Example: Pikachu's (source) encoding

source → AABCAD → message → Source Encoder → pi pi pi pi pi ka pi chu
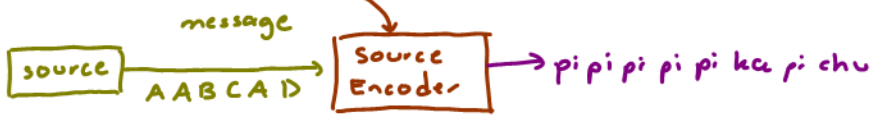
Source Alphabet
$\{A, B, C, D\}$

Each of these is called a source symbol.

x	$c(x)$
A	pipi
B	pika
C	pichu
D	ka

} dictionary or codebook

Each of these is a codeword

Each of these is a code symbol

code alphabet: $\{pi, ka, chu\}$

For some groups, the code alphabet is
$\{pi, ka, chu, \underline{\ }\}$

'Pause' between consecutive codeword

Example: English text

source → message → source encoder → 000100110

In our example, we used ASCII Encoder.

"codebook"

x	$c(x)$
⋮	﹏
E	1000101
⋮	﹏

Source Alphabet
$\{a, b, c, \ldots, z,$
$A, B, C, \ldots, Z,$
$1, 2, 3, \ldots, 9, 0,$
$?, @, \ldots\}$

Code Alphabet
$\{0, 1\}$

2   Source Coding

In this chapter, we look at the "source encoder" part of the system. This part removes redundancy from the message stream or sequence. We will focus only on <u>binary</u> source coding.

2.1. The material in this chapter is based on [C & T Ch 2, 4, and 5].

2.1   General Concepts

Example 2.2. Suppose your message is a paragraph of (written natural) text in English.

- Approximately 100 possibilities for characters/symbols.

    - For example, a character-encoding scheme called **ASCII** (American Standard Code for Information Interchange) originally[1] had 128 specified characters – the numbers 0–9, the letters a–z and A–Z, some basic punctuation symbols[2], and a blank space.

- Do we need $\underset{=}{7}$ bits per characters?

    100 characters ⟶ 700 bits   too many?

2.3. A sentence of English–or of any other language–always has more information than you need to decipher it. The meaning of a message can remain unchanged even though parts of it are removed.

Example 2.4.

- "J-st tr- t- r–d th-s s-nt-nc-." [3]

- "Thanks to the redundancy of language, yxx cxn xndxrstxnd whxt x xm wrxtxng xvxn xf x rxplxcx xll thx vxwxls wxth xn 'x' (t gts lttl hrdr f y dn't vn kn whr th vwls r)." [4]

---

[1]Being American, it didn't originally support accented letters, nor any currency symbols other than the dollar. More advanced Unicode system was established in 1991.

[2]There are also some control codes that originated with Teletype machines. In fact, among the 128 characters, 33 are non-printing control characters (many now obsolete) that affect how text and space are processed and 95 printable characters, including the space.

[3]Charles Seife, Decoding the Universe. Penguin, 2007

[4]Steven Pinker, The Language Instinct: How the Mind Creates Language. William Morrow, 1994

2.5. It is estimated that we may only need about 1 bits per character in English text.

Simplified model of the source (to make the analysis easier)

Definition 2.6. Discrete Memoryless Sources (DMS): Let us be more specific about the information source.

$$\boxed{\text{Source}} \rightarrow X_1, X_2, X_3, \ldots$$

- The message that the information source produces can be represented by a **vector** of characters $X_1, X_2, \ldots, X_n$.

  ○ A perpetual message source would produce a never-ending **sequence** of characters $X_1, X_2, \ldots$.

- These X_k's are random variables (at least from the perspective of the decoder; otherwise, these is no need for communication).

- For simplicity, we will assume our source to be discrete and memoryless.

  ○ Assuming a **discrete** source means that the random variables are all discrete; that is, they have supports which are countable.

    * Recall that "countable" means "finite" or "countably infinite".
    * We will further assume that they all share the same support and that the support is finite.
      · This support is called the **source alphabet**. = *set of all possible source symbols*
      · See Example 2.7 for some examples.

  ○ Assuming a **memoryless** source means that there is no dependency among the characters in the sequence.

