How much can we compress?
- Shannon's Source Coding Theorem

On Probability and Entropy

Probability

- An ensemble X is a random variable x with a set of possible outcomes $\mathbf{A_x}$ with probabilities $\mathbf{P_x}$
- Probability of a subset T of $\mathbf{A_x}$

$$P(T) = \sum_{a_i \in T} P(x = a_i)$$

- A **joint ensemble** XY is an ensemble for which the outcomes are ordered pairs x,y where $x \in \mathbf{A_x}$ and $y \in \mathbf{A_y}$

Probability continued

- Marginal probability (from the joint probability P(x,y))

$$P(y) = \sum_{x \in A_x} P(x, y)$$

- Conditional probability

$$P(x = a_i \mid y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)}$$

Probability continued

- Product rule

$$P(x, y \mid H) = P(x \mid y, H)P(y \mid H)$$

- Sum rule

$$P(x \mid H) = \sum_{y} P(x, y \mid H)$$

$$= \sum_{y} P(x \mid y, H)P(y \mid H)$$

Bayes's theorem

$$P(y \mid x, H) = \frac{P(x \mid y, H)P(y \mid H)}{P(x \mid H)}$$

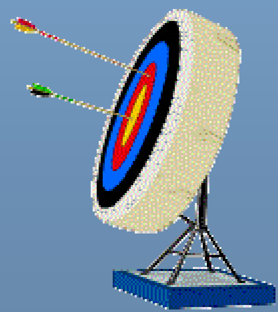
$$= \frac{P(x \mid y, H)P(y \mid H)}{\sum_{y'} P(x \mid y', H)P(y' \mid H)}$$

Bayesian view of probability!

Entropy

- The entropy of X is a measure of the information content or "uncertainty" of x
  - ✓ $H(X) \geq 0$ (= iff p_i=1 for one i)
  - ✓ $H(X) \leq \log(|X|)$ (= iff p_i=1 / |X| for all i)

$$H(X) \equiv \sum_{x \in A_X} P(x) \log \frac{1}{P(x)}$$



Binary entropy

$$H(X) \equiv \sum_i p_i \log_2 \frac{1}{p_i}$$

Information measure?

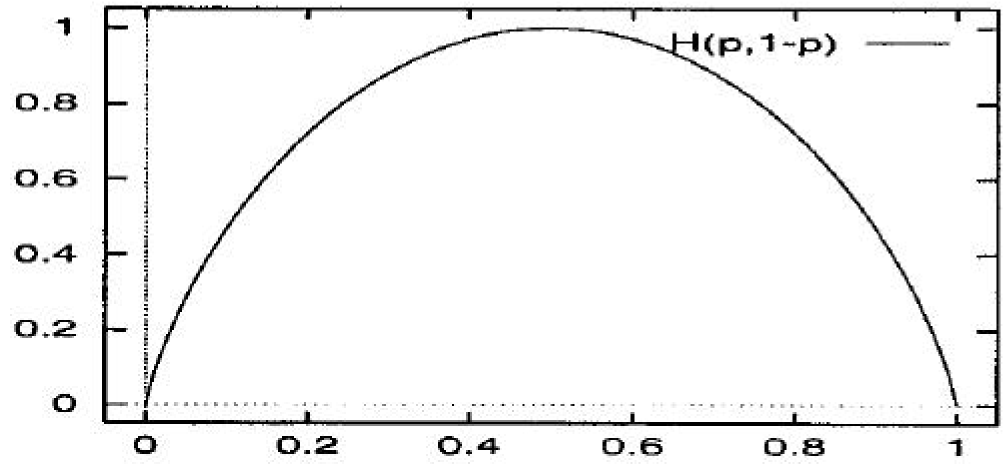
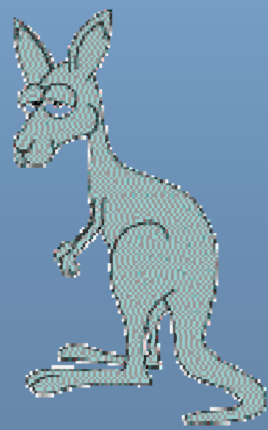


Figure 2.1. The binary entropy function $H_2(p) = H(p, 1-p) = p\log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{(1-p)}$ as a function of p.

