Chapter 3

General Lossless Data Compression Theorems

Po-Ning Chen

Department of Communications Engineering

National Chiao-Tung University

Hsin Chu, Taiwan 30050

Motivations

- In Volume I of the lecture notes, we already know that the entropy rate

$$\lim_{n\to\infty} \frac{1}{n} H(X^n)$$

  is the minimum data compression rate (nats per source symbol) for arbitrarily small data compression error for block coding of the stationary-ergodic source.

- We also mentioned that for a more complicated situations where the sources becomes non-stationary, the quantity $\lim_{n\to\infty}(1/n)H(X^n)$ may not exist, and can no longer be used to characterize the source compression.

- This results in the need to establish a new entropy measure which appropriately characterizes the operational limits of arbitrary stochastic systems, which was done in the previous chapter.

- Here, we have made an implicit assumption in the following derivation, which is the source alphabet $\mathcal{X}$ is *finite*.
- Actually, the following theorems also apply for sources with *countable* alphabets. We assume finite alphabets in order to avoid uninteresting cases (such as $\bar{H}_\varepsilon(\boldsymbol{X}) = \infty$) that might arise with countable alphabets.

Definition 3.1 (cf. Definition 4.2 and its associated footnote in volume I) An (n, M) block code for data compression is a set

$$\mathscr{C}_n \stackrel{\triangle}{=} \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_M\}$$

consisting of M sourcewords of block length n (and a binary-indexing codeword for each sourceword $\boldsymbol{c}_i$); each sourceword represents a group of source symbols of length n.

- In Definition 4.2 of volume I, the (n, M) block data compression code is defined by M codewords, where each codeword represents a group of sourcewords of length n. However, we can actually pick up one source symbol from each group, and equivalently define the code using these M representative sourcewords.
- Later, it will be shown that this viewpoint does facilitate the proving of the general source coding theorem.

Fixed-length codes for arbitrary sources

Definition 3.2 Fix $\varepsilon \in [0,1]$. R is an ε-achievable data compression rate for a source $\boldsymbol{X}$ if there exists a sequence of block data compression codes $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ with

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n \leq R,$$

and

$$\limsup_{n\to\infty} P_e(\mathscr{C}_n) \leq \varepsilon,$$

where $P_e(\mathscr{C}_n) \stackrel{\triangle}{=} Pr\left(X^n \notin \mathscr{C}_n\right)$ is the probability of decoding error.

The infimum of all ε-achievable data compression rate for $\boldsymbol{X}$ is denoted by $T_\varepsilon(\boldsymbol{X})$.

- Note that in conventional source coding theorem, one wants to find the minimum rate with arbitrary small error. This rate is exactly $\lim_{\varepsilon\downarrow 0} T_\varepsilon(\boldsymbol{X})$.
- As expected, for DMS, $\lim_{\varepsilon\downarrow 0} T_\varepsilon(\boldsymbol{X}) = H(X)$. Indeed in this case, $T_\varepsilon(\boldsymbol{X}) = H(X)$ for any $\varepsilon \in [0,1)$.

Lemma 3.3 Fix a positive integer n. There exists an (n, M_n) source block code $\mathscr{C}_n$ for P_{X^n} such that its error probability satisfies

$$P_e(\mathscr{C}_n) \leq Pr\left[\frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log M_n\right].$$

Proof: Observe that

$$\begin{aligned}
1 &\geq \sum_{\{x^n \in \mathcal{X}^n \,:\, (1/n)h_{X^n}(x^n) \leq (1/n)\log M_n\}} P_{X^n}(x^n) \\
&\geq \sum_{\{x^n \in \mathcal{X}^n \,:\, (1/n)h_{X^n}(x^n) \leq (1/n)\log M_n\}} \frac{1}{M_n} \\
&\geq \left|\left\{x^n \in \mathcal{X}^n \,:\, \frac{1}{n}h_{X^n}(x^n) \leq \frac{1}{n}\log M_n\right\}\right| \frac{1}{M_n}.
\end{aligned}$$

Therefore, $|\{x^n \in \mathcal{X}^n \,:\, (1/n)h_{X^n}(x^n) \leq (1/n)\log M_n\}| \leq M_n$. We can then choose a code

$$\mathscr{C}_n \supset \left\{x^n \in \mathcal{X}^n \,:\, \frac{1}{n}h_{X^n}(x^n) \leq \frac{1}{n}\log M_n\right\}$$

with $|\mathscr{C}_n| = M_n$ and

$$P_e(\mathscr{C}_n) = 1 - P_{X^n}\{\mathscr{C}_n\} \leq Pr\left[\frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log M_n\right].$$

□

Fixed-length codes for arbitrary sources

Lemma 3.4 Every (n, M_n) source block code $\mathscr{C}_n$ for P_{X^n} satisfies

$$P_e(\mathscr{C}_n) \geq Pr\left[\frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log M_n + \gamma\right] - \exp\{-n\gamma\},$$

for every $\gamma > 0$.

Proof: It suffices to prove that

$$1 - P_e(\mathscr{C}_n) = \Pr\{X^n \in \mathscr{C}_n\} < \Pr\left[\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right] + \exp\{-n\gamma\}.$$

Clearly,

$$\begin{aligned}\Pr\{X^n \in \mathscr{C}_n\} \;=\; & \Pr\left\{X^n \in \mathscr{C}_n \text{ and } \frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} \\ & + \Pr\left\{X^n \in \mathscr{C}_n \text{ and } \frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log M_n + \gamma\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq \Pr\left\{\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} \\
&\quad + \Pr\left\{X^n \in \mathscr{C}_n \text{ and } \frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log M_n + \gamma\right\} \\
&= \Pr\left\{\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} \\
&\quad + \sum_{x^n \in \mathscr{C}_n} P_{X^n}(x^n) \cdot \mathbf{1}\left\{\frac{1}{n}h_{X^n}(x^n) > \frac{1}{n}\log M_n + \gamma\right\} \\
&= \Pr\left\{\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} \\
&\quad + \sum_{x^n \in \mathscr{C}_n} P_{X^n}(x^n) \cdot \mathbf{1}\left\{P_{X^n}(x^n) < \frac{1}{M_n}\exp\{-n\gamma\}\right\} \\
&< \Pr\left\{\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} + |\mathscr{C}_n|\frac{1}{M_n}\exp\{-n\gamma\} \\
&= \Pr\left\{\frac{1}{n}h_{X^n}(X^n) \leq \frac{1}{n}\log M_n + \gamma\right\} + \exp\{-n\gamma\}.
\end{aligned}
$$

□

Fixed-length codes for arbitrary sources

We now apply Lemmas 3.3 and 3.4 to prove a *general* source coding theorems for block codes.

