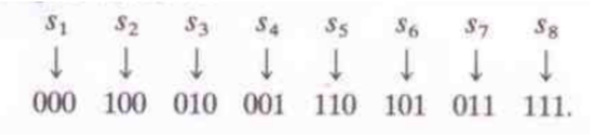
- Suppose we have a memoryless source $\mathcal{S}$ which emits symbols $\{s_1, s_2, \ldots, s_N\}$ with probabilities $\{p_1, p_2, \ldots, p_N\}$ respectively. We connect this source to a communication channel (here a BSC) with error probability p.
- We assume that the encoding into binary ($D = 2$) is noiseless and known to the decoder.



- The critical issue, solved the Shannon's second theorem deals with the problem of preventing the channel to corrupt data in such a way that the recipient will decode $\hat{s}_j$ different from the symbols s_j emitted and hence will lose part of the emitted message.

Illustrating the Issue

- Take $N = 8$ and the following optimal (i.e compact) encoding :

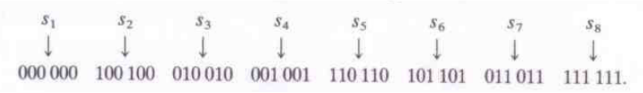


$$\begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 000 & 100 & 010 & 001 & 110 & 101 & 011 & 111. \end{array}$$

- Give the probabilities that any particular word is correctly transmitted and that a message of n words is correctly transmitted.
- The question is
  - Can we do better ?
  - At what cost ?

Illustrating the Issue (2)

- We now double up the encoding as follows :



$$\begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 000\,000 & 100\,100 & 010\,010 & 001\,001 & 110\,110 & 101\,101 & 011\,011 & 111\,111. \end{array}$$

- The decoding rule is
  - We decode only when the first three symbols and the second three symbols agree.
  - Otherwise we ask the word to be emitted again (provided that we can make such a request).
- Analyze the cost and the probability that an error remain undetected.
- Case of the k-repetition codes.
- Shannon's second theorem deals with the problem of achieving reliable transmission through a noisy channel at an optimal cost and without requiring the receiver has no feedback to the sender.

Agenda

1. Introduction : The Discrete Memoryless Channel
2. The Noisy Channel Issue
3. Codes and Decoding Rules
4. Channel Capacity
5. Shannon's Second Theorem on Noisy Coding
6. Conclusion
7. Bibliography

Codes and Decoding Rules

- Let us consider a memoryless channel with input alphabet σ_1 and output alphabet σ_2. A *code of length* n is any *collection* $\mathcal{C}$ of distinct n-sequences of symbols from σ_1.
- The elements are called the *codewords*.
- Given a code $\mathcal{C}$ of length n with codewords $c_1, c_2, \ldots, c_N$, a *decoding rule* is any partition of the set of possible received sequences into (disjoint) sets $R_1, R_2, \ldots, R_N$ with the rule that if $y \in R_j$ then y is decoded as c_j.
- The core of the problem is hence to choose a decoding rule (i.e. a partition) that make the *residual probability of error* as small as possible.

Decoding Models

- Our aim is to minimize the risk of residual error. This suggests decoding any received vector y into a codeword c_j such that

$$P(c_j \text{ sent}|y \text{ received}) \geq P(c_i \text{ sent}|y \text{ received}) \qquad \forall i$$

- This rule is called the *ideal-observer* or *minimal error* rule. But its major drawback is that it requires the knowledge of the probabilities that the c_j are used. Moreover is is difficult to use for a large number of codewords.
- We consider another rule called *maximum-likelihood* decoding rule. We decode a received y into a c_j that maximizes

$$P(y \text{ received}|c_j \text{ sent})$$

Theorem

If the codewords are equally probable, then the maximum-likelihood decoding rule agrees with the ideal-observer rule.

- Proof left as an exercice.

Hamming Distance

- For binary symmetric channel, there is a very simple implementation of the maximum-likelihood decoding rule.
- Let V_n denotes the set of all binary m-sequences (it is thus a m-dimensional vector space over $\mathbb{F}_2$. If x and y are two vectors in V_n, we define the *Hamming distance* $d(x, y)$ between x and y as $|\{i \in [1, n] | x_i \neq y - i\}|$.
- For the BSC, a natural decoding rule is the minimum-distance rule :
  - We decode any received vector y into a vector c that is a minimum Hamming distance from y. If there are more than one such codeword, we choose arbitrarily.

Theorem

For the BSC with error probability $p < \frac{1}{2}$, the minimum-distance rule is equivalent to the maximum-likelihood decoding rule.

- Proof left as an exercice.