    * More specifically,

$$p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1) \times p_{X_2}(x_2) \times \cdots \times p_{X_n}(x_n). \tag{1}$$

Recall: Two discrete RVs X and Y are independent iff

joint pmf   *marginal pmfs*

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$$

    * Practical sources would *not* be memoryless; there are some amount of dependence (structure) among the characters. For English text, this is demonstrated in Example 2.4.
      · Simple DMS model provides a good starting point to study.
      · We can take advantage of such dependency.

∗ We will further assume that all of the random variables share the same probability mass function (pmf)[5]. We denote this shared pmf by $p_X(x)$.

[are identically distributed]

$$p_{X_k}(x) \equiv p_X(x) = p(x)$$

In which case, (1) becomes

$$p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = p_X(x_1) \times p_X(x_2) \times \cdots \times p_X(x_n). \quad (2)$$

· We will also assume that the pmf $p_X(x)$ is known. In practice, there is an extra step of estimating this $p_X(x)$.

· To save space, we may see the pmf $p_X(x)$ written simply as $p(x)$, i.e. without the subscript part.

∗ The shared support of X which is usually denoted by S_X becomes the source alphabet. Note that we also often see the use of $\mathcal{X}$ to denote the support of X.

• Summary: A DMS produces a sequence (symbol by symbol) of i.i.d. RVs $X_1, X_2, \ldots$ all of which share the same pmf $p_X(x)$ whose support is called the source alphabet.

• Because our simplified source code can be characterized by a random variable X, we only need to specify its pmf $p_X(x)$.
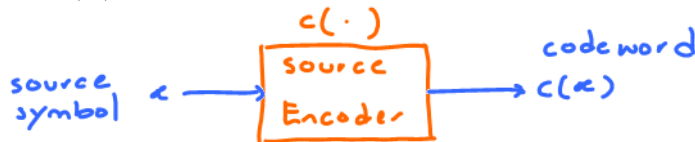
Example 2.7. Examples of (finite) source alphabets

(a) Collection of 95 symbols for English text.

(b) Collection of 128 symbols for string of ASCII symbols.

(c) Collection of four symbols {Yes, No, OK, Thank You} for crude conversation with Farang.

(d) Collection of four symbols {A, B, C, D} for answers of multiple-choice test.

---

[5]We often use the term "distribution" interchangably with pmf and pdf; that is, instead of saying "pmf of X", we may say "distribution of X".

Definition 2.8. An **encoder** $c(\cdot)$ is a function that maps each of the ~~char~~ ~~acter~~ in the source alphabet into a corresponding (binary) codeword.

(handwritten: source; symbol; If we assume binary codewords, then code alphabet $= \{0,1\}$)

- In particular, the codeword corresponding to a source character x is denoted by $c(x)$.

(handwritten diagram: source symbol → [Source Encoder $c(\cdot)$] → codeword $c(x)$)
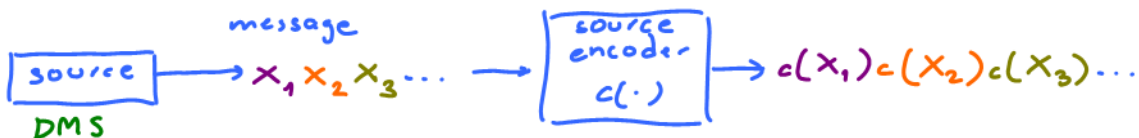
- Each codeword is constructed from a **code alphabet**.
  - A binary codeword is constructed from a two-symbol alphabet, wherein the two symbols are usually taken as 0 and 1.
  - It is possible to consider non-binary codeword. Morse code discussed in Example 2.13 is one such example.

- Mathematically, we write

$$\text{Encoder } c : S_X \to \{0,1\}^*$$

  where

$$\{0,1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, \ldots\}$$

  is the set of all finite-length binary strings.

(handwritten diagram: [source DMS] → message $x_1 x_2 x_3 \cdots$ → [source encoder $c(\cdot)$] → $c(x_1) c(x_2) c(x_3) \cdots$)

- The length of the codeword associated with source character x is denoted by $\ell(x)$.
  - In fact, writing this as $\ell(c(x))$ may be clearer because we can see that the length depends on the choice of the encoder. However, we shall follow the notation above[6].

Example 2.9. $c(\text{red}) = 00$, $c(\text{blue}) = 11$ is a source code for $S_X = \{\text{red}, \text{blue}\}$.

