INFORMATION AND MINIMAL LENGTH CODING
6.02 / Nov. 30, 2010

BASICS:

GIVEN: A PROBABILITY DISTRIBUTION ON A FINITE ALPHABET

(Ex.: $\{$"A", "♂", "♄"$\}$, $P_A = \frac{1}{4}$, $P_♂ = \frac{1}{3}$, $P_♄ = \frac{5}{12}$)

WANT: A VARIABLE LENGTH BIT CODE, i.e

A CORRESPONDENCE ALPHABET $\longrightarrow$ BIT SEQUENCES SUCH THA THE SEQUENCE CORRESPONDING TO ONE ~~ALP~~ SYMBOL IS NEVER A BEGINNING OF A SEQUENCE CORRESPONDING TO ANOTHER SYMBOL

Ex.:

$$\left.\begin{array}{l} A \to 0 \\ ♂ \to 10 \\ ♄ \to 11 \end{array}\right\} \text{is A VALID } \cancel{\text{ENCODING}} \text{ CODE} \qquad \left.\begin{array}{l} A \to 0 \\ ♂ \to 1 \\ ♄ \to 01 \end{array}\right\} \text{is NOT A VALID CODE}$$

PERFORMANCE ("AVERAGE LENGTH"):

$$A.L. = \sum_{\text{SYMBOLS}} \mathbb{P}(\text{SYMBOL}) \cdot \text{LENGTH (SYMBOL ENCODING)}$$

(Ex.: $1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{3} + 2 \cdot \frac{5}{12} = \frac{7}{4}$ FOR THE CODE ABOVE)

USEFUL RESULTS:

$$* \quad A.L. \geq \underbrace{\sum_{\text{SYMBOLS}} \mathbb{P}(\text{SYMBOL}) \cdot \overbrace{\log_2 (1/\mathbb{P}(\text{SYMBOL}))}^{\text{"INFORMATION" DELIVERED BY THE SYMBOL}}}_{\text{"INFORMATION ENTROPY" / "SHANNON'S ENTROPY"}}$$

$$* \quad \overbrace{A.L.}^{\text{MINIMAL}} \leq 1 + \sum_s \mathbb{P}(s) \log_2 (1/\mathbb{P}(s))$$

* MINIMAL A.L. IS DELIVERED BY HUFFMAN CODES

* FOR THE ALPHABET OF n-LENGTH WORDS $w = (s_1, s_2, \dots, s_n)$, WITH $\mathbb{P}((s_1 s_2 \dots s_n)) = \mathbb{P}(s_1)\mathbb{P}(s_2)\dots\mathbb{P}(s_n)$ THE ENTROPY IS $n *$ (ORIGINAL ENTROPY)

i.e. THE ENTROPY IS OPTIMAL AVERAGE NUMBER OF BITS $(n \to \infty$

Ex. 1 YOU ARE TOLD THAT TWO INDEPENDENT TOSSES OF A FAIR DIE (SIDES 1,2,3,4,5,6) HAVE YIELDED A SUM OF 4. HOW MUCH INFO IS THIS?

$\underbrace{3 \cdot \text{POSSIBILITIES}}_{(1+3,\,2+2,\,3+1)}$ LEFT OUT OF 36

$\text{INFO} = \log_2(36/3) \approx 3.58$

(IF THE SUM = 7, INFO = $\log_2(36/6) \approx$ ~~...~~ $3.58 - 1 = 2.58$

Ex. 2 UNFAIR COIN. ($\mathbb{P}(\text{TAIL}) = 0.02$)

NEED: TO ENCODE THE OUTCOME OF 100 TOSSES, TO MINIMIZE A.L. MINIMAL A.L. = ?

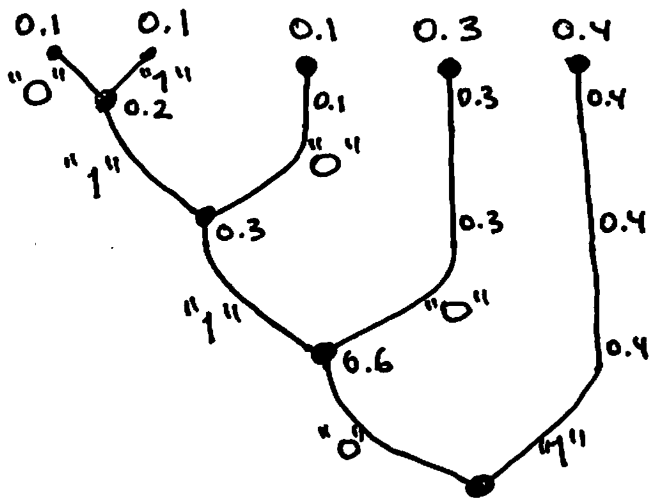
THE ALPHABET HAS 2^{100} SYMBOLS!

$$\text{ENTROPY} = 100 \cdot \left[0.02 \cdot \log \frac{1}{0.02} + 0.98 \cdot \log \frac{1}{0.98}\right] \approx 14.144$$

HENCE:

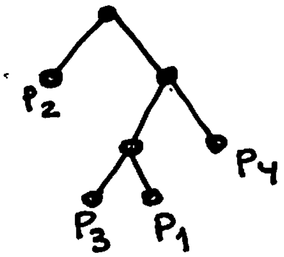
$$14.1 < \text{MINIMAL A.L} < 15.2$$

Ex. 3 HUFFMAN CODE FOR $(P_1, P_2, P_3, P_4, P_5) = (0.1, 0.1, 0.1, 0.3, 0.4)$



ASSIGN 0/1 ARBITRARILY AT EACH BRANCHING

Ex. 4



THIS IS A HUFFMAN CODE TREE. WHAT CAN BE SAID ABOUT P_1, P_2, P_3, P_4?

$$\left.\begin{array}{l} P_1 \le P_2 \\ P_1 \le P_4 \\ P_3 \le P_2 \\ P_3 \le P_4 \end{array}\right\}$$ BECAUSE P_1, P_3 ARE THE TWO LONGEST

$$\left.\begin{array}{l} P_1 + P_3 \le P_2 \\ P_4 \le P_2 \end{array}\right\}$$ BECAUSE (P_1, P_3) IS JOINED WITH P_4

IS THIS CONSISTENT WITH $P_2 = 0.3$?

NO, BECAUSE THEN $1 = P_1 + P_2 + P_3 + P_4 \le 3P_2 = 0.9$!

IS THIS CONSISTENT WITH $P_2 = 1/3$?

YES, HERE'S AN EXAMPLE: $P_2 = 1/3, P_4 = 1/3, P_1 = 1/6, P_3 = 1/6$

Ex. 5 $\mathbb{P}(S_1) = 1/2$

$\mathbb{P}(S_2) = \mathbb{P}(S_3) = \ldots = \mathbb{P}(S_{100}) = \frac{1}{198}$

WHAT IS THE LENGTH OF ENCODING S_1 IN THE HUFFMAN CODE?

ANSWER: 1

Ex. 6 $P_1 = 0.24, P_2 = 0.25, P_3 = 0.26, P_4 = 0.25$

HOW MANY OPTIMAL LENGTH ENCODINGS?

ANSWER: $4! = 24$