Course Report: Fundamental Concepts and Advanced Insights in Information Theory and statistics

邓愉静

July 12, 2025

1 Introduction to Information Theory Course Framework

This course, Information Theory and statistics, delves into the mathematical principles governing information processing, communication, and statistical inference. Spanning from basic information measures to advanced hypothesis testing with large deviations theory, it equips learners to analyze and design systems that efficiently handle data. By integrating theoretical bounds, such as entropy for compression limits, with practical algorithms like Huffman coding, the course bridges abstract concepts with real - world applications. This report synthesizes key knowledge, emphasizing the interconnections between information measures, lossless data compression, and hypothesis testing, while exploring their asymptotic behaviors.

2 Information Measures: The Building Blocks

2.1 Shannon Entropy and Its Properties

For a discrete random variable X with probability mass function P_X, the Shannon entropy is defined as:

$$H(X) = -\mathbb{E}\left[\log P_X(X)\right] = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

It serves as a fundamental measure of uncertainty associated with X.

- **Positivity and Bounds**: Entropy is non - negative, $H(X) \geq 0$, with equality if and only if X is a deterministic variable (i.e., one outcome has probability 1). For a random variable X taking values in a finite alphabet $\mathcal{X}$ with $|\mathcal{X}| = M$, $H(X) \leq \log M$. Equality holds when X is uniformly distributed over $\mathcal{X}$, representing the maximum uncertainty.

- **Chain Rule**: For two random variables X and Y, the joint entropy $H(X, Y)$ satisfies the chain rule: $H(X, Y) = H(X) + H(Y|X)$. This rule extends to multiple random variables, allowing us to decompose the uncertainty of a joint distribution into a sum of conditional entropies. It is crucial for analyzing sequential data, such as time - series or data streams in communication channels.

2.2 Kullback - Leibler Divergence and Mutual Information

2.2.1 Kullback - Leibler (KL) Divergence

The KL divergence between two probability distributions P and Q (where $P \ll Q$, meaning P is absolutely continuous with respect to Q) is given by:

$$D(P\|Q) = \mathbb{E}_Q \left[\frac{dP}{dQ} \log \frac{dP}{dQ} \right] = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad \text{(for discrete case)}$$

It quantifies the "distance" or dissimilarity between two distributions. A key property is the information inequality $D(P\|Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere. This divergence is fundamental in various applications, including model selection, where it helps compare a proposed model distribution with the true data distribution.

2.2.2 Mutual Information

Mutual information $I(X; Y)$ measures the dependence between two random variables X and Y. It is defined as:

$$I(X; Y) = D(P_{X,Y} \| P_X P_Y)$$

Equivalently, it can be expressed using entropies: $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Mutual information is symmetric, $I(X; Y) = I(Y; X)$, and non

- negative. If X and Y are independent, $I(X;Y) = 0$. In communication systems, it represents the amount of information that can be transmitted from one variable to another, forming the basis for concepts like channel capacity.

3 Lossless Data Compression: Theory and Practice

3.1 Source Coding Theorems

3.1.1 Asymptotic Equipartition Property (AEP)

For a discrete memoryless source generating i.i.d. samples $X^n = (X_1, \cdots, X_n)$, the AEP states that:

$$-\frac{1}{n} \log P_{X^n}(X^n) \xrightarrow{\text{a.s.}} H(X) \quad \text{as } n \to \infty$$

Most sequences (called typical sequences) have a probability approximately equal to $2^{-nH(X)}$, and the number of typical sequences is approximately $2^{nH(X)}$. This property implies that to losslessly compress the source, we need at least $nH(X)$ bits to represent these typical sequences, establishing $H(X)$ as the theoretical lower bound for lossless compression.

3.1.2 Shannon Source Coding Theorem

This theorem formalizes the AEP result. For any $\epsilon > 0$, there exists a prefix code (a uniquely decodable code) such that the average length of the code for encoding n samples is at most $H(X) + \epsilon$ for sufficiently large n. Conversely, no prefix code can have an average length less than $H(X)$. This theorem provides a fundamental limit for lossless data compression, guiding the design of efficient compression algorithms.

3.2 Coding Schemes

3.2.1 Huffman Coding

Huffman coding is an algorithm for constructing optimal prefix codes for a given probability distribution. It works by repeatedly merging the two least probable symbols into a new symbol with their combined probability, until a single

tree is formed. The resulting code assigns shorter bitstrings to more probable symbols, achieving an average length within 1 bit of the entropy $H(X)$. It leverages the Kraft - McMillan inequality, which states that a set of codewords is a prefix code if and only if the sum of their codeword lengths' probabilities (in the form of 2^{-l_i}, where l_i is the length of codeword i) is less than or equal to 1.

- **Kraft - McMillan Inequality**: For a set of codewords with lengths $l_1, l_2, \cdots, l_k$, $\sum_{i=1}^{k} 2^{-l_i} \leq 1$ if and only if the code is a prefix code. This inequality is a key tool in analyzing the validity and efficiency of Huffman codes.

3.2.2 Universal Compression

Universal compression algorithms, such as the Lempel - Ziv algorithm, are designed to work without prior knowledge of the source distribution. For a stationary ergodic source with entropy rate $H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n)$, these algorithms guarantee that the average length per symbol of the encoded data approaches $H(\mathbb{X})$ as the number of samples n becomes large. This makes them suitable for real - world applications where the source distribution is unknown or changing, such as compressing text files or network traffic.