---

[6]which is used by the standard textbooks in information theory.

11

Example 2.10. Suppose the message is a sequence of basic English words which happen according to the probabilities provided in the table below.

x	$p(x)$	Codeword $c(x)$	$\ell(x)$
Yes	4%	00	2
No	3%	01	2
OK	90%	10	2
Thank You	3%	11	2

$$\mathbb{E}[\ell(X)] = \sum_x p(x)\,\underbrace{\ell(x)}_{2}$$
$$= 2\left(\underbrace{\sum_x p_x(x)}_{1}\right)$$
$$= 2 \quad \text{bits/source symbol}$$

Definition 2.11. The **expected length** of a code $c(\cdot)$ for (a DMS source which is characterized by) a random variable X with probability mass function $p_X(x)$ is given by

$$\mathbb{E}[\ell(X)] = \sum_{x \in S_X} p_X(x)\ell(x).$$

Example 2.12. Back to Example 2.10. Consider a new encoder:

x	$p(x)$	Codeword $c(x)$	$\ell(x)$
Yes	4%	01	2
No	3%	001	3
OK	90%	1	1
Thank You	3%	0001	4

$$\mathbb{E}[\ell(X)] = 2(0.04) + 3(0.03)$$
$$+ 1(0.90) + 4(0.03)$$
$$= 1.19 \quad \text{bits/source symbol}$$

Observe the following:

- Data compression can be achieved by assigning short descriptions to the most frequent outcomes of the data source, and necessarily longer descriptions to the less frequent outcomes.

- When we calculate the expected length, we don't really use the fact that the source alphabet is the set {Yes, No, OK, Thank You}. We would get the same answer if it is replaced by the set $\{1, 2, 3, 4\}$, or the set $\{a, b, c, d\}$. All that matters is that the alphabet size is 4, and the corresponding probabilities are $\{0.04, 0.03, 0.9, 0.03\}$.

  Therefore, for brevity, we often find DMS source defined only by its alphabet size and the list of probabilities.

Example 2.13. The Morse code is a reasonably efficient code for the English alphabet using an alphabet of four symbols: a dot, a dash, a letter space, and a word space. [See Slides]

- <mark>Short sequences represent frequent letters</mark> (e.g., a single dot represents E) and <mark>long sequences represent infrequent letters</mark> (e.g., Q is represented by "dash,dash,dot,dash").

Example 2.14. Thought experiment: Let's consider the following code

x	$p(x)$	Codeword $c(x)$	$\ell(x)$
1	4%	0	1
2	3%	1	1
3	90%	0	1
4	3%	1	1

$\mathbb{E}\left[\ell(x)\right] = 1$ bit/source symbol)

Not nonsingular

singular

This code is bad because we have ambiguity at the decoder. When a codeword "0" is received, we don't know whether to decode it as the source symbol "1" or the source symbol "3". If we want to have <mark>lossless source coding</mark>, this ambiguity is not allowed.

Definition 2.15. A code is <mark>nonsingular</mark> if <mark>every source symbol in the source alphabet has different codeword</mark>.

Equivalently, for any $\alpha_1, \alpha_2 \in S_X$, if $\alpha_1 \neq \alpha_2$, then $c(\alpha_1) \neq c(\alpha_2)$.

As seen from Example 2.14, nonsingularity is an important concept. However, it turns out that this property is not enough.

Example 2.16. Another thought experiment: Let's consider the following code

x	$p(x)$	Codeword $c(x)$	$\ell(x)$
1	4%	01	2
2	3%	010	3
3	90%	0	1
4	3%	10	2

$\mathbb{E}\left[\ell(x)\right] = 2(0.04) + 3(0.03)$
$\qquad + 1(0.90) + 2(0.03)$
$= 1.13$

$010 = "2"$
$010 \rightarrow 01|0 = "34"$ not UD
$01|0 = "13"$

2.17. We usually wish to convey a sequence (string) of source symbols. So, we will need to consider **concatenation** of codewords; that is, if our source string is

$$X_1, X_2, X_3, \ldots$$

then the corresponding encoded string is

$$c(X_1)c(X_2)c(X_3)\cdots.$$

In such cases, to ensure decodability, we may

(a) use fixed-length code (as in Example 2.10), or

(b) use variable-length code and

    (i) add a special symbol (a "comma" or a "space") between any two codewords

$$0 \text{✶} 10 \text{✶} 010 \text{✶} \cdots$$

    or

    (ii) use uniquely decodable codes.

