Lossless Data Compression

"Today is Wednesday" a message as a sequence of letter

a letter $\longleftrightarrow$ bytes $\{1 \cdots 256\} \longleftrightarrow$ 8 bits $\{0,1\}^8$

"a" 97 2^6 110 0001

See table below

Binary	Oct	Dec	Hex	Glyph	Binary	Oct	Dec	Hex	Glyph	Binary	Oct	Dec	Hex	Glyph
010 0000	040	32	20	(space)	100 0000	100	64	40	@	110 0000	140	96	60	`
010 0001	041	33	21	!	100 0001	101	65	41	A	110 0001	141	97	61	a
010 0010	042	34	22	"	100 0010	102	66	42	B	110 0010	142	98	62	b
010 0011	043	35	23	#	100 0011	103	67	43	C	110 0011	143	99	63	c
010 0100	044	36	24	$	100 0100	104	68	44	D	110 0100	144	100	64	d
010 0101	045	37	25	%	100 0101	105	69	45	E	110 0101	145	101	65	e
010 0110	046	38	26	&	100 0110	106	70	46	F	110 0110	146	102	66	f
010 0111	047	39	27	'	100 0111	107	71	47	G	110 0111	147	103	67	g
010 1000	050	40	28	(	100 1000	110	72	48	H	110 1000	150	104	68	h

Is this optimal in # of bits? No, if only english words we only need

$2^5 = 32 < 26 \times 2 = 52 < 64 = 2^6$

one letter — 6 bits

Lossless compression

Source X (alphabet) $\xrightarrow[\text{compressor or encoder}]{f}$ $\{0,1\}^*$ $\xrightarrow[\text{Decompressor or decodor}]{g}$ X

① $\{0,1\}^* = \{\phi, 0, 1, 00, 01, 10, 11, 000 \cdots\}$ $\{$ bit string

② for $x \in X$, $f(x) \in \{0,1\}^*$ code word. $\{f(x) \mid x \in X\}$ code book (countable)

③ Lossless compression: $g \circ f = I_X \Rightarrow f$ injective

need $|f(X)| = |X|$ code word

④ length function: $l : \{0,1\}^* \to \mathbb{N}$, e.g. $l(01101) = 5$

an alphabet X need code word with maximal length

$\sup l(f(x)) = \log_2 |X|$

Can we compress more? In terms of maximal codeword length, no
expected codeword length, Yes

Example: $X = \{a, b, c, d\} \xrightarrow{w} \{0,1\}^2$

$a \to 00 \quad b \to 01 \quad c \to 10 \quad d \to 11$

each 2 bits $\qquad \forall x,\ L(w(x)) = 2$ (length)

Now given the fact

$P(a) = 1/2 \quad P(b) = 1/8 \quad P(c) = 1/4, \quad P(d) = 1/8$

Consider $a \to 0, \quad b \to 110, \quad c \to 10, \quad d \to 111$ ← variable length code

$0011010111 \longleftrightarrow a\ a\ b\ c\ d\ a$

Expected codeword length

$$\frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) = \frac{7}{4} < 2 \quad !$$

In the long turn, do better than 2 bits/letter.

What does "$P(a) = \frac{1}{2} \quad P(b) = \frac{1}{8}$ - - - - " mean?

frequency of letters in two different English novel are approximately the same.

English text $\xrightarrow{\text{empirical}}$ frequency of letters $\xrightarrow{\text{modeled by}}$ probabilistic feature

distribution on Alphabet
R.V.

—Shannon

Objective: minimize $\nearrow$ $\sup \ell[f(X)]$
$\searrow$ $\mathbb{E}\ \ell[f(X)]$

There is an optimal compressor f^* minimize both!

Main idea — Assigen short code words for more probable symbols

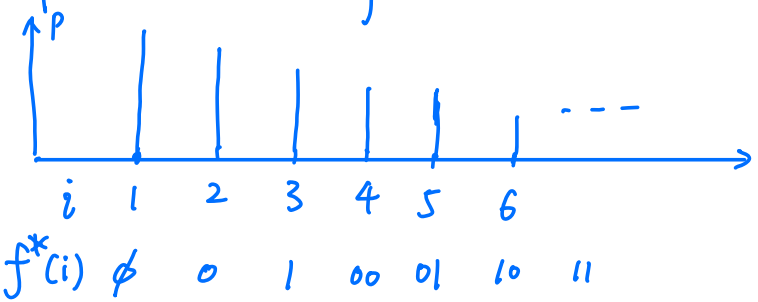
Longer - - - - - - less probable symbols.

