Chapter 4

Block Source Coding

Discussion about general sources

The compression (or representation) of a sequence of symbols from an alphabet $\mathcal{A}$ depends on whether (and how much) do we know the underlying model that generates this sequence.

An Artificial Language Model

- A language uses a finite alphabet $\mathcal{A}$.
- For a sufficiently long sequence of this language, we can count the frequency of each symbol in $\mathcal{A}$ and apply Huffman coding to represent the sequence.
- Suppose that all the words are of length n and form a k-dimensional subspace of $\mathcal{A}^n$.
  - If the word space is known, by means of a basis, we can represent each word by its coordinate with respect to the basis. To compress a sequence, parse the sequence into words and represent these words by their coordinates.
  - If only the word length is known, but not the word space, we can first parse the sequence into words of n symbols and then apply the Huffman coding on words.
  - What if we don't even know the word length?

The compression ratio can be further reduced as not all combinations of words are reasonable sentences of this language. In other words, if we know better the structure of the language, lower the compression ratio can be achieved. Note that lower compression ratio can only be achieved for relatively longer sequences. If only sequences of n symbols are given for compression, you cannot do better than k/n.

People may judge whether a sequence is from a language or not, but we still don't know how to make a machine that can do the same. What can you do for compression if you are not given any information of the language? If a sequence of n symbols is given, you may not be possible to compress. If a much longer sequence is given, it is surprisingly that we can learn the structure online and can compress almost as good as you known the language model, which will be discussed in steaming source coding.

In this section, we discuss a memoryless source, i.e., symbols are generated i.i.d. from $\mathcal{A}$. But a general source is not memoryless. A language sequence of symbols from an alphabet $\mathcal{A}$ is not i.i.d. Not all sequences in $\mathcal{A}^*$ are reasonable language.

4.1 Block Source Coding

ASCII

- ASCII stands for American Standard Code for Information Interchange.
- A 7-bit character code where each one represents a unique character.
- An 8-bit extension of ASCII, Windows-12523, is probably the most-used encoding in the world.

English writing uses mainly the 2^7 symbols represented by ASCII, but other symbols, e.g., α and á, appear occasionally. In other words, this representation tries to us a short, fixed length sequence to represent the most frequently used symbols.

Block Codes

- A *block* code for representing an alphabet $\mathcal{A}$ is a pair of mappings

$$f : \mathcal{A} \to \{1, \dots, M\}, \quad \phi : \{1, \dots, M\} \to \mathcal{A},$$

where f is called the encoder and ϕ is called the decoder.

For example, let $\mathcal{A}$ be all the symbols that can be used in English writing, which has definitely larger than 128 symbols. ASCII is a code for $\mathcal{A}$ with $M = 128$, where 128 most frequently used symbols x in $\mathcal{A}$ are given a unique mapping by the encoding and decoding function, i.e., $\phi(f(x)) = x$. Other symbols x' in $\mathcal{A}$ can be mapped to a special symbol, e.g., '?', so that $\phi(f(x')) =?$.

GB2312 Character Set for Simplified Chinese

- Uses two bytes (16 bits) to represents $6,763$ Chinese characters.
- Covers over 99% of the characters of contemporary usage.

GB2312 is sufficient for most of our daily conversations, but is not enough to for historical text and many names.

Equivalent Definition of Block Codes

- A block code for representing a finite alphabet $\mathcal{A}$ is a finite subset $\mathcal{C}$ of $\mathcal{A}$.
- $M = |\mathcal{C}|$.
- The encoding and decoding functions are implied by the set $\mathcal{C}$.
- We will study the behavior of block codes for $\mathcal{A}^n$ when n tends to infinity.

4.2 Discrete Memoryless Source

Discrete Memoryless Source

- A discrete memoryless source (DMS) is a sequence $\{X_i\}_{i=1}^{\infty}$ of random variables taking values in a finite set $\mathcal{A}$, drawn i.i.d. according to $p(x)$.
- Let X be the generic random variable with distribution $p(x)$.

Error Probability

- Consider a sequence of codes $\mathcal{C}_n \subset \mathcal{A}^n$.
- The normalized *rate* of $\mathcal{C}_n$ is $\frac{\log |\mathcal{C}_n|}{n}$.
- The *error probability* of $\mathcal{C}_n$ is

$$P_e = 1 - \sum_{\mathbf{x} \in \mathcal{C}_n} p(\mathbf{x}).$$

- We are interested in codes with small rate and small probability of error.