Theorem 3.5 (general source coding theorem) For any source $\boldsymbol{X}$,

$$T_\varepsilon(\boldsymbol{X}) = \begin{cases} \lim\limits_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}), & \text{for } \varepsilon \in [0,1); \\ 0, & \text{for } \varepsilon = 1. \end{cases}$$

Proof: The case of $\varepsilon = 1$ follows directly from its definition; hence, the proof only focus on the case of $\varepsilon \in [0,1)$.

1. *Forward part (achievability):* $T_\varepsilon(\boldsymbol{X}) \leq \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X})$

We need to prove the existence of a sequence of block codes $\{\mathcal{C}_n\}_{n\geq 1}$ such that for every $\gamma > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{C}_n| \leq \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) + \gamma \quad \text{and} \quad \limsup_{n\to\infty} P_e(\mathcal{C}_n) \leq \varepsilon.$$

Lemma 3.3 ensures the existence (for any $\gamma > 0$) of a source block code $\mathcal{C}_n = (n, M_n = \lceil \exp\{n(\lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) + \gamma)\}\rceil)$ with error probability

$$\begin{aligned} P_e(\mathcal{C}_n) &\leq Pr\left\{\frac{1}{n} h_{X^n}(X^n) > \frac{1}{n}\log M_n\right\} \\ &\leq Pr\left\{\frac{1}{n} h_{X^n}(X^n) > \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) + \gamma\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned}\limsup_{n\to\infty} P_e(\mathscr{C}_n) &\le \limsup_{n\to\infty} Pr\left\{\frac{1}{n}h_{X^n}(X^n) > \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) + \gamma\right\} \\ &= 1 - \liminf_{n\to\infty} Pr\left\{\frac{1}{n}h_{X^n}(X^n) \le \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) + \gamma\right\} \\ &\le 1 - (1-\varepsilon) = \varepsilon,\end{aligned}$$

where the last inequality follows from

$$\lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) = \sup\left\{\theta : \liminf_{n\to\infty} Pr\left[\frac{1}{n}h_{X^n}(X^n) \le \theta\right] < 1-\varepsilon\right\}. \tag{3.1.1}$$

2. *Converse part:* $T_\varepsilon(\boldsymbol{X}) \ge \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X})$
Assume without loss of generality that $\lim_{\gamma\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) > 0$. We will prove the converse by contradiction. Suppose that $T_\varepsilon(\boldsymbol{X}) < \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X})$. Then $(\exists\ \gamma > 0)\ T_\varepsilon(\boldsymbol{X}) < \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) - 4\gamma$. By definition of $T_\varepsilon(\boldsymbol{X})$, there exists a sequence of codes $\mathscr{C}_n$ such that

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathscr{C}_n| \le \left(\lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) - 4\gamma\right) + \gamma \tag{3.1.2}$$

and

$$\limsup_{n\to\infty} P_e(\mathscr{C}_n) \le \varepsilon. \tag{3.1.3}$$

(3.1.2) implies that

$$\frac{1}{n}\log|\mathscr{C}_n| \le \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) - 2\gamma$$

for all sufficiently large n. Hence, for those n satisfying the above inequality and also by Lemma 3.4,

$$\begin{aligned} P_e(\mathscr{C}_n) &\ge Pr\left[\frac{1}{n}h_{X^n}(X^n) > \frac{1}{n}\log|\mathscr{C}_n| + \gamma\right] - e^{-n\gamma} \\ &\ge Pr\left[\frac{1}{n}h_{X^n}(X^n) > \left(\lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) - 2\gamma\right) + \gamma\right] - e^{-n\gamma}. \end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} P_e(\mathscr{C}_n) \ge 1 - \liminf_{n\to\infty} Pr\left[\frac{1}{n}h_{X^n}(X^n) \le \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) - \gamma\right] > \varepsilon,$$

where the last inequality follows from (3.1.1). Thus, a contradiction to (3.1.3) is obtained. □

Fixed-length codes for arbitrary sources

A few remarks are made based on the previous theorem.

- Note that as $\varepsilon = 0$,

$$T_0(\boldsymbol{X}) = \lim_{\delta \uparrow (1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}).$$

  Hence, the minimum (asymptotic) lossless fixed-length source coding rate of any finite-alphabet source is $\bar{H}(\boldsymbol{X})$.

- Consider the special case where

$$-\frac{1}{n} \log P_{X^n}(X^n) \text{ converges in probability to a constant } H \text{ (entropy rate)},$$

  which holds for all *information stable* sources. In this case, both the inf- and sup-spectrums of $\boldsymbol{X}$ degenerate to a unit step function:

$$u(\theta) = \begin{cases} 1, & \text{if } \theta > H; \\ 0, & \text{if } \theta < H. \end{cases}$$

  Thus, $\bar{H}_\varepsilon(\boldsymbol{X}) = H$ for all $\varepsilon \in [0, 1)$. Hence, general source coding theorem reduces to the conventional source coding theorem.

Fixed-length codes for arbitrary sources

– A source

$$\boldsymbol{X} = \left\{ X^n = \left(X_1^{(n)}, \ldots, X_n^{(n)} \right) \right\}_{n=1}^{\infty}$$

is said to be *information stable* if

$$H(X^n) = E\left[-\log P_{X^n}(x^n) \right] > 0 \text{ for all } n,$$

and

$$\lim_{n \to \infty} Pr\left(\left| \frac{-\log P_{X^n}(x^n)}{H(X^n)} - 1 \right| > \varepsilon \right) = 0,$$

for every $\varepsilon > 0$.