Definition 2.18. A code is called **uniquely decodable** (UD) if any encoded string has only one possible source string producing it.

Example 2.19. The code used in Example 2.16 is not uniquely decodable because source string "2", source string "34", and source string "13" share the same code string "010".

2.20. It may not be easy to check unique decodability of a code. (See Example 2.28.) Also, even when a code is uniquely decodable, one may have to look at the entire string to determine even the first symbol in the corresponding source string. Therefore, we focus on a subset of uniquely decodable codes called prefix code.

Definition 2.21. A code is called a **prefix code** if no codeword is a prefix[7] of any other codeword.

- Equivalently, a code is called a **prefix code** if you can put all the codewords into a binary tree where all of them are leaves.

- A more appropriate name would be "**prefix-free**" code.

- The codeword corresponding to a symbol is the string of labels on the path from the root to the corresponding leaf.

Example 2.22.

x	Codeword $c(x)$
1	10
2	110
3	0
4	111

So, this code is a prefix code.

$$1\,0\,1\,1\,0\,0\,1\,1\,0\,0$$
$$1\quad 2\quad 3\quad 2\quad 1$$

---
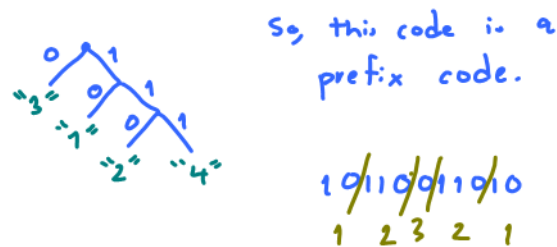
[7]String s_1 is a prefix of string s_2 if there exist a string s_3, possibly empty, such that $s_2 = s_1 s_3$.

Example 2.23. The code used in Example 2.12 <u>is</u> a prefix code.

x	Codeword $c(x)$
1	01
2	001
3	1
4	0001

2.24. Any prefix code is uniquely decodable.

- The end of a codeword is immediately recognizable.

- Each source symbol can be decoded as soon as we come to the end of the codeword corresponding to it. In particular, we need not wait to see the codewords that come later.

- Therefore, another name for "prefix code" is **instantaneous code**.

Example 2.25. The codes used in Example 2.12 (Example 2.23) and Example 2.22 are prefix codes and hence they are uniquely decodable.

2.26. The nesting relationship among all the types of source codes is shown in Figure 2.
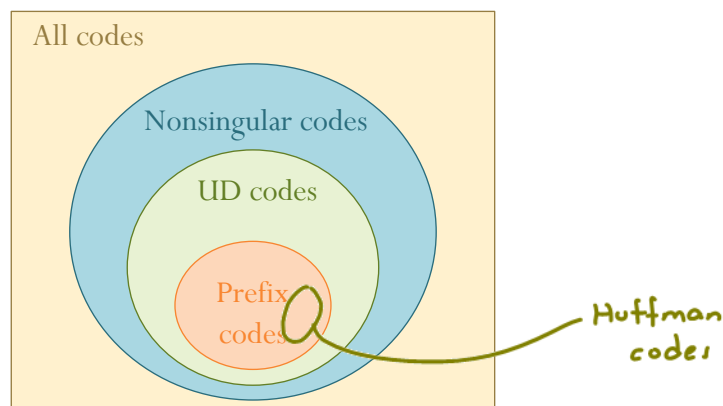

Figure 2: Classes of codes

Example 2.27.

x	Codeword $c(x)$
1	1
2	10
3	100
4	1000

observe that a "1" signifies
the beginning of every
codeword.

Problem: Need to look at
the first bit of
the next codeword
to determine the
end of the current
codeword.

Try to decode 100100011101001 11

Example 2.28. [5, p 106–107]

x	Codeword $c(x)$
1	10
2	00
3	11
4	110

This code is not a prefix code because codeword "11" is a prefix of codeword "110".

This code is uniquely decodable. To see that it is uniquely decodable, take any code string and start from the beginning.

- If the first two bits are 00 or 10, they can be decoded immediately.

- If the first two bits are 11, we must look at the following bit(s).

  ○ If the next bit is a 1, the first source symbol is a 3.

  ○ If the next bit is a 0, we need to count how many 0s are there before 1 shows up again.