Information content

- First attempt: number of possible outcomes $|A_x|$
  - ✓ not additive: for xy we have $|A_x||A_y|$
- Perfect information content
  - ✓ additive, but no probabilistic element

$$H_0(X) = \log_2 |A_X|$$



Shannon information

- looking for an information content of the event $x=a_i$

$$h(x) = \log_2 \frac{1}{p_i}$$

Example: letter distribution

i	a_i	p_i	$\log_2 \frac{1}{p_i}$
1	a	0.06	4.1
2	b	0.01	6.3
3	c	0.03	5.2
4	d	0.03	5.1
5	e	0.09	3.5
6	f	0.02	5.9
7	g	0.01	6.2
8	h	0.03	5.0
9	i	0.06	4.1
10	j	0.00	10.7
11	k	0.01	6.9
12	l	0.04	4.9
13	m	0.02	5.4
14	n	0.06	4.1
15	o	0.07	3.9
16	p	0.02	5.7
17	q	0.01	10.3
18	r	0.05	4.3
19	s	0.06	4.1
20	t	0.07	3.8
21	u	0.03	4.9
22	v	0.01	7.2
23	w	0.01	6.4
24	x	0.01	7.1
25	y	0.02	5.9
26	z	0.00	10.4
27	-	0.19	2.4
	$\sum_i p_i \log_2 \frac{1}{p_i}$		4.11