– By the definition, any stationary-ergodic source with finite n-fold entropy is information stable; hence, it can be viewed a generalized source model for stationary-ergodic sources.

Fixed-length codes for arbitrary sources

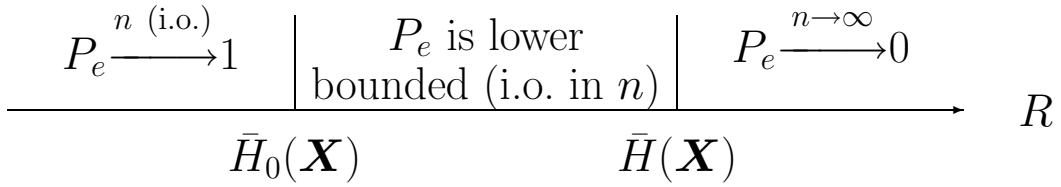
- If
$$-\frac{1}{n}\log P_{X^n}(X^n) \text{ converges in probability to a random variable } Z$$
whose cdf is $F_Z(\cdot)$, then the minimum achievable data compression rate subject to decoding error being no greater than ε is
$$T_\varepsilon(\boldsymbol{X}) = \lim_{\delta\uparrow(1-\varepsilon)} \bar{H}_\delta(\boldsymbol{X}) = \sup\{R \ : \ F_Z(R) < 1-\varepsilon\}.$$

Example 3.6 Consider a binary source $\boldsymbol{X}$ with each X^n is Bernoulli(Θ) distributed, where Θ is a random variable defined over $(0,1)$. By ergodic decomposition theorem (which states that any stationary source can be viewed as a mixture of stationary-ergodic sources) that
$$-\frac{1}{n}\log P_{X^n}(X^n) \text{ converges in probability to} h_b(\Theta),$$
where $h_b(x) \stackrel{\triangle}{=} -x\log_2(x) - (1-x)\log_2(1-x)$. Consequently,
$$T_\varepsilon(\boldsymbol{X}) = \sup\{R \ : \ Pr\{h_b(\Theta) \le R\} < 1-\varepsilon\}.$$

Fixed-length codes for arbitrary sources

- From the above example, or from Theorem 3.5, it shows that the *strong converse theorem* (which states that codes with rate below entropy rate will ultimately have decompression error approaching one, cf. Theorem 4.6 of Volume I of the lecture notes) does not hold in general. However, one can always claim the *weak converse statement* for arbitrary sources.

  Theorem 3.7 (weak converse theorem) For any block code sequence of ultimate rate $R < \bar{H}(\boldsymbol{X})$, the probability of block decoding failure P_e cannot be made arbitrarily small. In other words, there exists $\varepsilon > 0$ such that P_e is lower bounded by ε infinitely often in block length n.



Behavior of the probability of block decoding error as block length n goes to infinity for an *arbitrary* source $\boldsymbol{X}$.

Theorem 3.8 (generalized asymptotic equipartition property for arbitrary sources) Fix $\varepsilon \in [0,1)$. Given an arbitrary source $\boldsymbol{X}$, define

$$\mathcal{T}_n[R] \stackrel{\triangle}{=} \left\{ x^n \in \mathcal{X}^n : -\frac{1}{n} \log P_{X^n}(x^n) \leq R \right\}.$$

Then for any $\delta > 0$, the following statements hold.

1.
$$\liminf_{n\to\infty} Pr\left\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) - \delta]\right\} \leq \varepsilon \tag{3.1.4}$$

2.
$$\liminf_{n\to\infty} Pr\left\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) + \delta]\right\} > \varepsilon \tag{3.1.5}$$

3. The number of elements in

$$\mathcal{F}_n(\delta;\varepsilon) \stackrel{\triangle}{=} \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) + \delta] \setminus \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) - \delta],$$

denoted by $|\mathcal{F}_n(\delta;\varepsilon)|$, satisfies

$$|\mathcal{F}_n(\delta;\varepsilon)| \leq \exp\left\{n(\bar{H}_\varepsilon(\boldsymbol{X}) + \delta)\right\}, \tag{3.1.6}$$

where the operation $\mathcal{A}\setminus\mathcal{B}$ between two sets $\mathcal{A}$ and $\mathcal{B}$ is defined by $\mathcal{A}\setminus\mathcal{B} \stackrel{\triangle}{=} \mathcal{A}\cap\mathcal{B}^c$ with $\mathcal{B}^c$ denoting the complement set of $\mathcal{B}$.

Generalized AEP theorem

4. There exists $\rho = \rho(\delta) > 0$ and a subsequence $\{n_j\}_{j=1}^{\infty}$ such that

$$|\mathcal{F}_n(\delta;\varepsilon)| > \rho \cdot \exp\left\{n_j(\bar{H}_\varepsilon(\boldsymbol{X}) - \delta)\right\}. \tag{3.1.7}$$

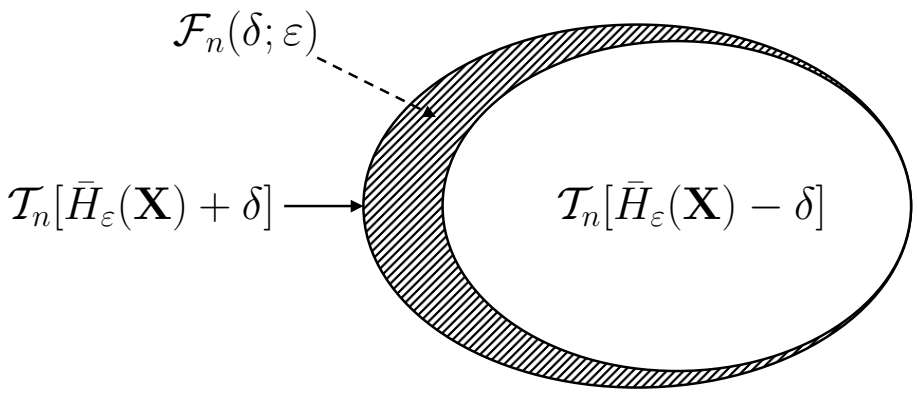


Illustration of generalized AEP Theorem. $\mathcal{F}_n(\delta;\varepsilon) \stackrel{\triangle}{=} \mathcal{T}_n[\bar{H}_\varepsilon(\mathbf{X}) + \delta] \setminus \mathcal{T}_n[\bar{H}_\varepsilon(\mathbf{X}) - \delta]$ is the dashed region.