  ○ If the length of the string of 0's immediately following the 11 is even, the first source symbol is a 3.

  ○ If the length of the string of 0's immediately following the 11 is odd, the first codeword must be 110 and the first source symbol must be 4.

By repeating this argument, we can see that this code is uniquely decodable.

16

ECS452 2018/2   Part I.2   Dr.Prapun

2.29. For our present purposes, a better code is one that is uniquely decodable and has a shorter expected length than other uniquely decodable codes. We do not consider other issues of encoding/decoding complexity or of the relative advantages of block codes or variable length codes. [6, p 57]

2.2   Optimal Source Coding: Huffman Coding

In this section we describe a very popular source coding algorithm called the Huffman coding.

Definition 2.30. Given a source with known probabilities of occurrence for symbols in its alphabet, to construct a binary Huffman code, create a binary tree by repeatedly combining[8] the probabilities of the two least likely symbols.

- Developed by David Huffman as part of a class assignment[9].

steps
① Combine the two least likely symbols
② Repeat ① until you cannot repeat.

---

[8]The Huffman algorithm performs _repeated source reduction_ [6, p 63]:

- At each step, two source symbols are combined into a new symbol, having a probability that is the sum of the probabilities of the two symbols being replaced, and the new reduced source now has one fewer symbol.

- At each step, the two symbols to combine into a new symbol have the two lowest probabilities.

   ◦ If there are more than two such symbols, select any two.

[9]The class was the first ever in the area of information theory and was taught by Robert Fano at MIT in 1951.

   ◦ Huffman wrote a term paper in lieu of taking a final examination.

   ◦ It should be noted that in the late 1940s, Fano himself (and independently, also Claude Shannon) had developed a similar, but suboptimal, algorithm known today as the ShannonFano method. The difference between the two algorithms is that the ShannonFano code tree is built from the top down, while the Huffman code tree is constructed from the bottom up.

- By construction, Huffman code is a prefix code.

Example 2.31.

$\frac{1}{2^1} = 1/2$

$\frac{1}{2^2} = 1/4$

$\frac{1}{2^3} = 1/8$

$\frac{1}{2^3} = 1/8$

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
A	0.5	0	1
B	0.25	10	2
C	0.125	110	3
D	0.125	111	3

$\mathbb{E}[\ell(X)] = 1(0.5) + 2(0.25) + 3(0.125) + 3(0.125) = 1.75$ bits/source symbol

Note that for this particular example, the values of $2^{\ell(x)}$ from the Huffman encoding is inversely proportional to $p_X(x)$:

$$p_X(x) = \frac{1}{2^{\ell(x)}}.$$

In other words,

$$\ell(x) = \log_2 \frac{1}{p_X(x)} = -\log_2(p_X(x)).$$

Therefore,

$$\mathbb{E}[\ell(X)] = \sum_x p_X(x)\ell(x) = \sum_x p_X(x)\left(-\log_2(p_X(x))\right) = -\sum_x p_X(x)\log_2 p_X(x)$$

Example 2.32.

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
'a'	0.4	0	1
'b'	0.3	10	2
'c'	0.1	110	3
'd'	0.1	1110	4
'e'	0.06	11110	5
'f'	0.04	11111	5

$\mathbb{E}[\ell(X)] = 1(0.4) + 2(0.3) + 3(0.1) + 4(0.1) + 5(0.06) + 5(0.04)$

$= 2.2$ bits/source symbol

$-\sum_x p_X(x)\log_2 p_X(x) = -\left(0.4\log_2 0.4 + 0.3\log_2 0.3 + (2\times 0.1\log_2 0.1) + 0.06\log_2 0.06 + 0.04\log_2 0.04\right)$

18

$= 2.1435 < 2.2$

Example 2.33.

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
1	0.25	10	2
2	0.25	00	2
3	0.2	01	2
4	0.15	110	3
5	0.15	111	3

$$\mathbb{E}\left[\ell(X)\right] = 2.3 \quad \text{bits/symbol}$$

Example 2.34.

