Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 21, Number 3 (2025), pp. 560-564

http://www.ripublication.com/gjpam.htm

Mathematics of Information Theory

Narinder
Associate Professor in Mathematics
naridermohanikamath@gmail.com
SPPND GDC Samba, University of Jammu, INDIA

Abstract

Information theory, founded by Claude Shannon, provides a mathematical framework for quantifying information, uncertainty, and communication efficiency. This paper presents the foundational concepts of information theory including entropy, mutual information, and channel capacity, along with rigorous proofs of key theorems.

Introduction

Information theory studies the transmission, processing, and quantification of information. It is grounded in probability theory and has applications in data compression, cryptography, and communication systems.

Entropy

2.1 Definition

Let X be a discrete random variable with probability mass function p(x). The entropy of X is defined as:

$H(X) = -\sum p(x) \log p(x)$

Entropy measures the average uncertainty in the outcome of X.

2.2 Properties

- $H(X) \geq 0$
- $H(X) = 0$ if and only if X is deterministic
- $H(X)$ is maximized when X is uniformly distributed

2.3 Maximum Entropy

Let X take values in a finite set $\{x_1, ..., x_n\}$.

Claim: $H(X)$ is maximized when $p(x_i) = 1/n$ for all i.

Proof: Using Jensen's inequality for the concave function $f(p) = -p \log p$:

$H(X) = -\sum p(x_i) \log p(x_i) \leq -\sum (1/n) \log (1/n) = \log n$

Equality holds when $p(x_i) = 1/n$ for all i.

Joint and Conditional Entropy
3.1 Joint Entropy
For random variables X and Y
$H(X, Y) = -\sum p(x, y) \log p(x, y)$

3.2 Conditional Entropy
$H(Y|X) = -\sum p(x, y) \log p(y|x)$

3.3 Chain Rule
$H(X,Y) = H(X) + H(Y|X)$
Proof: $H(X, Y) = -\sum p(x, y) \log p(x, y)$
$= -\sum p(x, y) \log[p(x)p(y|x)]$
$= -\sum p(x, y) \log p(x) - \sum p(x, y) \log p(y|x)$
$= -\sum p(x) \log p(x) - \sum p(x, y) \log p(y|x)$
$= H(X) + H(Y|X)$

Mutual Information
4.1 Definition
$I(X;Y) = \sum p(x, y) (\log (p(x, y) / p(x)p(y))$
Mutual information quantifies the amount of information shared between X and Y.

4.2 Alternative Form
$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
Proof: From the chain rule: $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
Then: $I(X;Y) = H(X) + H(Y) - H(X,Y)$

Kullback-Leibler Divergence
5.1 Definition
For distributions P and Q over the same support:
$D_{KL}(P\|Q) = \sum P(x) \log [P(x) / Q(x)]$
5.2 Properties
• $D_{KL}((P\|Q) \geq 0$
$D_{KL}(P\|Q) = 0$ iff P = Q
Proof: By Gibbs' inequality
: $\sum P(x) \log (P(x) / Q(x)) \geq 0$
with equality iff $P(x) = Q(x) \forall x$

Channel Capacity
6.1 Definition
Let X be the input and Y the output of a communication channel. The channel capacity C is defined as:
$C = \max_{p(x)} I(X;Y)$

7 Shannon's Noisy Channel Coding Theorem

7.1 Definition:

Let X and Y be finite input and output alphabets, and p(y|x) define a discrete memoryless channel DMC. The channel capacity C is defined as:

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

where I(X;Y) is the mutual information between input X and output Y.

7.2 Shannon's Noisy Channel Coding Theorem

Theorem Statement: Theorem: For any rate R < C, there exists a sequence of codes with block length n and rate R such that the probability of decoding error tends to zero as n → ∞. Conversely, for any rate R > C, the probability of error is bounded away from zero for all codes.

Proof: Fix p(x) such that $C = max_{p(x)}\, I(X;Y)$. Generate M = 2^{nR} code words $\{x^n(1), x^n(2), ..., x^n(M)\}$ independently according to $\prod_1^n p(x_i)$.

To send message m ∈ {1, ..., M}, we transmit codeword $x^n(m)$.

Upon receiving y^n, the decoder searches for a unique m such that $(x^n(m), y^n)$ is jointly typical. If none or more than one such m exists, we declare an error.

Now we do Error Analysis

Let P_e be the average probability of error. We bound P_e by considering two types of errors:

• Type I Error: $(x^n(m), y^n)$ is not jointly typical.

• Type II Error: There exists $m' \neq m$ such that $(x^n(m'), y^n)$ is jointly typical.

By the Asymptotic Equipartition Property (AEP), the probability of Type I error tends to zero as n → ∞.