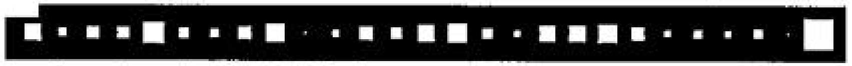


Figure 1.16. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The frequently asked questions manual for Linux*). The picture shows the probabilities by the sizes of white squares.

Entropy continued

- The joint entropy of X,Y

$$H(X,Y) \equiv \sum_{xy \in A_X A_Y} P(x,y) \log \frac{1}{P(x,y)}$$

- The conditional entropy of X given Y

$$H(X \mid Y) \equiv \sum_{y \in A_Y} P(y) \left[\sum_{x \in A_X} P(x \mid y) \log \frac{1}{P(x \mid y)} \right]$$

$$= \sum_{xy \in A_X A_Y} P(x,y) \log \frac{1}{P(x \mid y)}$$

"Average uncertainty that remains about x when y is known"

Entropy continued

- Chain rule for entropy

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

- Mutual information

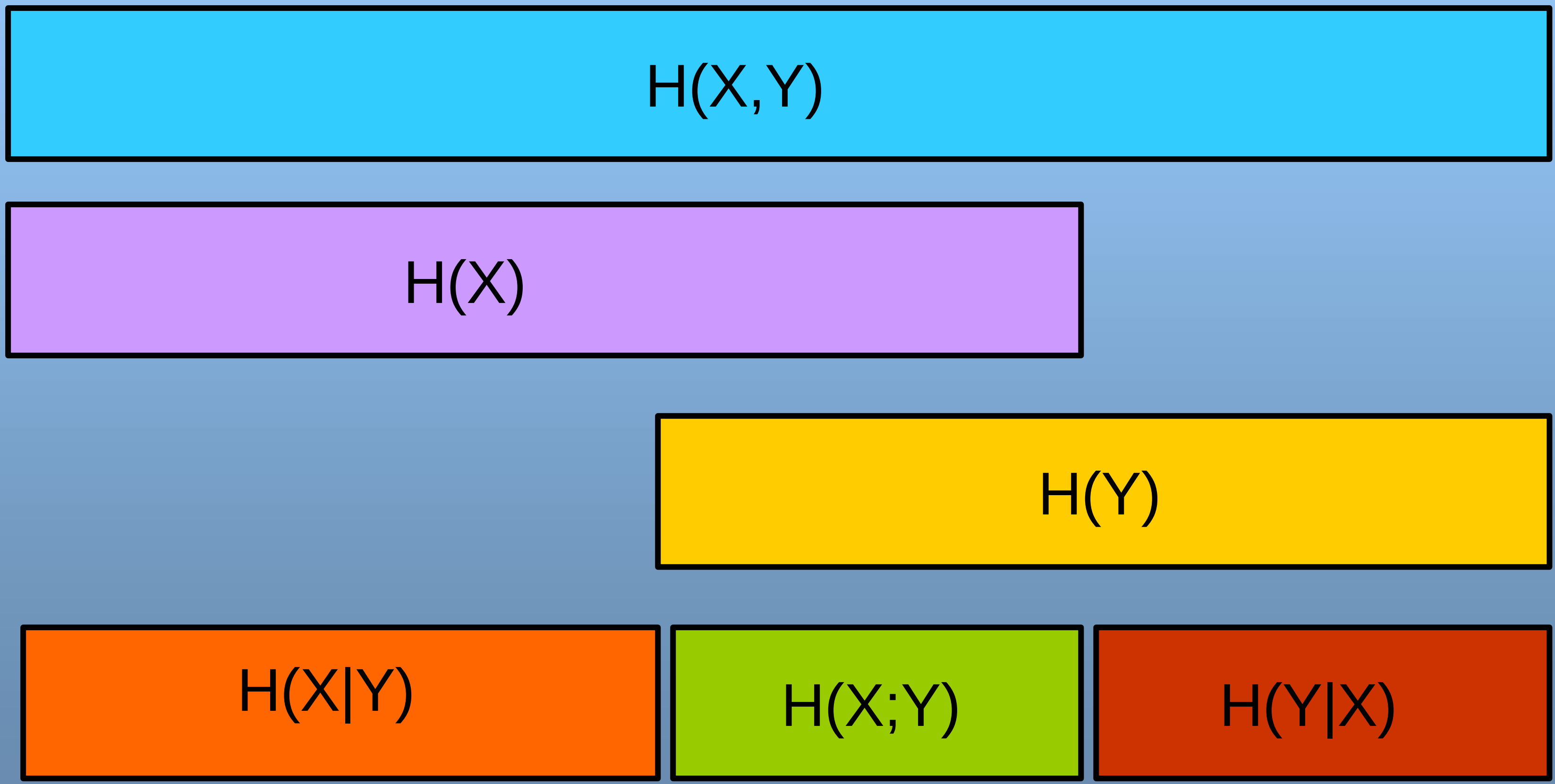
"Average reduction in uncertainty of x when learning the value of y