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
A	1/3	00	2
B	1/3	01	2
C	1/4	10	2
D	1/12	11	2

$$\mathbb{E}\left[\ell(X)\right] = \mathbb{E}\left[2\right] = 2 \quad \text{bits/source symbol}$$

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
i	1/3	0	1
ii	1/3	10	2
iii	1/4	110	3
iv	1/12	111	3

$$\mathbb{E}\left[\ell(X)\right] = 1\times\frac{1}{3} + 2\times\frac{1}{3} + 3\times\left(\frac{1}{4}+\frac{1}{12}\right) = 2 \quad \text{bits/source symbol}$$

2.35. The set of codeword lengths for Huffman encoding is not unique. There may be more than one set of lengths but all of them will give the same value of expected length.

Definition 2.36. A code is **optimal** for a given source (with known pmf) if it is uniquely decodable and its corresponding expected length is the shortest among all possible uniquely decodable codes for that source.

2.37. The Huffman code is optimal.

2.3 Source Extension (Extension Coding)

2.38. One can usually (not always) do better in terms of expected length (per source symbol) by encoding blocks of several source symbols.
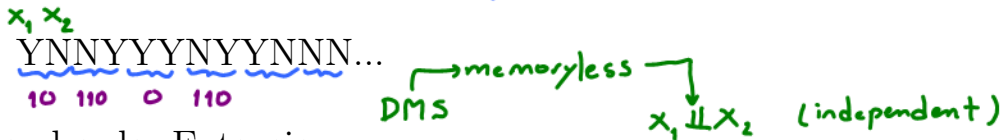
Definition 2.39. In, an **n-th extension** coding, n successive source symbols are grouped into blocks and the encoder operates on the blocks rather than on individual symbols. [4, p. 777]

Example 2.40.

x	$p_X(x)$	Codeword $c(x)$	$\ell(x)$
Y(es)	0.9 ⟩0,1	0	1
N(o)	0.1	1	1

(a) First-order extension:

$$\mathbb{E}[\ell(X)] = \mathbb{E}[1] = 1 \quad \text{bit/source symbol.}$$

$$\overset{x_1\,x_2}{\underbrace{YN}_{10}\underbrace{NY}_{110}\underbrace{YY}_{0}\underbrace{YNNN}_{110}\ldots} \quad \longrightarrow \text{memoryless} \longrightarrow$$

$$\text{DMS} \qquad X_1 \perp\!\!\!\perp X_2 \ \text{(independent)}$$

(b) Second-order Extension:

$x_1 x_2$	$p_{X_1,X_2}(x_1,x_2) = P_{X_1}(x_1)P_{X_2}(x_2)$	$c(x_1,x_2)$	$\ell(x_1,x_2)$
YY	$0.9 \times 0.9 = 0.81$	0	1
YN	$0.9 \times 0.1 = 0.09$	10	2
NY	$0.1 \times 0.9 = 0.09$	110	3
NN	$0.1 \times 0.1 = 0.01$	111	3

Let L_n be the expected length per (one) source symbol when n^{th}-extension of Huffman coding is applied.

$$\mathbb{E}[\ell(X_1,X_2)] = 1 \times 0.81 + \underbrace{2 \times 0.09}_{0.18} + \underbrace{3 \times (\overbrace{0.09+0.01}^{0.1})}_{0.3} = 1.29 \quad \text{bits/2 source symbols}$$

$$L_2 = \frac{1.29}{2} = 0.645 \ \text{bits per source symbol.}$$

(c) Third-order Extension:

$x_1 x_2 x_3$	$p_{X_1,X_2,X_3}(x_1,x_2,x_3)$	$c(x_1,x_2,x_3)$	$\ell(x_1,x_2,x_3)$
YYY	$0.9 \times 0.9 \times 0.9 = 0.729$		
YYN	$0.9 \times 0.9 \times 0.1 = 0.081$		
YNY	$0.9 \times 0.1 \times 0.9 = 0.081$		
$\vdots$	[HW]		

$$\mathbb{E}[\ell(X_1,X_2,X_3)] = 1.5930 \ \text{bits per 3 source symbols}$$

$$L_3 = \frac{1.5930}{3} = 0.5327 \ \text{bits per source symbol.}$$

20

ECS452 2018/2   Part I.3   Dr.Prapun

2.4 (Shannon) Entropy for Discrete Random Variables

Entropy is a measure of uncertainty of a random variable [5, p 13].

Entropy quantifies/measures the amount of uncertainty in a RV (randomness)

It arises as the answer to a number of natural questions. One such question that will be important for us is "What is the average length of the shortest *description* of the random variable?"