Agenda

1 Introduction : The Discrete Memoryless Channel

2 The Noisy Channel Issue

3 Codes and Decoding Rules

4 Channel Capacity

5 Shannon's Second Theorem on Noisy Coding

6 Conclusion

7 Bibliography

Channel Capacity

- The capacity of a communication channel is a measure of its ability to transmit information. If we transmit more than its capacity, we are bound to lose information (think to a large diameter tap connected to a small diameter pipe).
- Suppose that we have a DMC with input alphabet σ_1 and output alphabet σ_2 and a channel matrix $P = [p_{ij}] = [P(b_j \text{ received}|a_i \text{ sent})]$. We attach a memoryless source $\mathcal{S}$ to this DMC. $\mathcal{S}$ emits symbols $\{a_1, a_2, \ldots, a_m\}$ with probabilities $(p_1, p_2, \ldots, p_m)$ respectively.
- The DMC output can be considered as a source $\hat{\mathcal{S}}$ which emits symbols $\{b_1, b_2, \ldots, b_n\}$ with probabilities $(q_1, q_2, \ldots, q_n)$ respectively where

$$q_j = \sum_{i=1}^{m} P(b_j \text{ received}|a_i \text{ sent}).P(a_i \text{ sent}) = \sum_{i=1}^{m} p_i p_{ij}$$

Channel Capacity (2)

- The information about $\mathcal{S}$ provided by $\hat{\mathcal{S}}$ is (refer to the first course) defined by

$$I(\mathcal{S}|\hat{\mathcal{S}}) = H(\mathcal{S}) - H(\mathcal{S}|\hat{\mathcal{S}}) = H(\mathcal{S}) + H(\hat{\mathcal{S}}) - H(\mathcal{S}, \hat{\mathcal{S}})$$

  It is a function of the source distribution $(p_1, p_2, \ldots, p_m)$ and matrix P only.

- It is then natural to define the *capacity* of a channel by

$$C = \mathsf{sup}\{I(\mathcal{S}|\hat{\mathcal{S}})\}$$

  whose sum is taken over all memoryless sources $\mathcal{S}$ or equivalently over all possible input distributions $(p_1, p_2, \ldots, p_m)$.

Channel Capacity : Theorem

- C is mathematically well defined since f is a continuous function on a closed and bounded subset of $\mathbb{R}^m$ and since any continuous function on such a set attains its supremum on the set, we can then rewrite

$$C = \max\{I(\mathcal{S}|\hat{\mathcal{S}})\}$$

- C is a quantity defined by P entirely. It is analogous to the conductance of a resistor in electrical network theory. Its units is then clearly *bits per second* or *bits per symbol*, depending on the context.
- Let us give the theorem which shows how to computer capacity more explicitely.

BSC Capacity

The capacity of a BSC of error probability p is given by

$$C(p) = 1 + p\log_2(p) + q\log_2(q)$$

where $q = 1 - p$.

Channel Capacity : Theorem for r-extension

- We have $C(0) = 1$ (perfect transmission) and $C(\frac{1}{2}) = 0$ (perfect scrambler).
- Working out the capacities or general channels is far from being trivial, unless having special properties.
- Let us give the generalized version of the theorem for r-extensions of BSC.

Capacity for r-extensions

If a memoryless channel has capacity C then its r-extension has capacity rC.

Agenda

1. Introduction : The Discrete Memoryless Channel
2. The Noisy Channel Issue
3. Codes and Decoding Rules
4. Channel Capacity
5. Shannon's Second Theorem on Noisy Coding
6. Conclusion
7. Bibliography

Introduction

- We have shown that it is possible to achieve arbitrarily high reliability (think to k-repetition codes).
- Shannon's second theorem shows that, provided that one keeps the transmission rate below the channel capacity, we can achieve high reliability.
- Wlog we focus on BSC.
- Given any code $\mathcal{C}$ and any decoding scheme for $\mathcal{C}$, the *error probability* $e(\mathcal{C})$ is usually defined as the average probability of error when considering equally distributed codewords. If there are M codewords $c_1, c_2, \ldots, c_M$ in $\mathcal{C}$, then (under the assumption that maximum-likelihood decoding rule is used),

$$e(\mathcal{C}) = \frac{1}{M}\sum_{i=1}^{M} P(\text{error}|c_i \text{ transmitted})$$

Introduction (2)

- The goal is to find codes with small average error probability. However it is not sufficient. Why (compare to average and worst complexity cases) ?
- We in fact require a much stronger property : the *maximum error probability* is small. It is defined by

$$\hat{e}(\mathcal{C}) = \mathrm{max}_i\{P(\mathrm{error}|c_i \text{ transmitted})\}$$

- Clearly we have

$$\hat{e}(\mathcal{C}) \geq e(\mathcal{C})$$

Shannon's Second Theorem on Noisy Coding

Shannon's Noisy Coding Theorem

Given a binary symmetric channel of capacity C and any R, with $0 < R < C$, then, if $(M_n : 1 \leq n < \infty)$ is any sequence of integers satisfying

$$1 \leq M_n \leq 2^{Rn} \qquad 1 \leq n < \infty,$$

and $\epsilon > 0$ any given positive quantity, there exists a sequence of codes $(\mathcal{C}_n : 1 \leq n < \infty)$ and an integer $N_0(\epsilon)$ with $\mathcal{C}_n$ having M_n codewords of length n and with maximum error probability

$$\hat{e}(\mathcal{C}) \leq \epsilon$$

for all $n \geq N_0(\epsilon)$

- Proof can be omitted.