$$H(X;Y) \equiv H(X) - H(X|Y)$$

- Entropy distance

$$D_H(X,Y) \equiv H(X,Y) - H(X;Y)$$

Entropy relationships



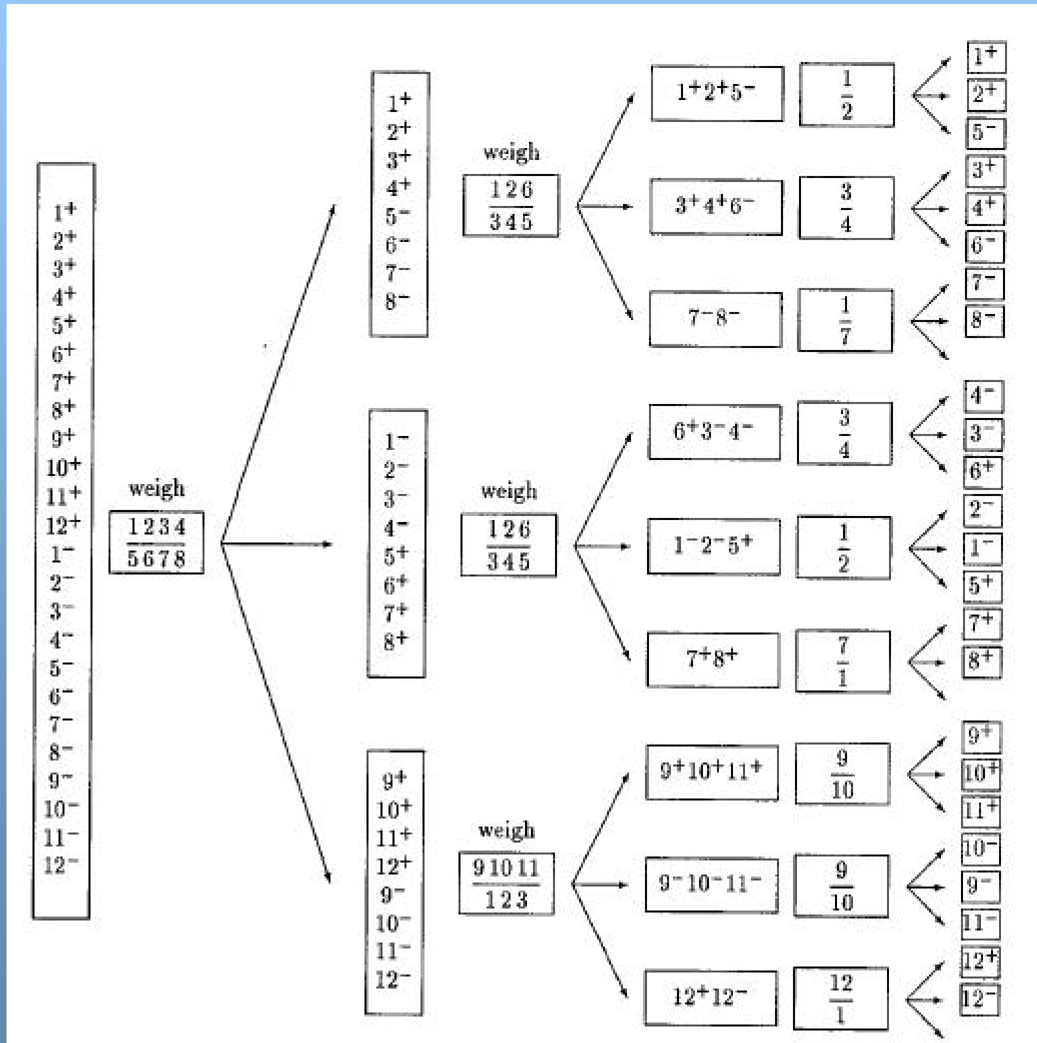
Kullback-Leibler divergence

- Also known as “relative entropy”

$$D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Not strictly a “distance”

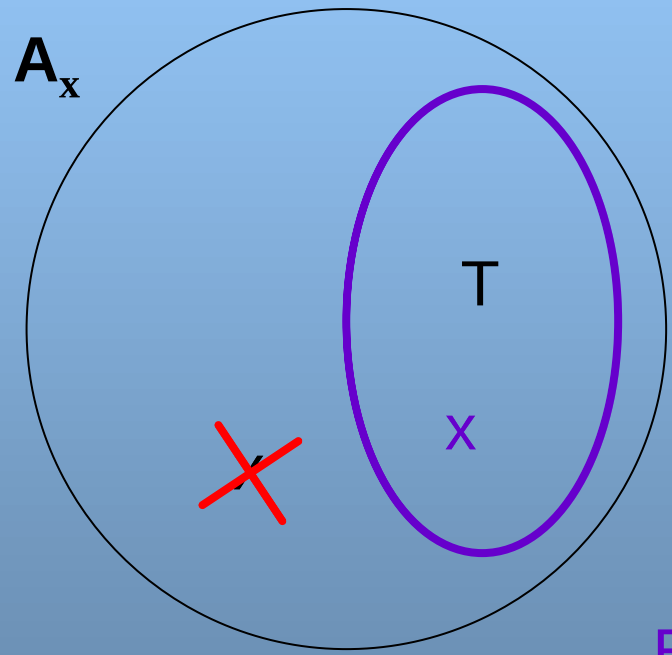
Weighting problem



Idea

- Some symbols have a smaller probability
- gamble that the rare symbols won't occur
- encode the observations in a smaller code (alphabet) C_X
- measure $\log_2|C_X|$
- the larger the risk, the smaller the alphabet

Formalize the idea



Smallest T s.t.

$$P(x \notin T) < d$$

Essential information

$$H_d(X) = \log_2 \min\{|T| : T \subseteq A_X, P(x \in T) \geq 1-d\}$$

Example

$\mathbf{x} = (x_1,\ldots,x_N)$, $x = \{0,1\}$ with probabilities $p_0 = .9$, $p_1 = .1$
Let $r(\mathbf{x})$ be the number of 1's in **x**