Generalized AEP theorem

- The set

$$\begin{aligned}\mathcal{F}_n(\delta;\varepsilon) &\stackrel{\triangle}{=} \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\delta] \setminus \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\delta] \\ &= \left\{x^n \in \mathcal{X}^n : \left|-\frac{1}{n}\log P_{X^n}(x^n) - \bar{H}_\varepsilon(\boldsymbol{X})\right| \le \delta\right\}\end{aligned}$$

  is nothing but the weakly δ-typical set.

- $q_n \stackrel{\triangle}{=} Pr\{\mathcal{F}_n(\delta;\varepsilon)\} > 0$ infinitely often in n.

- $$|\mathcal{F}_n(\delta;\varepsilon)| \approx e^{n\bar{H}_\varepsilon(\boldsymbol{X})},$$

  and the probability of each sequence in $\mathcal{F}_n(\delta;\varepsilon)$ can be estimated by $q_n \cdot \exp\left\{-n\bar{H}_\varepsilon(\boldsymbol{X})\right\}$.

- In particular, if $\boldsymbol{X}$ is a stationary-ergodic source, then $\bar{H}_\varepsilon(\boldsymbol{X})$ is independent of $\varepsilon \in [0,1)$ and, $\bar{H}_\varepsilon(\boldsymbol{X}) = \underline{H}_\varepsilon(\boldsymbol{X}) = H$ for all $\varepsilon \in [0,1)$, where H is the source entropy rate

$$H = \lim_{n\to\infty} \frac{1}{n} E\left[-\log P_{X^n}(X^n)\right].$$

  In this case, the generalized AEP reduces to the conventional AEP.

VL codes with smallest weighted codelength

Criterion for optimality of codes

- Conventional source coding theorem: minimize the average codeword length.
  - Meaning: system cost varies *linearly* with codeword length.
- General source coding theorem: minimize the weighted codeword length.
  - Example. Exponential cost function

$$L(t) \stackrel{\triangle}{=} \frac{1}{t} \log \left(\sum_{x \in \mathcal{X}} P_X(x) e^{t \cdot \ell(\boldsymbol{c}_x)} \right),$$

    where t is a chosen positive constant, P_X is the distribution function of source X, $\boldsymbol{c}_x$ is the binary code for source symbol x, and $\ell(\cdot)$ is the length of the binary codeword.
  - Optimality: *a code is said to be optimal if its cost $L(t)$ is the smallest among all possible codes.*
  - Meaning:
    * When $t \to 0$, $L(0) = \sum_{x \in \mathcal{X}} P_X(x) \ell(\boldsymbol{c}_x)$ which is the average codeword length.
    * In the case of $t \to \infty$, $L(\infty) = \max_{x \in \mathcal{X}} \ell(\boldsymbol{c}_x)$, which is the maximum codeword length for all binary codewords.

VL codes with smallest weighted codelength

– Weight function:

$$\sum_{x\in\mathcal{X}} P_X(x)e^{t\ell(\boldsymbol{c}_x)} = \text{the weight for codeword } \boldsymbol{c}_x \text{ is } e^{t\ell(\boldsymbol{c}_x)}.$$

For the minimization operation, it is obvious that events with smaller weight is more preferable since it contributes less in the sum of $L(t)$.

– Therefore, with the minimization of $L(t)$, it is less likely to have codewords with long code lengths. In practice, system with long codewords usually introduce complexity in encoding and decoding, and hence is somewhat non-feasible. Consequently, the new criterion, to some extent, is more suitable to physical considerations.

Source coding theorem for Rényi's entropy

Theorem 3.9 (source coding theorem for Rényi's entropy) The minimum cost $L(t)$ attainable for uniquely decodable codes is the Rényi's entropy of order $1/(1+t)$, i.e.,

$$H\left(X; \frac{1}{(1+t)}\right) = \frac{1+t}{t} \log \left(\sum_{x \in \mathcal{X}} P_X^{1/(1+t)}(x) \right).$$

Example 3.10 Given a source X. If we want to design an optimal lossless code with $\max_{x \in \mathcal{X}} \ell(\boldsymbol{c}_x)$ being smallest, the cost of the optimal code is $H(X; 0) = \log |\mathcal{X}|$.

Entropy of English

- The compression of English text is not only a practical application but also an interesting research topic.
- One of the main problem in the compression of English text (in principle) is that its statistical model is unclear.
- Therefore, its entropy rate cannot be immediately computable.
- To estimate the data compression bound of English text, various stochastic approximations to English have been proposed. One can then design a code, according to the estimated stochastic model, to compress English text. It is obvious that the better the stochastic approximation, the better the approximation.

Assumption 3.11 For data compression, we assume that the English text contains only 26 letters and the space symbol. In other words, the upper case letter is treated the same as its lower case counterpart, and special symbols, such as punctuation, will be ignored.

Markov estimate of entropy rate of English text

- According to the source coding theorem, the first thing that a data compression code designer shall do is to estimate the (δ-*sup*) *entropy rate* of the English text.
- One can start from modeling the English text as a Markov source, and compute the entropy rate according to the estimated Markov statistics.

zero-order Markov approximation. It has been shown that the frequency of letters in English is far from uniform; for example, the most common letter, E, has $P_{\text{empirical}}(E) \approx 0.13$ but the least common letters, Q and Z, have $P_{\text{empirical}}(Q) \approx P_{\text{empirical}}(Z) \approx 0.001$. Therefore, zero-order Markov approximation apparently does not fit our need in estimating the entropy rate of the English text.

