Basics of Information Theory
8. Channel Coding Theorem

Markku Juntti

Overview The channel coding theorem is derived.
It is extended also to feedback channels and joint
source-channel coding framework.
Source The material is mainly based on Sections
8.4–8.10 and 8.12–8.13 of the course book [1].

Outline of the Lecture
• Review of the last lecture
  – Properties of channel capacity
  – Preview of channel coding theorem
  – Definitions
  – Jointly typical sequences
• Channel coding theorem
• Feedback capacity
• Joint source channel coding theorem
• Summary

Course Overview
Basic concepts and tools
1 Introduction
2 Entropy, relative entropy and mutual information
3 Asymptotic equipartition property
4 Entropy rates of a stochastic process
Source coding or data compression
5 Data compression
Channel capacity
8 Channel capacity
9 Differential entropy
10 The Gaussian channel
Other applications
11 Maximum entropy and spectral estimation
13 Rate distortion theory
14 Network information theory

Review of Last Lecture
• The fundamental questions behind information
  theory:
  – What is the ultimate data compression rate?
    • Answer: entropy.
  – What is the ultimate data transmission rate?
    • Answer: channel capacity.
• A communicates with B: physical acts of A induce
  desired physical state in B.
  – A and B should agree on what was sent.
• The maximum number of distinguishable signals
  channel capacity = data rate limit for reliable
  communication.
**Channel Coding and Capacity Problem**

Information theoretic model of a communication system.
- Message is encoded, transmitted through channel, output of which is then decoded.
  - Allows introduction of controlled redundancy.
  - Compare to source coding which attempts to remove redundancy.
- Sequences of channel symbols are considered.
- The channel capacity is solved as the maximum number of distinguishable messages per channel use.
- The basic information theoretic channel model is a conditional probability mass function.
  - Assumed to be known both by encoder and decoder.

**Properties of Channel Capacity**

1. \( C \geq 0 \), since \( \mathcal{I}(X;Y) \geq 0 \).
2. \( C \leq \log(|X|) \), since \( C = \max \mathcal{I}(X;Y) \leq \max \mathcal{H}(X) = \log(|X|) \).
3. \( C \leq \log(|Y|) \).
4. \( \mathcal{I}(X;Y) \) is a continuous function of \( p(x) \).
5. \( \mathcal{I}(X;Y) \) is a concave function of \( p(x) \).

Local maximum is a global one and finite.
- Algorithms available to find the maximum:
  - constrained maximization using calculus (Kuhn-Tucker conditions)
  - Frank-Wolfe gradient search algorithm
  - iterative Blahut-Arimoto algorithm.

**Definitions**

- **Discrete channel** consists of
  - input alphabet \( X \)
  - output alphabet \( Y \)
  - probability transition (matrix) \( p(y|x) \).
- Channel is **memoryless**, if the output is conditionally independent of the previous inputs and outputs.
- "**Information**" channel capacity:
  \[
  C = \max_{p(x)} I(X;Y).
  \]

**Preview of Channel Coding Theorem**

- Intuition behind the channel coding theorem:
  - Use sequences of \( n \) symbols for communications.
  - Use only uniquely distinguishable sequences, if possible.
- For each input sequence: \( 2^{n\mathcal{H}(Y)} \) typical output sequences.
- The total number of typical output sequences: \( 2^{n\mathcal{H}(Y)} \).
- Use only uniquely distinguishable input sequences. Divide the set of \( 2^{n\mathcal{H}(Y)} \) sequences into disjoint sets of size \( 2^{n\mathcal{H}(Y)} \).
  The number of such sets \( \leq 2^{n\mathcal{H}(Y) - \mathcal{H}(Y|X)} = 2^{n\mathcal{I}(X;Y)} \).
- Channel coding theorem: above bound is achievable.
Distinguishable Sequences

Definitions

- **Memoryless** channel:
  \[ p(y_k | x_k, y_{k-1}) = p(y_k | x_k), \quad k = 1,2,\ldots,n. \]
- Probability of error:
  - \( i \)th message: \( \lambda_i = \Pr[g(y^n) \neq i | X^n = X^n(i)] \).
  - average: \( p_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \).
  - maximum: \( \lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i \).
- Code rate \( R \) is **achievable**, if there exists a sequence of \( (\lceil 2^{nR} \rceil, n) \) codes so that \( \lambda^{(n)} \rightarrow 0 \), as \( n \rightarrow \infty \).
- "**Operational**" channel capacity: supremum of achievable rates.

