Information, Control and Computing

Based on Course Notes and "Information Theory and Coding (2nd Ed.)"

Contents

1 Introduction

1.1 Development of Information Theory

Information theory was founded by Claude E. Shannon in 1948 with his seminal paper "A Mathematical Theory of Communication." This work laid the foundation for quantifying information, establishing limits on data compression and reliable transmission over noisy channels. Over the decades, information theory has evolved to encompass network information theory, quantum information, and applications in machine learning, cryptography, and biology.

The second edition of the referenced book includes updates on modern topics such as network coding, multi-user information theory, and practical coding schemes like LDPC and Turbo codes, reflecting advancements since the first edition in 2010.

1.2 Information and Uncertainty

Information theory is an applied science based on probability and statistics, dealing with the transmission, storage, and processing of information.

- **Information vs. Message:** A message (e.g., text, voice, image) is the carrier. Information is the abstract content, defined by Shannon as "that which eliminates uncertainty."
- **Measurement:** Since uncertainty is probabilistic, information is measured using probability functions.

Basic concepts include the distinction between syntactic, semantic, and pragmatic information, but Shannon's theory primarily focuses on the syntactic level, measuring the amount of uncertainty eliminated.

1.3 Communication System Model

The fundamental model of a communication system, as established by Shannon, consists of the following components:

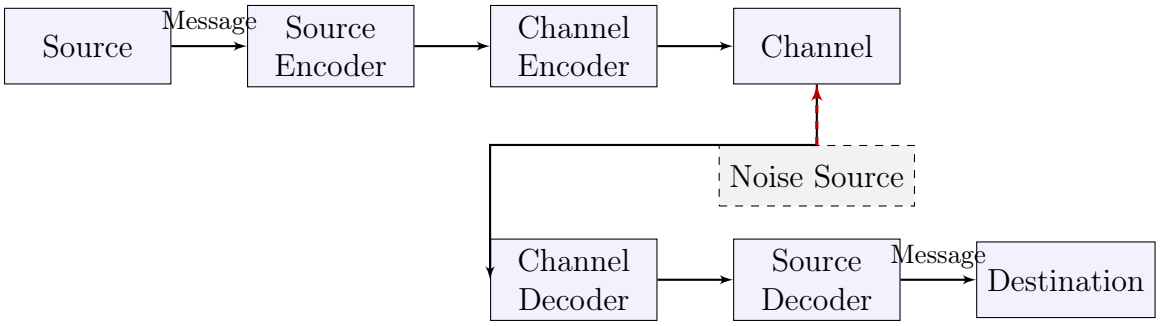


Figure 1: Shannon's General Communication System Model

1. **Source:** Generates the message (discrete sequence or continuous signal).
2. **Source Encoder:** Removes redundancy to compress the message (matching source rate to channel).

3. **Channel Encoder:** Adds redundancy to protect against noise (error correction).
4. **Channel:** The medium of transmission (introduces noise/interference).
5. **Decoders:** Perform the inverse operations of the encoders.

In modern extensions, multiple sources, channels, or users may be considered, leading to network models.

2 Entropy and Mutual Information

2.1 Information Measures for Random Variables

2.1.1 Self-Information

For a random event x occurring with probability $p(x)$, the self-information is:

$$I(x) = -\log_a p(x) \tag{1}$$

Units: **bit** ($a = 2$), **nat** ($a = e$), **Hartley** ($a = 10$).

Example: For a fair coin flip, $p(x) = 0.5$, $I(x) = 1$ bit.