1st-order Markov approximate. The frequency of pairs of letters is also far from uniform; the most common pair TH has frequency about 0.037. It is fun to know that Q is always followed by U.

A higher order approximation is possible. However, the database may be too large to be handled. For example, a 2nd-order approximation requires $27^3 = 19683$ entries, and one may need millions of samples to make an accurate estimate of its probability.

Markov estimate of entropy rate of English text

Here are some examples of Markov approximations to English from Shannon's original paper. Note that the sequence of English letters is generated according to the approximated statistics.

1. Zero-order Markov approximation: The symbols are drawn independently with equiprobable distribution.

   Example: XFOML RXKHRJEFJUJ ZLPWCFWKCYJ EFJEYVKCQSGX YD QPAAMKBZAACIBZLHJQD

2. 1st-order Markov approximation: The symbols are drawn independently. Frequency of letters matches the 1st-order Markov approximation of English text.

   Example: OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

3. 2nd-order Markov approximation: Frequency of pairs of letters matches English text.

   Example: ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN AN DY TOBE SEACE CTISBE

Markov estimate of entropy rate of English text

4. 3rd-order Markov approximation: Frequency of triples of letters matches English text.

   Example: IN NO IST LAT WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

5. 4th-order Markov approximation: Frequency of quadruples of letters matches English text.

   Example: THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE, ABOVERY UPONDULTS WELL THE CODE RST IN THESTICAL IT DO HOCK BOTHE MERG.

Markov estimate of entropy rate of English text

Another way to simulate the *randomness* of English text is to use word-approximation.

1. 1st-order word approximation.
   Example: REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.
2. 2nd-order word approximation.
   Example: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

From the above results, it is obvious that the approximations get closer and closer to resembling English for higher order approximation.

Using the above model, we can then compute the empirical entropy rate of English text.

order of the letter approximation model	entropy rate
zero order	$\log_2 27 = 4.76$ bits per letter
1st order	4.03 bits per letter
4th order	2.8 bits per letter

Markov estimate of entropy rate of English text

One final remark on Markov estimate of English statistics is that the results are not only useful in compression but also helpful in decryption (such as by letter guessing).

Gambling estimate of entropy rate of English text

A good gambler is also a good data compressor!

A) Sequential gambling

- Given an observed sequence of letters,

$$x_1, x_2, \ldots, x_k,$$

a gambler needs to bet on the next letter X_{k+1} (which is now a random variable) with all the money in hand.

  - It is not necessary for him to put all the money on the same outcome (there are 27 of them, i.e., 26 letters plus "space"). For example, he is allowed to place part of his money on one possible outcome and the rest of the money on another. The only constraint is that he should bet all his money.

Gambling estimate of entropy rate of English text

- Let $b(x_{k+1}|x_1, \ldots, x_k)$ be the ratio of his money, which be bet on the letter x_{k+1}, and assume that at first, the amount of money that the gambler has is 1. Then

$$\sum_{x_{k+1} \in \mathcal{X}} b(x_{k+1}|x_1, \ldots, x_k) = 1,$$

and

$$(\forall\ x_{k+1} \in \{a, b, \ldots, z, \text{SPACE}\})\ b(x_{k+1}|x_1, \ldots, x_k) \geq 0.$$

- When the next letter appears, the gambler will be paid 27 times the bet on the letter.

- Let S_n be the wealth of the gambler after n bets. Then

$$\begin{aligned}
S_1 &= 27 \cdot b_1(x_1) \\
S_2 &= 27 \cdot [b_2(x_2|x_1) \times S_1] \\
S_3 &= 27 \cdot [b_3(x_3|x_1, x_2) \times S_2] \\
&\vdots \\
S_n &= 27^n \cdot \prod_{k=1}^{n} b_k(x_k|x_1, \ldots, x_{k-1}) \\
&= 27^n \cdot b(x_1, \ldots, x_n),
\end{aligned}$$

where

$$b(x^n) = b(x_1, \ldots, x_n) = \prod_{k=1}^{n} b_k(x_k | x_1, \ldots, x_{k-1}).$$

We now wish to show that *high value of* S_n lead to *high data compression*. Specifically, if a gambler with some gambling policy yields wealth S_n, the data can be saved up to $\log_2 S_n$ bits.

Lemma 3.12 If a proportional gambling policy results in wealth $E[\log_2 S_n]$, there exists a data compression code for English-text source X^n which yields average codeword length being smaller than

$$\log_2(27) - \frac{1}{n} E\left[\log_2 S_n\right] + \frac{2}{n} \text{ bits},$$

where X^n represents the random variables of the n bet outcomes.

Proof:

1. Ordering : Index the English letter as

Gambling estimate of entropy rate of English text

$$\begin{aligned}
\text{index}(a) &= 0 \\
\text{index}(b) &= 1 \\
\text{index}(c) &= 2 \\
&\vdots \\
\text{index}(z) &= 25 \\
\text{index(SPACE)} &= 26.
\end{aligned}$$

For any two sequences x^n and $\hat{x}^n$ in $\mathcal{X} \stackrel{\triangle}{=} \{a, b, \ldots, z, \text{SPACE}\}$, we say $x^n \geq \hat{x}^n$ if

$$\sum_{i=1}^{n} \text{index}(x_i) \times 27^{i-1} \geq \sum_{i=1}^{n} \text{index}(\hat{x}_i) \times 27^{i-1}.$$

2. Shannon-Fano-Elias coder : Apply Shannon-Fano-Elias coder to the gambling policy $b(x^n)$ according to the ordering defined in step 1. The the codeword length for x^n is

$$(\lceil -\log_2(b(x^n)) \rceil + 1) \text{ bits.}$$

3. Data compression : Now observe that

$$\begin{aligned}
E\left[\log_2 S_n\right] &= E\left[\log_2(27^n b(x^n))\right] \\
&= n \log_2(27) + E[\log_2 b(X^n)].
\end{aligned}$$

Gambling estimate of entropy rate of English text

Hence, the average codeword length $\bar{\ell}$ is upper bounded by

$$\begin{aligned}
\bar{\ell} &\leq \frac{1}{n}\left(-\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\log_2 b(x^n) + 2\right) \\
&= -\frac{1}{n}E\left[\log_2 b(X^n)\right] + \frac{2}{n} \\
&= \log_2(27) - \frac{1}{n}E[\log_2 S_n] + \frac{2}{n}.
\end{aligned}$$

□

Gambling estimate of entropy rate of English text

- According to the concept behind the source coding theorem, the entropy rate of the English text should be upper bounded by the average codeword length of any (variable-length) code, which in turns should be bounded above by

$$\log_2(27) - \frac{1}{n}E[\log_2 S_n] + \frac{2}{n}.$$

- Equipped with the proportional gambling model, one can then find the bound of the entropy rate of English text by a properly designed gambling policy.

