Coding and Information Theory
Chapter 3
Entropy

Xuejun Liang

2019 Fall

Chapter 3: Entropy

3.1 Information and Entropy

3.2 Properties of the Entropy Function

3.3 Entropy and Average Word-length

3.4 Shannon-Fane Coding

3.5 Entropy of Extensions and Products

3.6 Shannon's First Theorem

3.7 An Example of Shannon's First Theorem

The aim of this chapter

- Introduce the entropy function
  - which measures the amount of information emitted by a source
- Examine the basic properties of this function
- Show how it is related to the average word lengths of encodings of the source

3.1 Information and Entropy

- Define a number $I(s_i)$, for each $s_i \in S$, which represents
  - How much information is gained by knowing that S has emitted s_i
  - Our prior uncertainty as to whether s_i will be emitted and our surprise on learning that it has been emitted
- Therefore require that:
  1) $I(s_i)$ is a decreasing function of the probability p_i of s_i, with $I(s_i) = 0$ if $p_i = 1$;
  2) $I(s_i s_j) = I(s_i) + I(s_j)$, where S emits s_i and s_j consecutively and independently.

Entropy Function

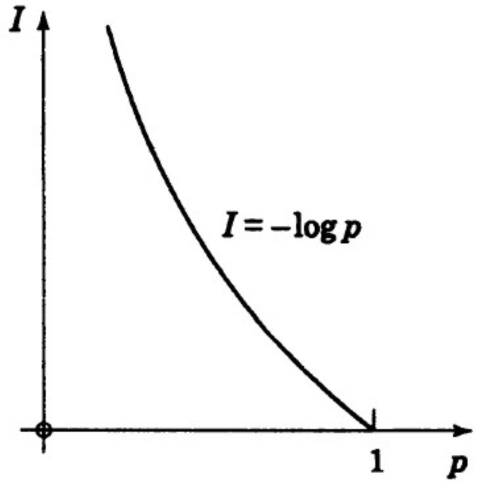

Figure 3.1

- We define

$$I(s_i) = -\log p_i = \log \frac{1}{p_i} \quad (3.1)$$

where $p_i = \Pr(s_i)$. So that I satisfies (1) and (2)

- Example 3.1
  - Let S be an unbiased coin, with s_1 and s_2 representing heads and tails. Then $I(s_1) = ?$ and $I(s_2) = ?$

The r-ary Entropy of S

- The average amount of information conveyed by S (per source-symbol) is given by the function

$$H_r(\mathcal{S}) = \sum_{i=1}^{q} p_i I_r(s_i) = \sum_{i=1}^{q} p_i \log_r \frac{1}{p_i} = -\sum_{i=1}^{q} p_i \log_r p_i$$

- Called the r-ary entropy of S.
- Base r is often omitted

$$H(\mathcal{S}) = \sum_{i=1}^{q} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{q} p_i \log p_i$$

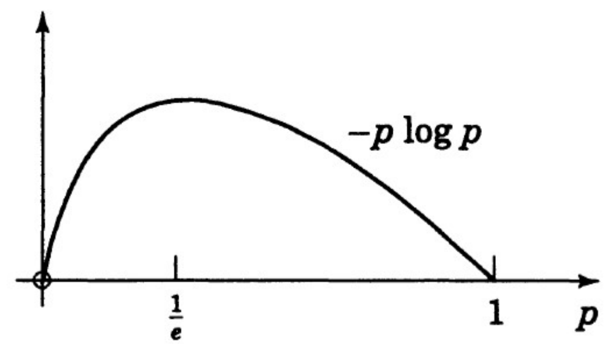

Figure 3.2

Examples

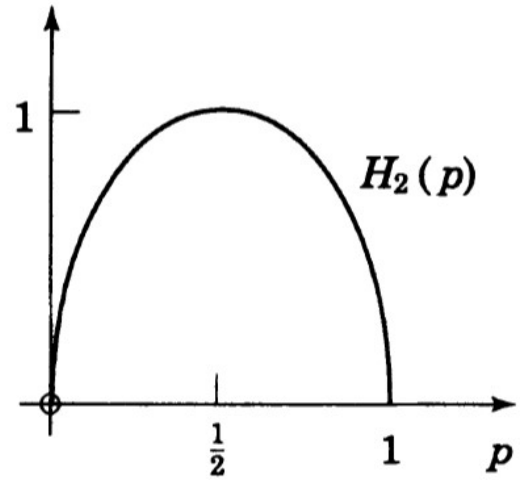

Figure 3.3

- Example 3.2
  - Let S have q = 2 symbols, with probabilities p and 1 - p
  - Let $\bar{p} = 1 - p$. Then

$$H(S) = -p \log p - \bar{p} \log \bar{p} . \qquad H(p) = -p \log p - \bar{p} \log \bar{p}.$$

  - $H(p)$ is maximal when p = ½
  - Compute $H(p)$ when p = ½ and p = 2⁄3
- Example 3.3
  - If S has q = 5 symbols with probabilities p_i = 0.3, 0.2, 0.2, 0.2, 0.1, as in §2.2, Example 2.5, we find that $H_2(S) = 2.246$.

