Lecture 2. Lossless Source Coding

Lin Zhang

Tsinghua-Berkeley Shenzhen Institute

Shenzhen, China, 2017

Outline

The definition of AEP

- **Theorem 2.1** If X_1,X_2,...,X_n are i.i.d random variables, generated by the distribution of $p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \longrightarrow H(X) \tag{1}$$

  in probability.
- The empirical mean converges to the expected value of $-\log p(x)$, i.e, $H(X) = \mathbb{E}\left[\log \frac{1}{p(x)}\right]$.

Typical sequence

- **Definition 2.1** A sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ generated by $X \sim p(x)$, the typical set $\mathcal{A}_\epsilon^n$ is a collection of sequences $x_1, x_2, \ldots, x_n$ on the condition that

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}. \quad (2)$$

- The sequences belonging to the typical set $\mathcal{A}_\epsilon^{(n)}$ are typical sequences.
- There are other definitions of typicality, including Strong Typicality, and Robust Typicality.

Properties of a typical sequence

- **Theorem 2.2** The properties of a typical sequence are listed as follows
  1. If $\mathbf{x} \in \mathcal{A}_\epsilon^n$, then $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$.
  2. $Pr\{\mathcal{A}_\epsilon^{(n)}\} > 1 - \epsilon$, if n is sufficiently large.
  3. $| \mathcal{A}_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.
  4. $|\mathcal{A}_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$
- Note that $|A|$ denotes the cardinality of the set A, i.e, the number of elements in this set.

Outline

Shannon Coding with AEP

- **Theorem 2.3** A sequence X^n is generated by i.i.d discrete probability distribution $X \sim p(x)$. $\forall \epsilon \geq 0$, there exists n and a mapping from X^n to a binary code, satisfying

$$\mathbb{E}\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \epsilon, \tag{3}$$

where $l(X^n)$ is the length of the binary codewords for X^n.

Outline

The definition

- **Definition 2.2** The entropy rate of a random process $\{X_i\}_{i=1}^{n}$ is defined as

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n), \quad (4)$$

when the limit exists.

- An alternative definition is

$$H^{'}(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1). \quad (5)$$

The entropy rate of stationary sources

- **Theorem 2.5** For stationary sources, the conditional entropy $H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$ is monotonically non-increasing function of n, and the limit $H^{'}(\mathcal{X})$ exists when n goes to infinity.
- **Theorem 2.6** The two definitions of entropy rate are equivalent for stationary sources.

Discrete markov process

- **Definition 2.3** $\forall x_1, x_2, \ldots, x_n \in \mathcal{X}$, a discrete stochastic process $X_1, X_2, \ldots, X_n$ forms a markov chain on the condition that

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x_{n+1} | X_n = x_n). \quad (6)$$

- **Definition 2.4** The markov process is time-invariant, which implies the transition probability

$$\Pr(X_{n+1} = b | X_n = a) = \Pr(X_2 = b | X_1 = a) \quad (7)$$

holds for $\forall n = 1, 2, \ldots$ and $\forall a, b$.

The entropy rate of Markovian processes

- **Theorem 2.7** The entropy rate of a stationary markov process equals

$$H(\mathcal{X}) = H^{'}(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1) = H(X_2|X_1). \qquad (8)$$

The redundancy of sources

- The maximum entropy $H(\mathcal{X}) \leq \log K$ holds, where K is the number of alphabets.
- **Definition 2.5** The redundancy of sources is defined as $\log K - H(X)$, and the relative redundancy equals $1 - \frac{H(X)}{\log K}$.

Outline

Average coding length

- **Definition 2.6** The assigned codeword C representing a random variable X is considered as a mapping from the set $\mathcal{X}$ to $\mathcal{C}$, where $\mathcal{C}$ is a collection of D-ary codewords with finite-length. Note that $C(x)$ is the codeword assigned to x and $l(x)$ is the length of this codeword.
- The average length of codewords $C(x)$ for a random variable X is defined as

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x), \tag{9}$$

i.e., the expected code length.

Non-singular codes and unique decoding

- **Definition 2.7 Non-singular code** is the code that the codeword C maps from the set $\mathcal{X}$ to the codeword space $\mathcal{C}$ on the condition that

$$x \neq x^{'} \Rightarrow C(x) \neq C(x^{'}). \tag{10}$$

- **Definition 2.8 Extension of code**. As a mapping to a finite-length D-ary sequence, the codeword C^* is the extension of C on the condition that

$$C(x_1, x_2, \ldots, x_n) = C(x_1)C(x_2)\cdots C(x_n). \tag{11}$$

- **Definition 2.9 Uniquely decodable codes** The codeword C is uniquely decodable when its extension is non-singular.
- **Definition 2.10 Prefix code** is a code set that there does not exist a codeword that is the prefix of others.