- **Stationary Ergodic Source**: A source is stationary if its statistical properties do not change over time, and ergodic if the time average of any sample function is equal to the ensemble average. These properties ensure that the entropy rate $H(\mathbb{X})$ is well - defined and can be approached by universal compression algorithms.

4 Hypothesis Testing: Error Exponents and Large Deviations

4.1 Neyman - Pearson Lemma and Optimal Tests

For simple hypotheses $H_0 : P$ vs. $H_1 : Q$, the Neyman - Pearson lemma is a cornerstone of hypothesis testing. The likelihood ratio test is defined as:

$$\phi(X^n) = \begin{cases} 1 & \text{if } \frac{1}{n} \sum_{i=1}^{n} \log \frac{Q(X_i)}{P(X_i)} \geq \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where τ is a threshold. This test is most powerful for a fixed type - I error probability $\alpha = \mathbb{P}(\phi = 1 \mid H_0)$. In other words, among all possible tests with type - I error probability α, the Neyman - Pearson test minimizes the type - II error probability $\beta = \mathbb{P}(\phi = 0 \mid H_1)$. This optimality result has far - reaching implications in statistical inference, from simple binary detection problems to complex model selection tasks in machine learning.

- **Type - I and Type - II Errors**: Type - I error is the probability of rejecting the null hypothesis H_0 when it is true, and type - II error is the probability of accepting H_0 when the alternative hypothesis H_1 is true. The Neyman - Pearson lemma balances these two types of errors for optimal testing.

4.2 Error Exponents and Large Deviations

4.2.1 Error Exponent Analysis

Error exponents characterize the exponential decay of error probabilities as the number of samples $n \to \infty$. For type - I error α_n and type - II error β_n, we are interested in the limits:

$$\lim_{n\to\infty} -\frac{1}{n}\log \alpha_n, \quad \lim_{n\to\infty} -\frac{1}{n}\log \beta_n$$

These exponents provide a way to compare different hypothesis testing procedures in terms of their asymptotic performance. A larger error exponent implies a faster decay of the error probability with increasing n, which is desirable.

4.2.2 Chernoff Bound

The Chernoff bound gives the optimal exponent for symmetric testing. It is defined as:

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\text{error}) = \min_{0\le\lambda\le 1} \left(-\log \mathbb{E}\left[P(X)^{1-\lambda}Q(X)^{\lambda}\right]\right)$$

This bound connects hypothesis testing to the theory of large deviations, where the rate function (such as the KL divergence) quantifies the exponent of rare events. The Chernoff bound is widely used in analyzing the performance of communication systems, where it helps determine the reliability of signal detection in the presence of noise.

- **Rate Function in Large Deviations**: The rate function, often related to the KL divergence in the context of hypothesis testing, determines how likely it is to observe rare events (deviations from the expected behavior). A lower rate function value for a particular deviation implies a higher probability of that deviation occurring.

4.3 Sanov's Theorem: Rare Event Probabilities

Sanov's theorem generalizes the results to composite hypotheses. It states that the probability of the empirical distribution $\hat{P}_n$ deviating from the true distribution P is dominated by the KL divergence to the "closest" distribution in the alternative hypothesis set. Mathematically,

$$\mathbb{P}\left(\hat{P}_n \in \mathcal{A} \mid P\right) \asymp \exp\left(-n \min_{Q \in \mathcal{A}} D(Q\|P)\right) \quad \text{as } n \to \infty$$

This theorem provides a unified framework for analyzing rare events, which is not only useful in hypothesis testing but also in anomaly detection, where we need to identify rare patterns that deviate from the normal behavior of a system.

- **Composite Hypotheses**: In contrast to simple hypotheses (where both H_0 and H_1 specify a single distribution), composite hypotheses involve sets of distributions. Sanov's theorem allows us to handle such complex scenarios by considering the minimum KL divergence within the alternative hypothesis set.

5 Interconnections and Course Insights

The course highlights the deep interconnections between different topics:

- **Entropy and Compression**: Entropy, as a measure of uncertainty, provides the theoretical lower bound for lossless data compression. Algorithms like Huffman coding and universal compression schemes strive to approach this bound, demonstrating the practical significance of entropy in data processing.

- **Mutual Information and Communication**: Mutual information quantifies the amount of information that can be transmitted between two variables, forming the basis for channel capacity in communication systems.

It helps in designing efficient communication protocols that maximize the transfer of information while minimizing errors.

- **KL Divergence and Hypothesis Testing**: The KL divergence plays a central role in hypothesis testing, quantifying the distinguishability between different distributions. Error exponents, analyzed using large deviations theory, are closely related to the KL divergence, providing a way to assess the performance of hypothesis tests asymptotically.

By understanding these interconnections, learners can develop a holistic view of information theory and its applications in various fields, including communication engineering, data science, and machine learning.

6 Conclusion and Future Directions

The course, Information Theory and statistics, covers a rich set of fundamental concepts in information theory, from information measures to lossless data compression and hypothesis testing. These concepts form the bedrock for analyzing and designing efficient information processing systems.

In future directions, the course is likely to extend these ideas to:

- **Channel Coding**: Analyzing reliable communication over noisy channels, building on the concepts of entropy and mutual information to derive bounds on channel capacity and design error - correcting codes.

- **Information Theory in Machine Learning**: Applying information - theoretic concepts to problems such as model selection, generalization bounds in learning algorithms, and understanding the information content in neural network representations.

These extensions will further demonstrate the versatility of information theory in addressing modern challenges in technology and data - driven disciplines.