Suppose $p(x) > 0$, $x \in \mathcal{A}$. If we require $P_e = 0$, then $\mathcal{C}_n = \mathcal{A}^n$. When $M \geq |\mathcal{A}|^n$, we can design a code with $P_e = 0$. On the other hand, for a code with $P_e = 0$, $M \geq |\mathcal{A}|^n$, i.e., the rate of the block code is $\log |\mathcal{A}|$.

Optimal Rates

- If we allow arbitrarily small error probability, the DMS can be compressed by block codes.
- Let $M^*(n, \epsilon)$ be the smallest cardinality of an n-length block code with $P_e \leq \epsilon$.
- $M^*(n, \epsilon)$ is non-decreasing with n, but $\frac{\log M^*(n,\epsilon)}{n}$ may not be monotone with n.

■ **Example 4.1** Consider that $\mathcal{A} = \{a, b\}$ and $p(a) = 1 - p(b) = 0.4$. Draw the curve of $M^*(1, \epsilon)$ and $M^*(2, \epsilon)$ for $\epsilon \in [0, 1]$. ■

Exercise 4.1 Draw the curve of $M^*(3, \epsilon)$. ■

4.3 Weak Typical Set

Lemma 4.1 Consider a DMS $\mathcal{A}$ with distribution Q. The probability that an n-length sequence $\mathbf{x}$ is generated is

$$\begin{aligned} Q^n(\mathbf{x}) &= \prod_{a\in\mathcal{A}} Q(a)^{N(a|\mathbf{x})} \\ &= \prod_{a\in\mathcal{A}} Q(a)^{nP_{\mathbf{x}}(a)} \\ &= 2^{\sum_{a\in\mathcal{A}} nP_{\mathbf{x}(a)} \log Q(a)} \\ &= 2^{-nH(P_{\mathbf{x}})-nD(P_{\mathbf{x}}||Q)}. \end{aligned}$$

Lemma 4.2 For any type P of sequences in $\mathcal{A}^n$ and distribution Q on $\mathcal{A}$,

$$(n+1)^{-|\mathcal{A}|}2^{-nD(P||Q)} \le Q^n(T_P^n) \le 2^{-nD(P||Q)}. \tag{4.1}$$

Proof. By Lemma 4.1

$$Q^n(T_P) = |T_P|2^{-nH(P)-nD(P||Q)}. \tag{4.2}$$

The proof is compeled by the bound on $|T_P|$. ■

Convergence in Probability

- Let $X^n = (X_1, \dots, X_n)$ be the i.i.d. sequence with distribution p sampled from $\mathcal{A}$.
- By the weak law of large numbers, P_{X^n} converges in probability to p when $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \Pr\{|P_{X^n}(x) - p(x)| > \delta\} = 0, \forall x \in \mathcal{A}.$$

- Hence, $D(P_{X^n}||p) + H(P_{X^n}) = -\sum_{x\in\mathcal{A}} P_{X^n}(x) \log p(x) \to H(p)$ in probability as $n \to \infty$.

The above convergence says that the type we can obtain form samples of distribution p will be very similar to p as n is large, and hence, the empirical value of $-\sum_{x\in\mathcal{A}} P_{X^n}(x) \log p(x)$ converges to $H(p)$ in probability. (Note that in general, we do not have the convergence of entropy with respect to the convergence in divergence ([7, Problem 28-31, Chapter 2]) when $\mathcal{A}$ is countably infinite.)

Weak Typical Set

- Fix $\delta > 0$.
- Let

$$\begin{aligned} W_\delta^{(n)} &= \{\mathbf{x} \in \mathcal{A}^n : |D(P_{\mathbf{x}}||p) + H(P_{\mathbf{x}}) - H(p)| \le \delta\} \\ &= \bigcup_{\text{type } P \text{ of } \mathcal{A}^n : |D(P||p)+H(P)-H(p)|\le\delta} T_P^n. \end{aligned}$$

Lemma 4.3
1. For any $\delta > 0$, $\lim_{n\to\infty} \Pr\{X^n \in W_\delta^{(n)}\} = 1$.
2. For any $\delta > 0$ and sufficiently large n, $|W_\delta^{(n)}| \le 2^{n(H(p)+\delta)}$.

Proof. Property 1 follows that $D(P_{X^n}||p) + H(P_{X^n}) \to H(p)$ in probability.