Kraft inequality

- **Theorem 2.8** The D-ary prefix codes satisfies

$$\sum_i D^{-l_i} \leq 1 \tag{12}$$

  with lengths $l_1, l_2, \ldots, l_m$.
- Conversely, there exists such prefix codes if the code lengths satisfies the inequality above.

The length of prefix codes

- **Theorem 2.9** The average code length of D-ary prefix codes for a random variable X satisfies

$$L \geq H_D(X), \tag{13}$$

and the equality holds if and only if $D^{-l_i} = p_i$.

- **Theorem 2.10** To encode a random variable X using D-ary prefix codes, the optimal coding length satisfies

$$H_D(X) \leq L^* < H_D(X) + 1. \tag{14}$$

Coding length using large blocks

- **Theorem 2.11** The minimum expected coding length per symbol satisfies

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq L_n^* < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}. \tag{15}$$

- If the source is a stationary random process, the minimum expected length $L_n^* \to H(\mathcal{X})$ as $n \to \infty$.
- **Theorem 2.12** If the codeword length obeys $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ to encode $X \sim p(x)$, the average codeword length satisfies

$$H(p) + D(p\|q) \leq E_p l(x) < H(p) + D(p\|q) + 1. \tag{16}$$

The length constraint on uniquely decodable codewords

- **Theorem 2.13** The codeword length $l_1, l_2, \ldots, l_m$ satisfies the Kraft inequality

$$\sum_i D^{-l_i} \leq 1 \tag{17}$$

  for any D-ary codes that are uniquely decodable.
- Conversely, there exists such a code with codeword length $l_1, l_2, \ldots, l_m$ that satisfies Kraft inequality.

Huffman coding

- Shannon codes with $l_i = \lceil \log \frac{1}{p_i} \rceil$ is not optimal.
- Huffman codes proposed by David A. Huffman is optimal with the minimum average coding length, provided with the prior distribution of sources.
- **Lemma 2.14** There exists an optimal prefix code for a given probability distribution, which satisfies the following conditions.
  1. If $p_j > p_k$, then $l_j \leq l_k$.
  2. The lengths of the longest codewords are the same.
  3. There is only one-digit difference at last between the longest codewords, corresponding to the symbols with the minimum probabilities.

The optimality of Huffman coding

- **Theorem 2.15** Huffman coding is optimal, i.e., if C^* is a Huffman code and there exists $C^{'}$ that is uniquely decodable, the average coding length

$$L(C^*) \leq L(C^{'}) \tag{18}$$

  holds given prior distribution of sources.
- Remarks:
  1. Huffman coding is a greedy algorithm, constructing the Huffman tree.
  2. Luckily and interestingly, the local optimality leads to the global optimal solution.

Shannon-Fano-Elias codes

- **Theorem 2.16** The average coding length of Shannon-Fano-Elias codes is greater than $H(X) + 1$ and less than $H(X) + 2$.
- S-F-E codes can be applied to encode sequences. The key ingredient is to calculate the c.d.f and assign proper codewords.
- Arithmetic coding is the extension of S-F-E codes and is well-known for its practical applications in FAX, JPEG and JPEG 2000, etc.

Lempel-Ziv coding

- Due to the difficulties to obtain the prior distribution of sources as Huffman coding, Lempel-Ziv coding, e.g., LZ-78, is proposed to address this issue.
- **Theorem 2.20** Let $\{X_i\}_{i=1}^n$ be an ergodic binary sequence with entropy rate $H(\mathcal{X})$, LZ-78 satisfies

$$\limsup_{n\to\infty} \frac{c(n)\log c(n)}{n} \leq H(\mathcal{X}) \tag{19}$$

with probability one, where $c(n)$ is the number of partitions of sequence $\{X_i\}_{i=1}^n$.

Outline

1. Asymptotic Equipartition Property (AEP)
2. Fixed-length coding theorem
3. Entropy rate
4. Variable-length coding
5. Summary of Lecture 2

Summary

- Asymptotic equilpartitioning property (AEP)
- Typical sequences and typical set
- Fixed-length coding theorem
- Types of codes
- Kraft inequality
- Shannon codes and the optimal coding theorem
- Huffman coding and its optimality
- Shannon-Fano-Elias codes and LZ-78 universal coding.

Reference

- *Lin Zhang*, "Lecture notes on Fundamentals of applied information theory", *2014-spring, in Chinese.*
- *Claude E. Shannon*, "A Mathematical Theory of Communication", *Bell System Technical Journal, 1948.*
- *Cover T M, Thomas J A.*, "Elements of information theory", *John Wiley and Sons, 2012.*
- *Xuelong Zhu*, "Fundamentals of applied information theory", *Tsinghua Univ. Press, 2001, in Chinese.*
- *David A. Huffman*, "A method for the construction of minmum-redundancy codes", *Proceedings of the I.R.E., Sep 1952, pp. 1098-1102.*