For Type II error, using the union bound:

$$P_e^{(2)} \leq \sum_{m \neq m'} \Pr\left(\left(x^n(m'),\ y^n\right) Jointly\ Typical\right)$$

Each term is bounded by $2^{-n(I(X;Y)-\epsilon)}$ and there are M − 1 such terms:

$$P_e^{(2)} \leq (M-1)\, 2^{-n(I(X;Y)-\epsilon)} = 2^{nR} 2^{-n(I(X;Y)-\epsilon)} = 2^{-n(I(X;Y)-R-\epsilon)}$$

If R < I(X; Y), then $P_e^{(2)} \to 0$ as n → ∞.

Thus, for any R < C, there exists a code with vanishing error probability.

Conversly, we Assume a sequence of (n, M) codes with vanishing error probability. Let W be the message, X^n the codeword, and Y^n the received sequence.

By Fano's inequality: $H(W|Y^n) \leq 1 + P_e Log\ M$

Then: $I(W;Y^n) = H(W) - H(W|Y^n) \geq -1 - P_e Log\ M$

Since $W \to X^n \to Y^n$ forms a Markov chain:

$$I(W;Y^n) \leq I(X^n;Y^n) = \sum_{i=1}^{n} I\left(X_i;Y_i\right) \leq nC$$

From these above equations we find that:

$$Log\ M \leq nC + 1 + P_e Log\ M$$

As $P_e \rightarrow 0$, we get:

$$\frac{Log\ M}{n} \leq C + \frac{1}{n}$$

Thus, any achievable rate R satisfies R ≤ C

Applications of Information Theory

8.1 Data Compression

Information theory helps determine the limits of lossless and lossy compression. Entropy quantifies the minimum average number of bits needed to encode a source. Huffman coding and arithmetic coding are practical algorithms that approach this limit.

Example: ZIP and PNG formats use entropy-based compression to reduce file sizes without losing information

8.2 Error Detection and Correction

Information theory enables reliable communication over noisy channels. Shannon's noisy-channel coding theorem guarantees that error correcting codes can achieve reliable transmission up to the channel capacity. Techniques like Hamming codes and Reed–Solomon codes are based on this principle.

Example: Used in satellite communication, mobile networks, and QR codes.

8.3 Cryptography

Information theory provides tools to analyze and design secure communication systems. Entropy measures the uncertainty of keys and messages. Shannon's theory of secrecy systems shows that perfect secrecy requires the key entropy to match the message entropy. Example: One-time pad encryption achieves perfect secrecy under Shannon's criteria

8.4 Machine Learning

Information-theoretic measures guide model selection and feature relevance. Mutual information quantifies the dependency between features and target variables. It is used in decision trees, clustering, and dimensionality reduction.

Example: Entropy and information gain are used to split nodes in decision trees like ID3 and C4.5

8.5 Neuroscience

Information theory helps analyze neural coding and information flow in biological systems. Neurons transmit signals that can be modeled as stochastic processes. Mutual information measures how much information a neuron conveys about stimuli.

Example: Used to study sensory systems and brain–computer interfaces.

8.6 Genomics
Information theory aids in analyzing genetic sequences and biological data. Entropy measures the variability in DNA sequences. Mutual information helps detect correlations between genes or regulatory elements.
Example: Used in motif discovery and sequence alignment algorithms.

8.7 Linguistics and Natural Language Processing
Information theory models language structure and communication efficiency. Entropy measures the predictability of language. Zipf's law and Shannon's experiments quantify redundancy and information content in text.
Example: Used in text compression, speech recognition, and language modeling

8.8 Image/Video Processing
Information theory guides abstraction, compression, and similarity detection. Entropy and mutual information are used to extract key frames, detect patterns, and compress multimedia data.
Example: Video summarization algorithms use mutual information to identify representative frames

8.9 Statistical Inference
Information theory provides criteria for model selection and hypothesis testing. Kullback–Leibler divergence measures the difference between probability distributions. It is used in Bayesian inference and regularization.
Example: Akaike Information Criterion (AIC) is derived from in formation theory

8.10 Physics
Information theory connects entropy with physical systems. Statistical mechanics uses entropy to describe disorder. Landauer's principle links information erasure to heat dissipation.
Example: Used in quantum computing and black hole thermodynamics

References
[1] C.E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, 1948.
[2] R.M. Gray, Entropy and Information Theory, Springer, 2011. https://ee.stanford.edu/~gray/it.pdf
[3] A.Ya. Khinchin, Mathematical Foundations of Information Theory, Dover Publications, 1957. https://archive.org/details/khinchin-mathematical-foundations-of-information-theory
[4] S. Schäffler, Mathematics of Information: Theory and Applications of Shannon-Wiener Information, Springer, 2024. https://link.springer.com/book/10.1007/978-3-662-69102-1