WLOG, $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\} \subseteq \mathbb{N}$ and reorder p.m.f $\searrow$

$P_X(i+1) \le P_X(i)$

Theorem (Optimal Compressor)

Define the encoder f^*



Then

1. length of code word

$$\ell(f^*(i)) = \lfloor \log_2 i \rfloor \qquad (\lfloor a \rfloor := \text{largese integer} \le a)$$

2. Stochastically optimal: $\forall$ encoder f,

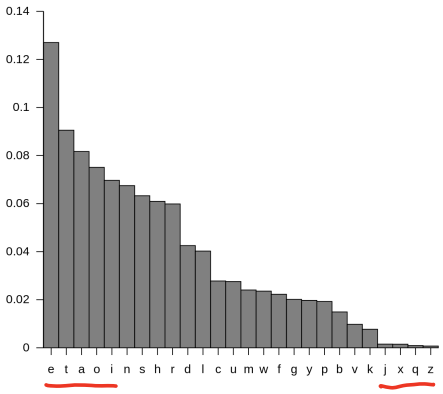
$$\forall k,\ \mathbb{P}[\ell(f(X)) \le k] \le \mathbb{P}[\ell(f^*(X)) \le k] \qquad (\ell f^*(X) \overset{st.}{\le} \ell f(X))$$

As a consequence,

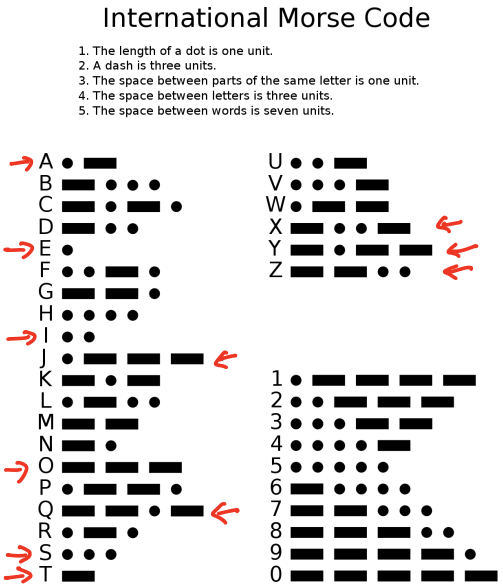
$$\sup \ell f(X) \ge \sup \ell f^*(X).\quad \mathbb{E}\,\ell f(X) \ge \mathbb{E}\,\ell f^*(X)$$

Example (Telegraph & Morse code)

Letter frequency in English text



Morse code



For the other side, we recall the lemma:

Lemma. For $Z \in \mathbb{N}$ and $\mathbb{E}Z < \infty$, $H(Z) \le \mathbb{E}[Z]\, h\left(\frac{1}{\mathbb{E}Z}\right)$

Entropy of Geometric distribution $Q_p(i) = p(1-p)^i$ $\quad H(Q_p) = \frac{h(p)}{p}$

$\mathbb{E}(Q_p) = \frac{1}{p}$

$\frac{1}{p} = \mathbb{E}Z \Rightarrow H(Z) \le H(Q_p)$

$Z \in \mathbb{N}$

Geometric distribution has largese entropy for integer valued RV of given mean $p = \mathbb{E}X$

How to prove? ① Lemma $\forall$ P,Q prob. $D(P\|Q) := \sum p(w) \log \frac{p(w)}{Q(w)} \ge 0$ (Relative entropy)

② Calculate $D(P\|Q_p) = \sum p(w) \log p(w) - p(w) \log Q(w)$

where $p = \frac{1}{\mathbb{E}X}$.

Pf: $|A_k| := |\{x \mid l \circ f(x) \le k\}| \le \sum_{k=0}^{k} 2^i = 2^k - 1$

$= |\{x \mid l^* \circ f(x) \le k\}| = |A_k^*|$

Because $A_k^* = \{1, 2 \cdots, 2^{k-1}\}$

$$P[l \circ f(x) \le k] = \sum_{x \in A_k} P_X(x) \le \sum_{i=1}^{2^{k-1}} P_X(i) = \sum_{x \in A_k^*} P_X(x) = P[l^* \circ f(x) \le k]$$

□

$X, Y : \Omega \to \mathbb{R}$ real R.Vs.