- An experiment using the book, *Jefferson the Virginian* by Dumas Malone, as the database resulted in an estimate of 1.34 bits per letter for the entropy rate of English.

Lempel-Ziv code revisited

In Section 4.3.4 of Volume I of the lecture notes, we have introduced the famous Lempel-Ziv coder, and states that the coder is universally *good* for stationary sources. In this section, we will establish the concept behind it.

For simplicity, we assume that the source alphabet is binary, i.e., $\mathcal{X} = \{0, 1\}$. The optimality of the Lempel-Ziv code can actually be extended to any stationary source with finite alphabet.

Lemma 3.13 The number $c(n)$ of distinct strings in the Lempel-Ziv parsing of a binary sequence satisfies

$$\sqrt{2n} - 1 \leq c(n) \leq \frac{2n}{\log_2 n},$$

where the upper bound holds for $n \geq 2^{13}$, and the lower bound is valid for every n. where $\varepsilon_n \to 0$ as $n \to \infty$.

Proof: The upper bound can be proved as follows.

For fixed n, the number of distinct strings is maximized when all the phrases are as short as possible. Hence, in the extreme case,

$$n = \overbrace{1 + 2 + 2 + 3 + 3 + 3 + 3 + \cdots}^{c(n) \text{ of them}},$$

which implies that

$$k2^{k+2}+2=\sum_{j=1}^{k+1} j2^j \geq n \geq \sum_{j=1}^{k} j2^j = (k-1)2^{k+1}+2, \tag{3.4.8}$$

where k is the integer satisfying

$$2^{k+1}-1 > c(n) \geq 2^k - 1. \tag{3.4.9}$$

Now from (3.4.8), we obtain for $k \geq 7$ (which will be justified later by $n \geq 2^{13}$),

$$n \leq k2^{k+2}+2 < 2^{2(k-1)} \quad \text{and} \quad n \geq (k-1)2^{k+1}+2 \geq (k-1)2^{k+1}.$$

The proof of the upper bond is completed by saying that the maximum $c(n)$ for fixed n should satisfy

$$c(n) < 2^{k+1}-1 \leq 2^{k+1} \leq \frac{n}{k-1} \leq \frac{2n}{\log_2(n)}.$$

Again for the lower bond, we note that the number of distinct strings is minimized when all the phrases are as long as possible. Hence, in the extreme case,

$$n = \overbrace{1+2+3+4+\cdots}^{c(n) \text{ of them}} \leq \sum_{j=1}^{c(n)} j = \frac{c(n)[c(n)+1]}{2} \leq \frac{[c(n)+1]^2}{2},$$

which implies that

$$c(n) \geq \sqrt{2n} - 1.$$

Note that when $n \geq 2^{13}$, $c(n) \geq 2^7 - 1$, which implies that the $k \geq 7$ assumption in (3.4.9) is valid for $n \geq 2^{13}$. □

- The condition $n \geq 2^{13}$ is equivalent to compressing a binary file of size larger than $1K$ bytes, which, in practice, is a frequently encountered situation. Since what concerns us is the *asymptotic* optimality of the coder as n goes to infinity, $n \geq 2^{13}$ certainly becomes insignificant in such considerations.

Lemma 3.14 (entropy upper bound by a function of its mean) A non-negative integer-valued source X with mean μ and entropy $H(X)$ satisfies

$$H(X) \leq (\mu + 1)\log_2(\mu + 1) - \mu \log_2 \mu.$$

Proof: The lemma follows directly from the result that the geometric distribution maximizes the entropy of non-negative integer-valued source with given mean, which is proved as follows.

For geometric distribution with mean μ,

$$P_Z(z) = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^z, \text{ for } z = 0, 1, 2, \ldots,$$

its entropy is

$$\begin{aligned} H(Z) &= \sum_{z=0}^{\infty} -P_Z(z)\log_2 P_Z(z) \\ &= \sum_{z=0}^{\infty} P_Z(z)\left[\log_2(1+\mu) + z\cdot\log_2\frac{1+\mu}{\mu}\right] \\ &= \log_2(1+\mu) + \mu\log_2\frac{1+\mu}{\mu} \\ &= \sum_{z=0}^{\infty} P_X(z)\left[\log_2(1+\mu) + z\cdot\log_2\frac{1+\mu}{\mu}\right], \end{aligned}$$

where the last equality holds for any non-negative integer-valued source X with mean μ. So,

$$\begin{aligned} H(X) - H(Z) &= \sum_{x=0}^{\infty} P_X(x)[-\log_2 P_X(x) + \log_2 P_Z(x)] \\ &= \sum_{x=0}^{\infty} P_X(x)\log_2\frac{P_Z(x)}{P_X(x)} \\ &= -D(P_X\|P_Z) \le 0, \end{aligned}$$

with equality holds if, and only if, $X \equiv Z$. □

Lempel-Ziv code revisited

- Before the introduction of the main theorems, we address some notations used in their proofs.

Give the source

$$x_{-(k-1)}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n,$$

and suppose $x_1, \ldots, x_n$ is Lempel-Ziv-parsed into c distinct strings, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_c$. Let ν_i be the location of the first bit of $\boldsymbol{y}_i$, i.e.,

$$\boldsymbol{y}_i \stackrel{\triangle}{=} x_{\nu_i}, \ldots, x_{\nu_{i+1}-1}.$$

1. Define

   $$\boldsymbol{s}_i = x_{\nu_i-k}, \ldots, x_{\nu_i-1}$$

   as the k bits preceding $\boldsymbol{y}_i$.