Jointly Typical Sequences

- The concept of typical sequences and typical sets can be generalized to two (or more) sequences.
  - Direct generalization.
  - Defines the set of typical sequences with most of the probability.
- This enables the concept of decoding based on joint typicality.
  - Often applied in information theory for analytical tractability benefits.
  - Not a practical decoding method.
The Set of Jointly Typical Sequences

- The set $A_x^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$, where $x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ and $y^n = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n$, with respect to $p(x,y)$ is defined as

$$A_x^{(n)} = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : -\frac{1}{n} \log p(x^n) - H(X) < \varepsilon, \right.$$

$$\left. -\frac{1}{n} \log p(y^n) - H(Y) < \varepsilon, \right.$$

$$\left. -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) < \varepsilon \right\},$$

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

Joint Asymptotic Equipartition Property

Let $(x^n, y^n) \sim p(x^n, y^n)$. Then

1. $\Pr[(X^n, Y^n) \in A_x^{(n)}] \to 1$, $n \to \infty$.
2. $|A_x^{(n)}| \leq 2^{n(H(X,Y) + \varepsilon)}$.
3. If $(\tilde{x}^n, \tilde{y}^n) \sim p(x^n)p(y^n)$

$$\Rightarrow \Pr[(\tilde{x}^n, \tilde{y}^n) \in A_x^{(n)}] \leq 2^{-n(I(X,Y) - 3\varepsilon)}.$$

4. For $n$ large enough:

$$\Pr[(\tilde{x}^n, \tilde{y}^n) \in A_x^{(n)}] \geq (1 - \varepsilon)2^{-n(I(X,Y) - 3\varepsilon)}.$$

Illustration of Jointly Typical Sequences

- The joint AEP states that the probability of two randomly picked sequences form a jointly typical sequence pair is about $2^{-nI(X,Y)}$.

There are about $2^{nI(X,Y)}$ uniquely distinguishable signals $x^n$.

Channel Coding Theorem

- The basic theorem of information theory.
  - Shannon's original paper [5], but rigorous proof only later.
  - In certain circumstances, all the errors encountered by the channel can be corrected.
    - Counterintuitive!
    - Built on a few basic ideas:
      - Allow arbitrarily small but positive probability of error.
      - Use the channel (infinitely) many times the law of large numbers comes into effect.
      - Calculate the probability of error over random codebooks symmetrization existence of at least one good code.
  - The proof builds on the joint AEP.
    - Not the optimal decoding method though.
The Channel Coding Theorem

- All rates below channel capacity $C$ are achievable.
  - For all rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes so that
    \[ \lambda(n) \to 0, \text{ as } n \to \infty. \]
- (Weak) converse: any sequence of achievable $(2^{nR}, n)$ codes:
  \[ R \leq C. \]
- Strong converse: for $R < C$, $P_e^{(n)} \to 0$, and $R > C$, $P_e^{(n)} \to 1$ exponentially.
- Proof of achievability in [1, Sect. 8.7], converse in [1, Sect. 8.9]; introduction to the converse in [1, Sect. 8.8].

Remarks on Channel Coding Theorem

- “Information” channel capacity
  \[ = \text{“operational” channel capacity}. \]
- Asymptotic result in the length of the code word.
- No rules for practical code design.
  - There exist good codes or at least one good code with exponentially small probability of error for long code block lengths.
  - Random codes are good on the average.
    \[ \text{Practically useless due to complex decoding.} \]

Feedback Capacity

Information theoretic model of a feedback channel.