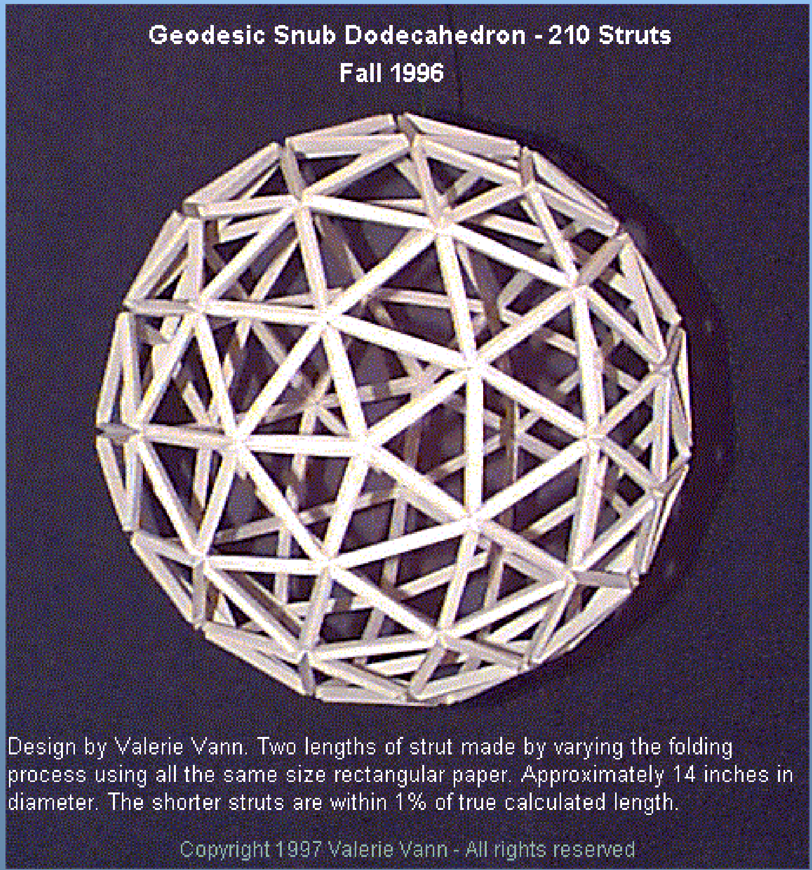
Probability of string **x**

$$P(\mathbf{x} \mid p_0, p_1) = p_0^{N-r(\mathbf{x})} p_1^{r(\mathbf{x})}$$

AEP and source coding

Asymptotic Equipartition Principle: for N i.i.d. random variables $X^N = \{X_1,\ldots,X_N\}$, with N sufficiently large, the outcome $\mathbf{x} = \{x_1,\ldots,x_N\}$ is almost certain to belong to a subset of $\mathbf{A}_{\mathbf{x}}^{\mathbf{N}}$ having only $2^{NH(X)}$ members all having probability close to $2^{-NH(X)}$

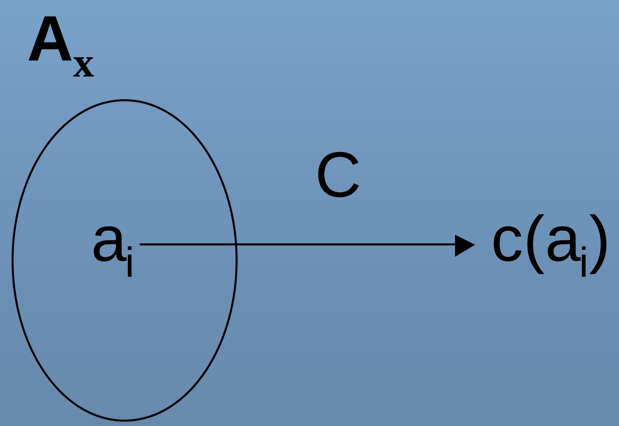
The Revenge of a Student - Symbol Codes



Symbol codes

- Notation: $\{0,1\}^+=\{0,1,00,01,10,11,000,\ldots\}$
- A symbol code C is a mapping from $\mathbf{A_x}$ to $\{0,1\}^+$

$$c^+(x_1x_2x_3\ldots x_N) = c(x_1)c(x_2)c(x_3)\ldots\ c(x_N)$$



$$l(x) = |x|$$

Decoding of symbol codes

- A code C(X) is uniquely decodable if

$$\forall \mathbf{x}, \mathbf{y} \in A_X^+, \mathbf{x} \neq \mathbf{y} \Rightarrow c^+(\mathbf{x}) \neq c^+(\mathbf{y})$$

- A code C(X) is a prefix code if no codeword is a prefix of any other codeword
- The expected length L(C,X) of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in A_x} P(x) l(x)$$

Example

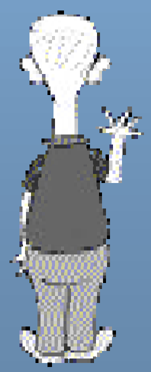
$\mathbf{A_x}$ = {1,2,3,4}, P_X = {1/2,1/4,1/8,1/8}