Examples (Cont.)

- If S has q equiprobable symbols, then $p_i = {}^1/_q$ for each i, so

$$H_r(\mathcal{S}) = q \cdot \frac{1}{q} \log_r q = \log_r q .$$

- Example 3.4 and 3.5
  - Let $q = 5$, $H_2(S) = log_2 5 \approx 2.321$
  - Let $q = 6$, $H_2(S) = log_2 6 \approx 2.586$
- Example 3.6.
  - Using the known frequencies of the letters of the alphabet, the entropy of English text has been computed as approximately 4.03.

Compare average word-length of binary Huffman coding with entropy

- As in Example 3.2 with $p = {}^{2}/_{3}$
  - $H_2(S) \approx 0.918$
  - $L(C^1) \approx 1, L(C^2)/2 \approx 0.944, L(C^3)/3 \approx 0.938$
- As in Example 3.3
  - $H_2(S) \approx 2.246$
  - $L(C^1) \approx 2.3$
- As in Example 3.4
  - $H_2(S) \approx 2.246$
  - $L(C^1) \approx 2.321$

3.2 Properties of the Entropy Function

- Theorem 3.7
  - $H_r(S) \geq 0$, with equality if and only if $p_i = 1$ for some i (so that $p_j = 0$ for all $j \neq i$).
- Lemma 3.8
  - For all $x > 0$ we have $\ln x \leq x - 1$, with equality if and only if $x = 1$.
  - Converting to some other base r, we have

$$log_r(x) \leq log_r(e) \cdot (x - 1)$$

with equality if and only if $x = 1$.

Properties of the Entropy Function

- Corollary 3.9
  - Let $x_i \geq 0$ and $y_i > 0$ for i =1, ..., q, and let $\sum_i x_i = \sum_i y_i = 1$ (so (x_i) and (y_i) are probability distributions, with $y_i \neq 0$). Then

$$\sum_{i=1}^{q} x_i \log_r \frac{1}{x_i} \leq \sum_{i=1}^{q} x_i \log_r \frac{1}{y_i} \, ,$$

  - (that is, $\sum_i x_i \log(y_i/x_i) \leq 0$), with equality if and only if $x_i = y_i$ for all i.
- Theorem 3.10
  - If a source S has q symbols then $H_r(S) \leq log_r q$, with equality if and only if the symbols are equiprobable.

3.3 Entropy and Average Word-length

- Theorem 3.11
  - If C is any uniquely decodable r-ary code for a source S, then $L(C) \geq H_r(S)$.
- The interpretation
  - Each symbol emitted by S carries $H_r(S)$ units of information, on average.
  - Each code-symbol conveys one unit of information, so on average each code-word of C must contain at least $H_r(S)$ code-symbols, that is, $L(C) \geq H_r(S)$.
  - In particular, sources emitting more information require longer code-words.

Entropy and Average Word-length (Cont.)

- Corollary 3.12
  - Given a source S with probabilities p_i, there is a uniquely decodable r-ary code C for S with $L(C) = H_r(S)$ if and only if $log_r(p_i)$ is an integer for each i , that is, each $p_i = r^{e_i}$ for some integer $e_i \leq 0$.
- Example 3.13
  - If S has q = 3 symbols s_i, with probabilities p_i = 1⁄4,1⁄2, and 1⁄4 (see Examples 1.2 and 2.1).
  - $H_2(S) =$
  - A binary Huffman code C for S:
  - $L(C) =$

More examples

- Example 3.14
  - Let S have q = 5 symbols, with probabilities $p_i = 0.3, 0.2, 0.2, 0.2, 0.1$, as in Example 2.5.
    - In Example 3.3, $H_2(S)$ = 2.246, and
    - in Example 2.5, $L(C) = 2.3$, C binary Huffman code for S
  - By Theorem 2.8, every uniquely decodable binary code C for S satisfies $L(C) \geq 2.3 > H_2(S)$.
  - Thus no such code satisfies $L(C) = H_r(S)$
  - What is the reason?
- Example 3.15
  - Let S have 3 symbols s_i, with probabilities $p_i = \frac{1}{2}, \frac{1}{2}, 0$.