2. Define $c_{\ell,\boldsymbol{s}}$ be the number of strings in $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_c$ with length ℓ and proceeding state $\boldsymbol{s}_i = \boldsymbol{s}$.

3. Define Q_k be the k-th order Markov approximation of the stationary source $\boldsymbol{X}$, i.e.,

   $$Q_k(x_1, \ldots, x_n | x_0, \ldots, x_{-(k-1)}) \stackrel{\triangle}{=} P_{X_1^n|X_{-(k-1)}^0}(x_n, \ldots, x_1 | x_0, \ldots, x_{-(k-1)}),$$

   where $P_{X_1^n|X_{-(k-1)}^0}$ is the true (stationary) distribution of the source.

Lempel-Ziv code revisited

It is easy to verify that

$$\sum_{\ell=1}^{n} \sum_{\boldsymbol{s} \in \mathcal{X}^k} c_{\ell,\boldsymbol{s}} = c, \quad \text{and} \quad \sum_{\ell=1}^{n} \sum_{\boldsymbol{s} \in \mathcal{X}^k} \ell \times c_{\ell,\boldsymbol{s}} = n. \tag{3.4.10}$$

For ease of understanding, these notations are graphically illustrated in the next slide.

Lempel-Ziv code revisited

$$\overbrace{x_{-(k-1)},\ldots,x_{-1}}^{\boldsymbol{s}_1}\underbrace{(x_1,\ldots,\overbrace{x_{\nu_2-(k-1)},\ldots,x_{\nu_2-1}}^{\boldsymbol{s}_2})}_{\boldsymbol{y}_1}\underbrace{(x_{\nu_2},\ldots,\overbrace{x_{\nu_3-(k-1)},\ldots,x_{\nu_3-1}}^{\boldsymbol{s}_3})}_{\boldsymbol{y}_2}$$

$$\ldots\underbrace{(x_{\nu_{c-1}},\ldots,\overbrace{x_{\nu_c-(k-1)},\ldots,x_{\nu_c-1}}^{\boldsymbol{s}_c})}_{\boldsymbol{y}_{c-1}}\underbrace{(x_{\nu_c},\ldots,x_n)}_{\boldsymbol{y}_c}$$

Figure: Notations used in Lempel-Ziv coder.

Lempel-Ziv code revisited

Lemma 3.15 For any Lempel-Ziv parsing of the source $x_1 \dots x_n$, we have

$$\log_2 Q_k(x_1, \dots, x_n | \boldsymbol{s}) \leq \sum_{\ell=1}^{n} -c_{\ell,\boldsymbol{s}} \log_2 c_{\ell,\boldsymbol{s}}. \tag{3.4.11}$$

Proof: By the k-th order Markov property of Q_k,

$$\begin{aligned}
& \log_2 Q_k(x_1, \ldots, x_n|\boldsymbol{s}) \\
= \; & \log_2 Q(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_c|\boldsymbol{x}) \\
= \; & \log_2 \left(\prod_{i=1}^{c} P_{X_1^{\nu_{i+1}-\nu_i}|X^0_{-(k-1)}}(\boldsymbol{y}_i|\boldsymbol{s}_i) \right) \\
= \; & \sum_{i=1}^{c} \log_2 P_{X_1^{\nu_{i+1}-\nu_i}|X^0_{-(k-1)}}(\boldsymbol{y}_i|\boldsymbol{s}_i) \\
= \; & \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} \left(\sum_{\{i \,:\, \nu_{i+1}-\nu_i=\ell \text{ and } \boldsymbol{s}_i=\boldsymbol{s}\}} \log_2 P_{X_1^{\ell}|X^0_{-(k-1)}}(\boldsymbol{y}_i|\boldsymbol{s}_i) \right) \\
= \; & \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} c_{\ell,\boldsymbol{s}} \left(\frac{1}{c_{\ell,\boldsymbol{s}}} \sum_{\{i \,:\, \nu_{i+1}-\nu_i=\ell \text{ and } \boldsymbol{s}_i=\boldsymbol{s}\}} \log_2 P_{X_1^{\ell}|X^0_{-(k-1)}}(\boldsymbol{y}_i|\boldsymbol{s}_i) \right) \\
\leq \; & \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} c_{\ell,\boldsymbol{s}} \log_2 \sum_{\{i \,:\, \nu_{i+1}-\nu_i=\ell \text{ and } \boldsymbol{s}_i=\boldsymbol{s}\}} \left(\frac{1}{c_{\ell,\boldsymbol{s}}} P_{X_1^{\ell}|X^0_{-(k-1)}}(\boldsymbol{y}_i|\boldsymbol{s}_i) \right) \quad (3.4.12) \\
\leq \; & \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} c_{\ell,\boldsymbol{s}} \log_2 \left(\frac{1}{c_{\ell,\boldsymbol{s}}} \right) \quad (3.4.13)
\end{aligned}$$

where (3.4.12) follows from Jensen's inequality and the concavity of $\log_2(\cdot)$, and (3.4.13) holds since probability sum is no greater than one. □

Theorem 3.16 Fix a stationary source $\boldsymbol{X}$. Given any observations x_1, x_2, x_3, ...,

$$\limsup_{n\to\infty} \frac{c \log_2 c}{n} \le H(X_{k+1}|X_k, \ldots, X_1),$$

for any integer k, where $c = c(n)$ is the number of distinct Lempel-Ziv parsed strings of $x_1, x_2, \ldots, x_n$.