Definition 2.41. The **entropy** $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in S_X} p_X(x) \log_2 p_X(x) = -\mathbb{E}[\log_2 p_X(X)].$$

Recall: $\log_2 a = \dfrac{\log_{10} a}{\log_{10} 2} = \dfrac{\log_b a}{\log_b 2} = \dfrac{\ln a}{\ln 2}$

- The log is to the base 2 and entropy is expressed in bits (per symbol).

  ○ The base of the logarithm used in defining H can be chosen to be any convenient real number $b > 1$ but if $b \neq 2$ the unit will not be in bits.

  ○ If the base of the logarithm is e, the entropy is measured in nats.

  ○ Unless otherwise specified, base 2 is our default base.

- Based on continuity arguments, we shall assume that $0 \ln 0 = 0$.

Back then, the probabilities were $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$

Example 2.42. The entropy of the random variable X in Example 2.31 is 1.75 bits (per symbol).

$$H(X) = -\left(\tfrac{1}{2}\log_2\tfrac{1}{2} + \tfrac{1}{4}\log_2\tfrac{1}{4} + 2\times\tfrac{1}{8}\log_2\tfrac{1}{8}\right) = \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{4} = 1.75 \text{ bits}/\text{symbol}$$

Example 2.43. The entropy of a fair coin toss is 1 bit (per toss).

The probabilities involved are $\frac{1}{2}, \frac{1}{2}$.

$$H(X) = -\left(\tfrac{1}{2}\log_2\tfrac{1}{2} + \tfrac{1}{2}\log_2\tfrac{1}{2}\right) = 1 \text{ bit}/\text{symbol}$$

2.44. Note that entropy is a functional of the (unordered) probabilities from the pmf of X. It does not depend on the actual values taken by the random variable X, Therefore, sometimes, we write $H(p_X)$ instead of $H(X)$ to emphasize this fact. Moreover, because we use only the probability values, we can use the row vector representation $\underline{\mathbf{p}}$ of the pmf p_X and simply express the entropy as $H(\mathbf{p})$.

In MATLAB, to calculate $H(X)$, we may define a row vector pX from the pmf p_X. Then, the value of the entropy is given by

$$\texttt{HX = -pX*(log2(pX))'}.$$

Example 2.45. The entropy of a uniform (discrete) random variable X on $\{1, 2, 3, \ldots, n\}$: $\rightarrow S_X$

$$p_X(x) = \begin{cases} 1/n, & x = 1, 2, \ldots, n, \\ 0, & \text{otherwise}. \end{cases}$$

$$H(X) = -\sum_x p_X(x)\log_2 p_X(x)$$
$$= -\sum_{x=1}^{n} \tfrac{1}{n}\log_2\tfrac{1}{n}$$
$$= -n\left(\tfrac{1}{n}\log_2\tfrac{1}{n}\right)$$
$$= \log_2 n = \log_2 |S_X|$$

Example 2.46. The entropy of a Bernoulli random variable X:

$$p_X(x) = \begin{cases} p, & x = 1, \\ 1-p & x = 0, \\ 0, & \text{otherwise}. \end{cases}$$

$$H(X) = -\left(p\log_2 p + (1-p)\log_2(1-p)\right)$$

Binary RV

$$p_X(x) = \begin{cases} p, & x = a, \\ 1-p, & x = b, \\ 0, & \text{otherwise}. \end{cases}$$