2.1.2 Entropy (Average Uncertainty)

The entropy $H(X)$ of a discrete random variable X is the expected value of its self-information:

$$H(X) = E[I(X)] = -\sum_{i=1}^{K} p(x_i) \log p(x_i) \tag{2}$$

Properties of Entropy:

1. **Non-negativity:** $H(X) \geq 0$.
2. **Extremal Property:** $H(X) \leq \log K$. Equality holds if X is uniformly distributed.
3. **Convexity:** $H(P)$ is a strictly concave ($\cap$) function of the probability distribution P.
4. **Additivity:** If X, Y are independent, $H(X, Y) = H(X) + H(Y)$.
5. **Continuity:** Entropy is continuous in the probabilities.
6. **Subadditivity:** $H(X, Y) \leq H(X) + H(Y)$.

Example: For a binary source with $p = 0.1$, $H(X) = -0.1 \log_2 0.1 - 0.9 \log_2 0.9 \approx 0.469$ bits.

2.1.3 Joint and Conditional Entropy

- **Joint Entropy:** $H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$.
- **Conditional Entropy:** $H(X|Y) = -\sum_{x,y} p(x, y) \log p(x|y)$.
- **Chain Rule:** $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

2.2 Mutual Information

Mutual information $I(X;Y)$ measures the amount of information one random variable contains about another.

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \tag{3}$$

Key Relationships:

$$I(X;Y) = H(X) - H(X|Y) \quad \text{(Information gain)} \tag{4}$$
$$I(X;Y) = H(Y) - H(Y|X) \tag{5}$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{6}$$

Properties: Non-negative, zero if independent, symmetric.

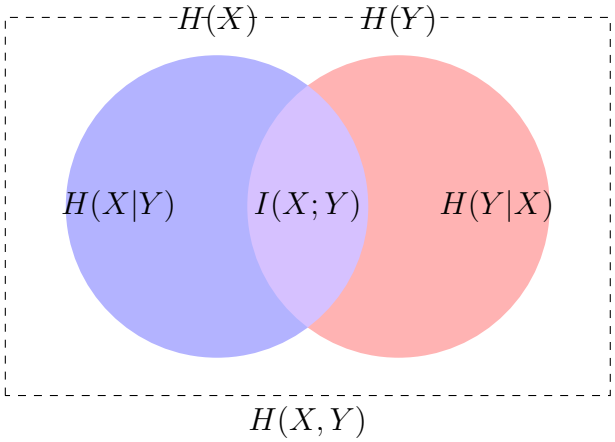


Figure 2: Venn Diagram of Information Measures

2.3 Relative Entropy (Kullback-Leibler Divergence)

A measure of the distance between two distributions $p(x)$ and $q(x)$:

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{7}$$

Properties:

- $D(p||q) \geq 0$ (Gibbs' Inequality).
- $D(p||q) = 0$ iff $p(x) = q(x)$.
- Not symmetric: $D(p||q) \neq D(q||p)$.

2.4 Fano's Inequality

Relates the probability of error P_e in estimating X given Y to the conditional entropy:

$$H(X|Y) \leq H(P_e) + P_e \log(K-1) \tag{8}$$

where K is the alphabet size of X.

2.5 Continuous Random Variables

- **Differential Entropy:** $h(X) = -\int p(x) \log p(x) dx$. Unlike discrete entropy, this can be negative and is coordinate-dependent.
- **Maximization:**
  - Peak limited $[a, b]$: Uniform distribution maximizes $h(X)$.
  - Power limited (σ^2): Gaussian distribution maximizes $h(X)$.
- **Entropy Power:** The power of a Gaussian variable with the same entropy as X.

$$\bar{\sigma}_x^2 = \frac{1}{2\pi e} e^{2h(X)} \leq \sigma^2 \tag{9}$$

For Gaussian $X \sim \mathcal{N}(0, \sigma^2)$, $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$.

3 Discrete Memoryless Source Coding

3.1 Lossless Coding

The goal is to represent the source sequence U^L of length L with a codeword b^N of length N using alphabet size D.

- Condition for unique decodability: $D^N \geq K^L$.
- Coding Rate $R = \frac{N}{L} \log D$.