Proof: Lemma 3.15 gives that

$$\begin{aligned}
\log_2 Q_k(x_1, \ldots, x_n|\boldsymbol{s}) &\le \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} -c_{\ell,\boldsymbol{s}} \log_2 c_{\ell,\boldsymbol{s}} \\
&= \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} -c_{\ell,\boldsymbol{s}} \log_2 \frac{c_{\ell,\boldsymbol{s}}}{c} + \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} -c_{\ell,\boldsymbol{s}} \log_2 c \\
&= -c \sum_{\ell=1}^{n} \sum_{\boldsymbol{s}\in\mathcal{X}^k} \frac{c_{\ell,\boldsymbol{s}}}{c} \log_2 \frac{c_{\ell,\boldsymbol{s}}}{c} - c \log_2 c. \qquad (3.4.14)
\end{aligned}$$

Denote by L and $\boldsymbol{S}$ the random variables with distribution

$$P_{L,\boldsymbol{S}}(\ell, \boldsymbol{s}) = \frac{c_{\ell,\boldsymbol{s}}}{c},$$

for which the sum-to-one property of the distribution is justified by (3.4.10). Also, from (3.4.10), we have

$$E[L] = \sum_{\ell=1}^{n} \sum_{\boldsymbol{s} \in \mathcal{X}^k} \ell \times \frac{c_{\ell,\boldsymbol{s}}}{c} = \frac{n}{c}.$$

Therefore, by independent bound for entropy (cf. Theorem 4.7 in Volume I of the lecture notes) and Lemma 3.14, we get

$$\begin{aligned} H(L, \boldsymbol{S}) &\leq H(L) + H(\boldsymbol{S}) \\ &\leq \{(E[L]+1)\log_2(E[L]+1) - E[L]\log_2 E[L]\} + \log_2 |\mathcal{X}|^k \\ &= \left[\left(\frac{n}{c}+1\right)\log_2\left(\frac{n}{c}+1\right) - \frac{n}{c}\log_2\frac{n}{c}\right] + k \\ &= \left[\log_2\left(\frac{n}{c}+1\right) + \frac{n}{c}\log_2\left(\frac{n/c+1}{n/c}\right)\right] + k \\ &= \log_2\left(\frac{n}{c}+1\right) + \frac{n}{c}\log_2\left(1+\frac{c}{n}\right) + k, \end{aligned}$$

which, together with Lemma 3.13, implies that for $n \geq 2^{13}$,

$$\begin{aligned} \frac{c}{n}H(L, \boldsymbol{S}) &\leq \frac{c}{n}\log_2\left(\frac{n}{c}+1\right) + \log_2\left(1+\frac{c}{n}\right) + \frac{c}{n}k \\ &\leq \frac{2}{\log_2 n}\log_2\left(\frac{n}{\sqrt{2n}-1}+1\right) + \log 2\left(1+\frac{2}{\log_2 n}\right) + \frac{2}{\log_2 n}k. \end{aligned}$$

Lempel-Ziv code revisited

Finally, we can re-write (3.4.14) as

$$\frac{c \log_2 c}{n} \leq -\frac{1}{n} \log_2 Q_k(x_1, \ldots, x_n | \boldsymbol{x}) + \frac{c}{n} H(L, \boldsymbol{S}).$$

As a consequence, by taking the expectation value with respect to $X_{-(k-1)}^n$ on both sides of the above inequality, we obtain

$$\begin{aligned} \limsup_{n\to\infty} \frac{c \log_2 c}{n} &\leq \limsup_{n\to\infty} \frac{1}{n} E\left[-\log_2 Q_k(X_1, \ldots, X_n | X_{-(k-1)}, \ldots, X_0)\right] \\ &= H(X_{k+1} | X_k, \ldots, X_1). \end{aligned}$$

□

Theorem 3.17 (Main result) Let $\ell(x_1, \ldots, x_n)$ be the Lempel-Ziv codeword length of an observatory sequence $x_1, \ldots, x_n$, which is drawn from a stationary source $\boldsymbol{X}$. Then

$$\limsup_{n\to\infty} \frac{1}{n} \ell(x_1, \ldots, x_n) \leq \lim_{n\to\infty} \frac{1}{n} H(X^n).$$

Proof: Let $c(n)$ be the number of the parsed distinct strings, then

$$\begin{aligned}\frac{1}{n}\ell(x_1,\ldots,x_n) &= \frac{1}{n}c(n)\left(\lceil\log_2 c(n)\rceil+1\right)\\ &\le \frac{1}{n}c(n)\left(\log_2 c(n)+2\right)\\ &= \frac{c(n)\log_2 c(n)}{n}+2\frac{c(n)}{n}.\end{aligned}$$

From Lemma 3.13, we have

$$\limsup_{n\to\infty}\frac{c(n)}{n}\le\limsup_{n\to\infty}\frac{2}{\log_2 n}=0.$$

From Theorem 3.16, we have for any integer k,

$$\limsup_{n\to\infty}\frac{c(n)\log_2 c(n)}{n}\le H(X_{k+1}|X_k,\ldots,X_1).$$

Hence,

$$\limsup_{n\to\infty}\frac{1}{n}\ell(x_1,\ldots,x_n)\le H(X_{k+1}|X_k,\ldots,X_1)$$

for any integer k. The theorem is completed by applying Theorem 6.5 in Volume I of the lecture notes, which states that *for a stationary source* $\boldsymbol{X}$*, its entropy rate always exists and is equal to*

$$\lim_{n\to\infty}\frac{1}{n}H(X^n)=\lim_{k\to\infty}H(X_{k+1}|X_k,\ldots,X_1).$$

Lempel-Ziv code revisited

□

- We conclude the discussion in the section into the next corollary.

Corollary 3.18 The Lempel-Ziv coders asymptotically achieves the entropy rate of any (unknown) stationary source.

Lempel-Ziv code revisited

- The Lempel-Ziv code is often used in practice to compress data which cannot be characterized in a simple statistical model, such as English text or computer source code.
- It is simple to implement, and has an asymptotic rate approaching the entropy rate (if it exists) of the source, which is known to be the lower bound of the lossless data compression code rate.
- This code can be used without knowledge of the source distribution provided the source is stationary.
- Some well-known examples of its implementation are the *compress* program in UNIX and the *arc* program in DOS, which typically compresses ASCII text files by about a factor of 2.