To prove Property 2,

$$\begin{aligned} 1 \ge p^n(W_\delta^{(n)}) &= \sum_{\mathbf{x}\in W_\delta^{(n)}} p^n(\mathbf{x}) \\ &= \sum_{\mathbf{x}\in W_\delta^{(n)}} 2^{-nH(P_{\mathbf{x}})-nD(P_{\mathbf{x}}||p)} \\ &\ge \sum_{\mathbf{x}\in W_\delta^{(n)}} 2^{-nH(p)-n\delta} \\ &= |W_\delta^{(n)}|2^{-nH(p)-n\delta}, \end{aligned}$$

■

4.4 Block Source Coding Theorem

Theorem 4.4 — Block Source Coding Theorem. For a discrete memoryless source with distribution p,

$$\lim_{n\to\infty} \frac{\log M^*(n,\epsilon)}{n} = H(X), \text{ for every } \epsilon \in (0,1).$$

Code Construction

- Let $\mathcal{C}_n = W_\delta^{(n)}$.
- For all sufficiently large n, (by Property 1)

$$P_e = \Pr\{X^n \notin \mathcal{W}_\delta^{(n)}\} \le \epsilon.$$

- So for any $\epsilon > 0$ and all sufficiently large n, $M^*(n,\epsilon) \le |W_\delta^{(n)}|$.
- Moreover, (by Property 2)

$$\lim_{n\to\infty} \frac{M^*(n,\epsilon)}{n} \le \lim_{n\to\infty} \frac{\log |W_\delta^{(n)}|}{n} \le H(p) + \delta.$$

Converse

- Consider a sequence of code $\mathcal{C}_n \subset \mathcal{A}^n$ with $\Pr\{X^n \in \mathcal{C}_n\} \ge 1-\epsilon$.
- As $\Pr\{X^n \notin W_\delta^{(n)}\} + \Pr\{X^n \in W_\delta^{(n)} \cap \mathcal{C}_n\} \ge P(\mathcal{C}_n) \ge 1-\epsilon$ and $\Pr\{X^n \notin W_\delta^{(n)}\} \to 0$ (Property 1), for sufficiently large n, $\Pr\{X^n \in W_\delta^{(n)} \cap \mathcal{C}_n\} \ge \frac{1-\epsilon}{2}$.
- Hence, for sufficiently large n

$$\begin{aligned}\frac{1-\epsilon}{2} &\le \Pr\{X^n \in W_\delta^{(n)} \cap \mathcal{C}_n\} \\ &\le |\mathcal{C}_n \cap W_\delta^{(n)}| 2^{-n(H(p)-\delta)} \\ &\le |\mathcal{C}_n| 2^{-n(H(p)-\delta)}.\end{aligned}$$

- So for every $\delta > 0$,

$$\lim_{n\to\infty} \frac{M^*(n,\epsilon)}{n} = \lim_{n\to\infty} \min_{A\subset\mathcal{X}^n:\Pr\{X^n\in\mathcal{C}_n\}\ge 1-\epsilon} \frac{\log |\mathcal{C}_n|}{n} \ge H(p) - \delta.$$

Universal Block Source Coding

Theorem 4.5 There exists a sequence of rate R codes such that $P_e \to 0$ for every DMS Q over $\mathcal{A}$ with $H(Q) < R$.

Proof. The coding problem is equivalent to find a set $\mathcal{A}_n \subset \mathcal{A}^n$ such that the sequences in $\mathcal{A}_n$ are decoded correctly.

Let $R_n = R - |\mathcal{A}|\frac{\log(n+1)}{n}$, and

$$\mathcal{A}_n = \bigcup_{\text{type } P:H(P)\le R_n} T_P. \tag{4.3}$$

We have

$$\begin{aligned}|\mathcal{A}_n| &= \sum_{P:\text{type},H(P)\le R_n} |T_p| &(4.4)\\ &\le (n+1)^{|\mathcal{X}|} 2^{nR_n} &(4.5)\\ &= 2^{nR}, &(4.6)\end{aligned}$$

and

$$\begin{aligned} P_e &= 1 - Q(\mathcal{A}_n) && (4.7) \\ &= Q\left(\cup_{P:\text{type},H(P)>R_n} T_p\right) && (4.8) \\ &= \sum_{P:\text{type},H(P)>R_n} Q(T_p) && (4.9) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P:H(P)>R_n} D(P||Q)} && (4.10) \end{aligned}$$

Since $R_n \uparrow R$, when n is sufficiently large, $H(Q) < R_n$ and hence $\min_{P:H(P)>R_n} D(P||Q)$ is strictly positive. ■

See more applications of types in [2, Chap. 11].