3.1.1 Shannon's First Theorem (Source Coding)

For a source with entropy $H(U)$:

- If $R > H(U)$, the probability of error $P_e \to 0$ as $L \to \infty$.
- If $R < H(U)$, $P_e \to 1$.

The theorem relies on the **Asymptotic Equipartition Property (AEP)**, which states that for large L, the source almost always produces a sequence from the **Typical Set** $A_\epsilon^{(L)}$, where $p(u^L) \approx 2^{-LH(U)}$.

Typical set size is approximately $2^{LH(U)}$.

3.2 Variable-Length Coding

- **Kraft Inequality:** For a prefix code with lengths $n_1, \ldots, n_K$, $\sum D^{-n_i} \leq 1$.
- **Optimal Code Length:** $\bar{n} \geq \frac{H(U)}{\log D}$.

3.2.1 Coding Algorithms

1. **Huffman Coding:** Optimal prefix code. Built bottom-up by combining least probable symbols. Example: For probabilities 0.4, 0.3, 0.2, 0.1, code lengths 2,2,2,3. 2. **Shannon-Fano Coding:** Top-down splitting or using lengths $l_i = \lceil -\log p_i \rceil$. 3. **Arithmetic**

Coding: Maps a sequence to a sub-interval of $[0, 1)$. Efficient for long sequences, avoids tree structure. 4. **Lempel-Ziv:** Universal coding (dictionary-based), used in ZIP/GZIP. Variants like LZ77, LZ78.

3.3 Rate Distortion Theory

When compression is lossy ($R < H(U)$), we minimize rate R for a given distortion D.

$$R(D) = \min_{p(\hat{x}|x):E[d(X,\hat{X})]\leq D} I(X;\hat{X}) \tag{10}$$

- **Binary Source (Hamming Distortion):** $R(D) = 1 - H(D)$ for $0 \leq D \leq \min(p, 1-p)$.
- **Gaussian Source (Squared Error):** $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$ for $D \leq \sigma^2$.

The rate-distortion function is convex and decreasing.

4 Channel Coding

4.1 Error-Correcting Codes

Channel coding adds controlled redundancy to detect and correct errors introduced by the channel.

4.1.1 Hamming Codes

A linear block code with minimum distance 3, capable of correcting 1 error. For (7,4) Hamming code, parity check matrix defines error positions.

4.1.2 Convolutional Codes

Tree-based codes with memory, decoded using Viterbi algorithm.

4.1.3 Modern Codes

- **LDPC Codes:** Low-Density Parity-Check codes, near-Shannon limit performance with iterative decoding. - **Turbo Codes:** Parallel concatenated convolutional codes, also near-capacity.

Shannon's second theorem (channel coding theorem) states that reliable communication is possible at rates below capacity C, with error probability approaching 0 as block length increases.

5 Channel Capacity

5.1 Discrete Memoryless Channel (DMC)

Defined by input X, output Y, and transition matrix $P(Y|X)$. **Channel Capacity** is the maximum mutual information:

$$C = \max_{p(x)} I(X;Y) \tag{11}$$

Computation often requires optimization, e.g., Blahut-Arimoto algorithm.

5.2 Standard Channels

5.2.1 Binary Symmetric Channel (BSC)

Bit flip probability p.

$$C_{BSC} = 1 - H(p) \tag{12}$$

5.2.2 Binary Erasure Channel (BEC)

Erasure probability ϵ.

$$C_{BEC} = 1 - \epsilon \tag{13}$$

Figure 3: Channel Models

Other channels: Z-channel, where only 0 flips to 1 with probability p, $C = log(1 + 2^{h(p)/(1-p)}(1-p))$.

5.3 Gaussian Channel Capacity

For an Additive White Gaussian Noise (AWGN) channel with bandwidth W, signal power P, and noise power spectral density N_0:

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \quad \text{bits/sec} \tag{14}$$

This is the Shannon-Hartley Theorem.

For infinite bandwidth, C approaches $P/(N_0 ln 2)$.