Interpretation

- Let us consider a simple but illustrative example. Suppose that the error probability is such that the channel capacity is $C(p) = 0.8$.
- If our message is a binary string, we know that for sufficiently large n, if we take $R = 0,75$, there exists a set of $2^{0.75.n}$ codewords of length n that have error probability less than a prescribed threshold.
- To encode the message stream from the source, the procedure is
  1. Break the message stream up into blocks of length m, where m is such that $3\lceil \frac{1}{4}.n \rceil = m \geq \frac{3}{4} N_0(\epsilon)$.
  2. Encode these m-blocks into the code $\mathcal{C}_n$ using a codeword of length $\frac{4}{3}.m$ for each block.
  3. Transmit the new encoded stream through the channel.

Shannon's Noisy Extension to General DMC

- We consider general DMC with arbitrary input and output alphabets. Shannon's theorem extends to this more general case.
- The sketch of the proof is the same
  1. Code messages randomly (*random coding*).
  2. Decode by the maximum-likelihood procedure.

Stronger Version of Shannon's Noisy Coding Theorem (Shannon, 1957)

If a discrete memoryless channel has capacity C and R is any quantity, with $0 < R < C$, there exists a sequences of codes $(\mathcal{C}_n : 1 \leq n < \infty)$ such that :

1. $\mathcal{C}_n$ has $\lfloor 2^{Rn} \rfloor$ codewords of length n.
2. $\hat{e}(\mathcal{C}_n)$ the maximum error probability of $\mathcal{C}_n$ satisfies $\hat{e}(\mathcal{C}_n) \leq A.e^{-Bn}$, where A and B depend only on the channel and on R.

- In other words, not only do there exists good codes but there are codes whose error probabilities decrease exponentially.

Capacity as Bound to Accurate Communication

We are now giving a converse theorem to Shannon's noisy coding theorem.

Welsh's Theorem (1988)

For a memoryless channel of capacity C and any $R > C$, there cannot exist a sequence of codes $(\mathcal{C}_n : 1 \leq n < \infty)$ with the property that $\mathcal{C}_n$ has 2^{nR} codewords of length n and error probability $e(\mathcal{C}_n)$ that tends towards 0 as $n \to \infty$.

- (Wolfowitz, 1961) proved a much stronger result : $\hat{e}(\mathcal{C}_n)$ converges to 1 as $n \to \infty$.

Fano's Inequality

- Let $\mathcal{C}$ be any code with M codewords $(c_1, c_2, \ldots, c_M)$ for a DMC. Let X be a random vector taking values in the set of codewords. Let Y denote the random output vector when X is transmitted through the the channel and decoded. Then, if p_E is the probability of an error, namely $p_E = P(X \neq Y)$, we have

Fano's Inequality

$$H(X|Y) \leq H(p_E, q_E) + p_E \log_2(M - 1)$$

where $p_E = 1 - q_E$.

- This inequality is useful to prove Welsh's and Wolfowitz's theorem.
- It also has a very natural interpretation : the term $H(p_E, q_E)$ is the information needed to decide whether or not there is an error, the second term $(p_E \log_2(M - 1))$ is the information needed to resolve the error.

Agenda

1. Introduction : The Discrete Memoryless Channel
2. The Noisy Channel Issue
3. Codes and Decoding Rules
4. Channel Capacity
5. Shannon's Second Theorem on Noisy Coding
6. Conclusion
7. Bibliography

Conclusion

- Shannon's theorem for noisy coding proves that we always can have a maximal transinformation (or a minimal residual error probability after decoding).
- It is therefore possible to manage noise over communication with reliably.
- Existence theorem only : Shannon's tells us only that such efficient codes exist.
- Shannon's theory gave birth to the error-correcting theory (how to build such codes practically) that will be exposed in Course INF5047.
- Shannon's noisy coding theory assumes that the noise is neither adaptative nor malicious (safety issue) and does not consider malicious and adaptative behaviour (security issue). This is (partly) covered by his 3rd theorem on *perfect secrecy* (exposed in the Cryptography Minor course).
- Go now to the computer room to practice with exercices.

Agenda

1. Introduction : The Discrete Memoryless Channel
2. The Noisy Channel Issue
3. Codes and Decoding Rules
4. Channel Capacity
5. Shannon's Second Theorem on Noisy Coding
6. Conclusion
7. Bibliography

Essential Bibliography

A few papers are available on the Moodle repository for this lecture.

- Fano, R.M. (1961). *Transmission of Information*, MIT Press.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.
- Shannon, C.E. (1957). Certain Results in Coding Theory for Noisy Channels. *Information and Control*, 1(1), pp. 6–25.
- Welsh, D. (1988). *Codes and Cryptography*, Oxford Science Publishing.
- Wolfowitz, J. (1961). *Coding Theorems of Information Theory*. Prentice Hall.