- The channel outputs are assumed to be available at the transmitter.
- A $(2^{nR}, n)$ feedback code is a sequence of mappings $x_i(W, Y_{i-1})$, where each symbol $x_i$ is a function only of the message $W$ and previous received values $Y_1, Y_2, ..., Y_{i-1}$.

Capacity with Feedback

- Capacity with feedback: supremum of achievable rates with feedback codes $C_{FB}$:
  \[ C_{FB} = C = \max_{\rho(x)} I(X; Y). \]
  Feedback does not increase the capacity.
- However, feedback may simplify achieving the capacity.
- Proof in [1, Sect. 8.12].
Joint Source-Channel Coding Theorem

- The fundamental questions behind information theory:
  - What is the ultimate data compression rate?
    - Answer: entropy.
  - What is the ultimate data transmission rate?
    - Answer: channel capacity.
- Now we consider these problems jointly.
  - Consider sending digital data or a digitized signal over a communication channel.
  - Should we keep the source coding / data compression separate from channel coding?

Source-Channel Coding Theorem

- \( \exists \) a source-channel code with \( P_e^{(n)} \rightarrow 0 \), if \( H(V) < C \), where the probability of error is
  \[
P_e^{(n)} = Pr(V \neq \hat{V}) = \sum_{y^n} \sum_{\hat{v}^n} p(y^n|v^n) p(v^n) i(g(y^n) \neq v^n),
\]
  where
  \( \hat{V} \) is the estimate of \( V \).
  \( i(x) \) is an indicator function: \( i(x) = \begin{cases} 1, & \text{if } x \text{ is true (non-zero)} \\ 0, & \text{if } x \text{ is false (zero)} \end{cases} \).
- Converse: If \( H(V) > C \), probability of error is bounded away from zero and reliable transmission is not possible.

Problem Definition

- Consider a stationary source described RV's \( V_1, V_2, \ldots, V_n \) with values from a finite alphabet \( \mathcal{V} \) forming a stochastic process that satisfies the AEP.
  - Examples:
    - IID RV's
    - irreducible Markov chains.
- The sequence is to be sent over the channel so that the receiver can reconstruct it.
  - Map the sequence to a codeword \( \mathcal{X}^n(V) \).
  - The receiver decodes it based on the received sequence \( Y^n \).

Basic Ideas of the Proof

- Proof: See textbook [1, Sect. 8.13].
- The data compression theorem is a consequence of the AEP: the subset of \( 2^{nH} \) data sequences contain most of the probability \( H \) bits per symbol suffice for the source code with a small probability of error.
- The data transmission theorem is a consequence of the joint AEP: the channel output sequence likely jointly typical with the input sequence while any other codeword is jointly typical with probability of about \( 2^{-nD} \) about \( 2^n \) codewords result in a small probability of error.
- The source and channel coding can be separated, and the result follows.
Remarks

A stationary ergodic source can be transmitted over a channel if its entropy rate is less than the capacity of the channel.

Source and channel coding can be separated.

- In general, limited to single-user channels
  - not true for all multiuser channels

- Asymptotically true only
  - with short block sizes joint source-channel coding may bring benefits.

- Sometimes the decoder may be able to utilize the source code.
  - Example: a human ear + brain can recognize language.

Summary

- "Information" channel capacity = "operational" channel capacity: \( C = \max_{P(X)} I(X; Y) \).
  - All rates below the capacity \((R<C)\) are achievable:
    \( \lambda^{(n)} \to 0 \), as \( n \to \infty \).
  - No rate above the capacity \((R>C)\) is achievable.

- Feedback does not increase the capacity.
  - Feedback can help achieving the capacity.

- A source with entropy rate below the capacity can be transmitted reliably. A source with entropy rate above the capacity cannot be transmitted reliably.