6 Theory of Computation

6.1 Models of Computation

6.1.1 Turing Machine

The theoretical foundation of modern computing. It consists of:

- An infinite tape (memory).
- A read/write head.
- A state register (q_i).

- A finite table of instructions (transition function): Given state q_i and symbol S_j, write S_k, move $L/R/N$, and go to state q_{next}.

Halting Problem: Determining if a program will finish execution is undecidable.

Universal Turing machine can simulate any other.

6.1.2 Von Neumann Architecture

The practical architecture for computers:

- Stored Program: Instructions and data share the same memory.
- Components: CPU (ALU + Control Unit), Memory, Input/Output.
- Bottleneck: Separation between CPU and memory affects speed.

Harvard architecture separates instruction and data memory.

6.2 Computational Complexity

- **P (Polynomial):** Problems solvable in polynomial time $O(n^k)$.
- **NP (Nondeterministic Polynomial):** Problems verifiable in polynomial time.
- **NP-Complete:** The hardest problems in NP (e.g., Traveling Salesman). If any NP-Complete problem is in P, then P=NP.

Other classes: EXP, PSPACE.

6.3 Kolmogorov Complexity

The Kolmogorov complexity $K(x)$ of a string x is the length of the shortest program that outputs x.

- $K(x) \approx \text{length}(x)$ for random strings (incompressible).
- $K(x) \ll \text{length}(x)$ for structured strings (compressible).

Incomputable, but upper bounds via compression.

7 Control Systems

7.1 System Classification

- **Open Loop:** Control action is independent of output.
- **Closed Loop (Feedback):** Control action depends on the error (Difference between reference and output).

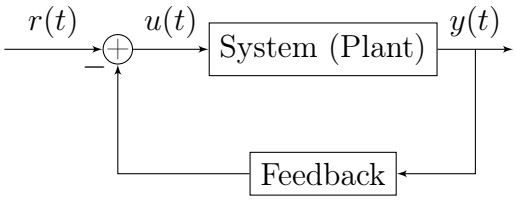


Figure 4: Feedback Control System

Advantages of feedback: Reduced sensitivity to disturbances, improved stability.

7.2 Modeling Representations

7.2.1 Transfer Function

Used for Linear Time-Invariant (LTI) systems in the frequency domain (Laplace Transform).

$$G(s) = \frac{Y(s)}{U(s)} = \frac{N(s)}{D(s)} \tag{15}$$

- **Poles:** Roots of $D(s) = 0$ (determine stability).
- **Zeros:** Roots of $N(s) = 0$.

Example: Second-order system $G(s) = \frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$.

7.2.2 State-Space Representation

Time-domain approach, valid for MIMO and non-linear systems.

$$\dot{x}(t) = Ax(t) + Bu(t) \quad \text{(State Equation)} \tag{16}$$
$$y(t) = Cx(t) + Du(t) \quad \text{(Output Equation)} \tag{17}$$

Convertible to transfer function via $G(s) = C(sI - A)^{-1}B + D$.

7.3 Controllability and Observability

- **Controllability:** Can the input $u(t)$ drive the state x from any initial to any final state?
- **Observability:** Can the internal state x be determined by observing output $y(t)$?
- **Kalman Rank Conditions:**

$$\text{Rank}[B, AB, \ldots, A^{n-1}B] = n \quad \text{(Controllable)} \tag{18}$$

$$\text{Rank}[C^T, A^TC^T, \ldots, (A^T)^{n-1}C^T] = n \quad \text{(Observable)} \tag{19}$$

7.4 Stability

- **BIBO Stability:** Bounded Input results in Bounded Output.
- **Routh-Hurwitz Criterion:** Determines number of unstable poles without solving the characteristic equation. For characteristic equation $a_n s^n + \cdots + a_0 = 0$, construct array and count sign changes.
- **Lyapunov Stability:**
  - Define a scalar "energy" function $V(x) > 0$.
  - If $\dot{V}(x) < 0$ along system trajectories, the system is asymptotically stable.