Def. We say X is stochastically dominated by Y, denoted $Y \overset{st.}{\le} X$ if

$\forall k$, $\quad P_X(X \le k) \le P_Y(Y \le k)$

Prop. If $Y \overset{st.}{\le} X$, then

i) $\sup Y \le \sup X$ and ii) $\mathbb{E}Y \le \mathbb{E}X$

Pf: i) $\sup X = \sup\{k \mid P_X(X \le k) < 1\} \ge \sup\{k \mid P_Y(X \le k)\} = \sup Y$

ii) special case: $X, Y : \Omega \to \mathbb{N}$.

$$\boxed{\mathbb{E}X = \sum_{n=1} P_X(X \ge n)}$$

Indeed, $\mathbb{E}X = \sum_{n=1} n\, P_X(n) = P_X(X \ge 1) + \sum_{n=2} P_X(X \ge n)$

$$= \sum_{n=1} P_X(X \ge n)$$

$$\mathbb{E}X = \sum_{n=1} P_X(X \ge n) \ge \sum_{n=1} P_Y(Y \ge n) = \mathbb{E}Y$$

Theorem (Optimal Average code length)
Given $X \in \mathbb{N}$ and $P_X(1) \geq P_X(2) \geq \cdots$. Then

① $\mathbb{E}[l(f^*(X))] = \sum_{k=1}^{\infty} P[X \geq 2^k]$

② $H(X) - \log_2(eH(X)+e) \leq \mathbb{E}[l(f^*(X))] \leq H(X)$

Pf: ① $\mathbb{E}[l(f^*(X))] = \mathbb{E}(\lfloor \log_2 X \rfloor) = \sum_{k\geq 1} P(\lfloor \log_2 X \rfloor \geq k)$

$= \sum_{k\geq 1} P(\log_2 X \geq k)$

② Denote $L(X) = l(f^*(X))$

$P_X(m) \leq \frac{1}{m}$ b/c $P_X(i)$ decreasing

$\Rightarrow l(f^*(m)) = \lfloor \log_2 m \rfloor \leq \log_2 \frac{1}{P_X(m)}$

$\Rightarrow \mathbb{E}(L(X)) \leq \mathbb{E}(\log \frac{1}{P_X(m)}) = H(X)$

For the other side, we recall the lemma:

Lemma. For $Z \in \mathbb{N}$ and $\mathbb{E}Z < \infty$, $H(Z) \leq \mathbb{E}[Z]\, h(\frac{1}{\mathbb{E}Z})$

$\{0,1,2,\cdots,n,\cdots\}$

$H(X) = H(X,L) = H(X|L) + H(L)$

$= \sum P_L(k) H(X|L=k) + h(\frac{1}{1+\mathbb{E}L})(1+\mathbb{E}L)$

$L \in \{0,1,2,\cdots\}$

$L+1 \in \mathbb{N}$

$\leq \sum P_L(k) \log 2^k + \cdots$

$\geq k$

$= \mathbb{E}L + \log(1+\mathbb{E}L) + (\mathbb{E}L)\log(1+\frac{1}{\mathbb{E}L})$

$$\leq \mathbb{E}L + \log_2(e(1+H(X)))$$

$$\left(x\log\left(1+\frac{1}{x}\right) \leq \log e \quad \forall x>0\right)$$

□

Cor. If $X=S^n$ i.i.d. sequence, $H(S^n) = nH(S)$

$$nH(S) - \log n + O(1) \leq \mathbb{E}[\ell^*(S^n)] \leq nH(S)$$

Hence $$\lim_{n\to\infty} \frac{\mathbb{E}[\ell^*(S^n)]}{n} = H(S) \text{ bits}$$

Expected length/message

($\mathbb{E}\ell^*(S^n) = nH(S) - \frac{1}{2}\log n + O(1)$ Szpankowski & Verdú . 2011)

Remark Actually. $\frac{\ell^*(S^n)}{n} \longrightarrow H(S)$ in probability by WLLN.