Code Efficiency and Redundancy

- If C is an r-ary code for a source S, its efficiency is defined to be

$$\eta = \frac{H_r(\mathcal{S})}{L(\mathcal{C})}, \tag{3.4}$$

  - So $0 \le \eta \le 1$ for every uniquely decodable code C for S
- The redundancy of C is defined to be $\bar{\eta} = 1 - \eta$.
  - Thus increasing redundancy reduces efficiency
- In Examples 3.13 and 3.14,
  - $\eta = 1$ and $\eta \approx 0.977$, respectively.

3.4 Shannon-Fano Coding

- Shannon-Fano codes
  - close to optimal, but easier to estimate their average word lengths.
- A Shannon-Fano code C for S has word lengths

$$l_i = \lceil \log_r(1/p_i) \rceil , \qquad (3.5)$$

- So, we have

$$\log_r \frac{1}{p_i} \leq l_i < 1 + \log_r \frac{1}{p_i} , \quad (3.6)$$

$$K = \sum_{i=1}^{q} r^{-l_i} \leq \sum_{i=1}^{q} p_i = 1,$$

So Theorem 1.20 (Kraft's inequality) implies that there is an instantaneous r-ary code C for S with these word-lengths l_i

Shannon-Fano Coding (Cont.)

- Theorem 3.16
  - Every r-ary Shannon-Fano code C for a source S satisfies

$$H_r(\mathcal{S}) \leq L(\mathcal{C}) \leq 1 + H_r(\mathcal{S})$$

- Corollary 3.17
  - Every optimal r-ary code D for a source S satisfies

$$H_r(\mathcal{S}) \leq L(\mathcal{D}) \leq 1 + H_r(\mathcal{S})$$

- Compute word length l_i of Shannon-Fano Code

$$l_i = \lceil \log_2(1/p_i) \rceil = \min\{n \in \mathbf{Z} \mid 2^n \geq 1/p_i\}$$

Examples

- Example 3.18
  - Let S have 5 symbols, with probabilities p_i= 0.3, 0.2, 0.2, 0.2, 0.1 as in Example 2.5
  - Compute Shannon-Fano code word length l_i, $L(C)$, η.
  - Compare with Huffman code.
- Example 3.19
  - If p_1 = 1 and p_i = 0 for all i > 1, then $H_r(S) = 0$. An r-ary optimal code D for S has average word-length $L(D) = 1$, so here the upper bound 1 + $H_r(S)$ is attained.

3.5 Entropy of Extensions and Products

- Recall from §2.6
  - S^n has q^n symbols $s_{i_1} \dots s_{i_n}$ with probabilities $p_{i_1} \dots p_{i_n}$.
- Theorem 3.20
  - If S is any source then $H_r(S^n) = nH_r(S)$.
- Lemma 3.21
  - If S and T are independent sources then $H_r(S \times T) = H_r(S) + H_r(T)$
- Corollary 3.22
  - If $S_1, \dots, S_n$ are independent sources then

    $$H_r(S_1 \times \cdots \times S_n) = H_r(S_1) + \cdots + H_r(S_n)$$

3.6 Shannon's First Theorem

- Theorem 3.23
  - By encoding S^n with n sufficiently large, one can find uniquely decodable r-ary encodings of a source S with average word-lengths arbitrarily close to the entropy $H_r(S)$.
- Recall that
  - if a code for S^n has average word-length L_n, then as an encoding of S it has average word-length L_n/n.
- Note that
  - the encoding process of S^n for a large n are complicated and time-consuming.
  - the decoding process involves delays

3.7 An Example of Shannon's First Theorem

- Let S be a source with two symbols s_1, s_2 of probabilities $p_i = 2/3, 1/3$, as in Example 3.2.
  - In §3.1, we have $\boxed{H_2(\mathcal{S}) = \log_2 3 - \frac{2}{3} \approx 0.918}$
  - In §2.6, using binary Huffman codes for S^n with n = 1, 2 and 3, we have $\boxed{L_n/n \approx 1, 0.944 \text{ and } 0.938}$
  - For larger n it is simpler to use Shannon-Fano codes, rather than Huffman codes.
    - Compute L_n for S^n
    - Verify $L_n/n \to H_2(S)$
    - You will need to use this formula $(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$