For LTI, stable if all poles have negative real parts.

8 Multi-user Information Theory (Book-Only Content)

8.1 Multi-user Source Coding

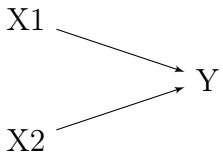
In multi-user scenarios, multiple sources may be correlated or independent. The goal is to compress them jointly or separately while minimizing rate.

- **Slepian-Wolf Theorem**: For two correlated sources X and Y, the total rate for lossless compression is $R_X + R_Y \geq H(X, Y)$, with individual rates $R_X \geq H(X|Y)$, $R_Y \geq H(Y|X)$. - **Distributed Compression**: Sources encode independently but decode jointly. Applications in sensor networks.

8.2 Multi-user Channel Capacity

For channels with multiple senders or receivers.

- **Multiple Access Channel (MAC)**: Capacity region is the set of rates (R_1, R_2) where $R_1 \leq I(X_1; Y|X_2)$, $R_2 \leq I(X_2; Y|X_1)$, $R_1 + R_2 \leq I(X_1, X_2; Y)$. - **Broadcast Channel**: One sender, multiple receivers. Capacity depends on channel type (degraded or not).



MAC

Figure 5: Multiple Access Channel

9 Network Coding (Book-Only Content)

9.1 Basics of Network Coding

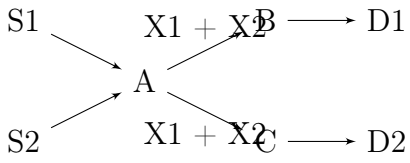
Traditional routing forwards packets, but network coding combines them at intermediate nodes to increase throughput.

- **Butterfly Network Example**: Two sources send to two sinks via a bottleneck; coding allows both to receive data simultaneously. - **Linear Network Coding**: Over finite fields, nodes compute linear combinations of inputs.

9.2 Random Network Coding

Random coefficients for coding; probabilistic success in decoding.

- **Max-Flow Min-Cut Theorem Extension**: Network coding achieves the min-cut capacity in multicast scenarios.



Butterfly Network

Figure 6: Network Coding Example

10 Information Security and Cryptography (Book-Only Content)

10.1 Symmetric Cryptography

Uses same key for encryption/decryption.

- **DES/AES**: Block ciphers; AES uses substitution-permutation network. - **Stream Ciphers**: RC4, generate keystream XOR with plaintext.

10.2 Public-Key Cryptography

Different keys for encryption/decryption.

- **RSA**: Based on factoring large numbers; public key (n, e), private d. - **Entropy in Cryptography**: Keys should have high entropy to resist attacks.

Properties: Confidentiality, integrity, authentication.

11 Quantum Information Theory (Book-Only Content)

11.1 Qubits and Entanglement

Qubit: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, with $|\alpha|^2 + |\beta|^2 = 1$.

- **Entropy**: Von Neumann entropy $S(\rho) = -\text{Tr}(\rho \log \rho)$. - **Entanglement**: Bell states, measure correlation beyond classical.

11.2 Quantum Channels

Noisy channels in quantum; capacity for classical/quantum information transmission.

- **Quantum Teleportation**: Uses entanglement to transmit qubit state.

12 Applications of Information Theory in Bioinformatics (Book-Only Content)

12.1 Sequence Alignment

Measure similarity using mutual information or entropy.

- **DNA Compression**: Use entropy to compress genomes, identifying repeats.

12.2 Phylogenetic Trees

Use information criteria (e.g., AIC) to select best tree models.

Applications: Gene prediction, protein structure.

13 Appendix (Book-Only Content)

13.1 Probability Basics

Review of probability, random variables, distributions.

13.2 Convex Optimization

Jensen's inequality, used in proofs of information inequalities.