22

Definition 2.47. Binary Entropy Function : We define $h_b(p)$, $h(p)$ or $H(p)$ to be $-p \log_2 p - (1-p) \log_2 (1-p)$, whose plot is shown in Figure 3.
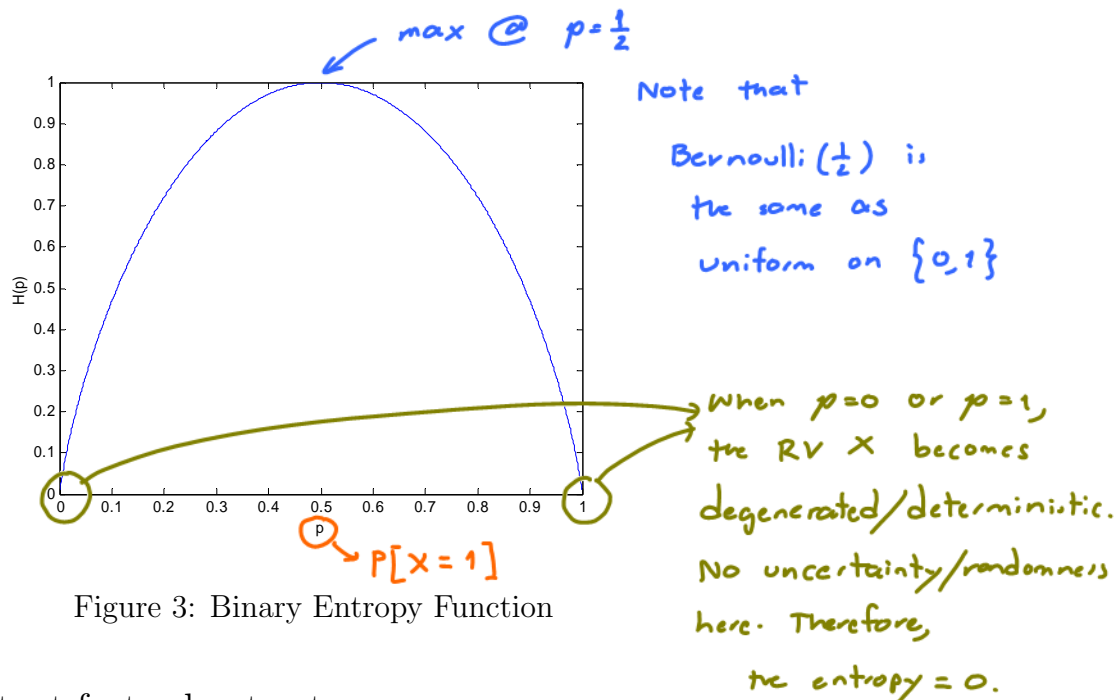

Figure 3: Binary Entropy Function

2.48. Two important facts about entropy:

(a) $H(X) \leq \log_2 |S_X|$ with equality if and only if X is a uniform random variable.

(b) $H(X) \geq 0$ with equality if and only if X is not random.

In summary,

$$\underbrace{0}_{\text{deterministic}} \leq H(X) \leq \underbrace{\log_2 |S_X|}_{\text{uniform}}.$$

Theorem 2.49. The expected length $\mathbb{E}[\ell(X)]$ of any uniquely decodable binary code for a random variable X is greater than or equal to the entropy $H(X)$; that is,

$$\mathbb{E}[\ell(X)] \geq H(X)$$

with equality if and only if $2^{-\ell(x)} = p_X(x)$. [5, Thm. 5.3.1]

Definition 2.50. Let $L(c, X)$ be the expected codeword length when random variable X is encoded by code c.

Let $L^*(X)$ be the minimum possible expected codeword length when random variable X is encoded by a uniquely decodable code c:

$$L^*(X) = \min_{\text{UD } c} L(c, X).$$

23

2.51. Given a random variable X, let c_{Huffman} be the Huffman code for this X. Then, from the optimality of Huffman code mentioned in 2.37,

$$L^*(X) = L(c_{\text{Huffman}}, X).$$

Theorem 2.52. The optimal code for a random variable X has an expected length less than $H(X) + 1$:

$$L^*(X) < H(X) + 1.$$

2.53. Combining Theorem 2.49 and Theorem 2.52, we have

$$H(X) \leq L^*(X) < H(X) + 1. \tag{3}$$

Definition 2.54. Let $L_n^*(X)$ be the minimum expected codeword length per symbol when the random variable X is encoded with n-th extension uniquely decodable coding. Of course, this can be achieve by using n-th extension Huffman coding.

2.55. An extension of (3):

$$H(X) \leq L_n^*(X) < H(X) + \frac{1}{n}. \tag{4}$$

In particular,

$$\lim_{n \to \infty} L_n^*(X) = H(X).$$

In otherwords, by using large block length, we can achieve an expected length per source symbol that is arbitrarily close to the value of the entropy.

2.56. Operational meaning of entropy: Entropy of a random variable is the average length of its shortest description.

2.57. References

- Section 16.1 in Carlson and Crilly [4]

- Chapters 2 and 5 in Cover and Thomas [5]

- Chapter 4 in Fine [6]

- Chapter 14 in Johnson, Sethares, and Klein [8]

- Section 11.2 in Ziemer and Tranter [18]