C: c(1) = 0, c(2) = 10, c(3) = 110, c(4) = 111

The entropy of X is 1.75 bits: L(C,X) is also 1.75 bits

Obs!

$$l_i = \log_2(1/p_i), p_i = 2^{-l_i}$$



Kraft inequality

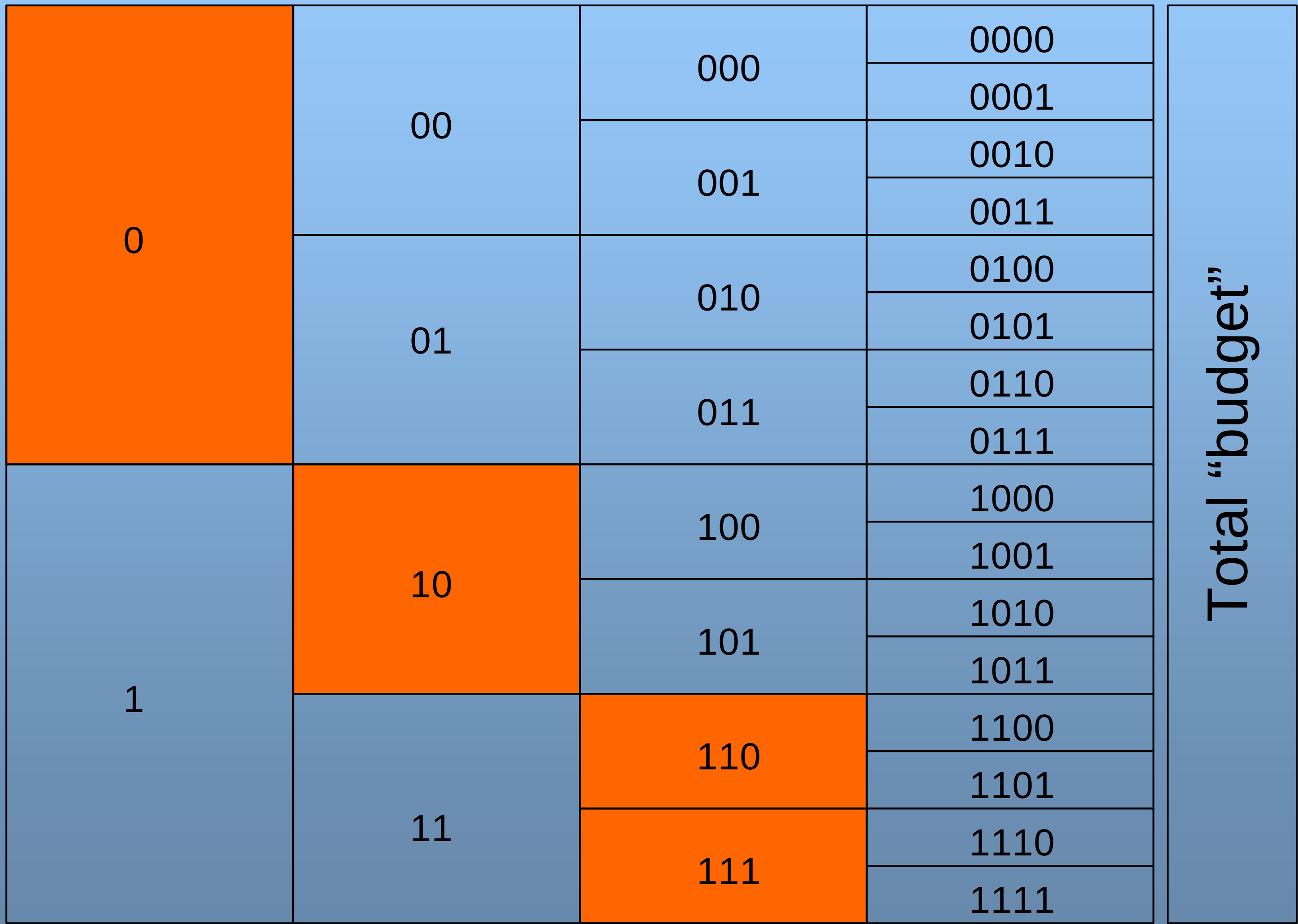
- Given a list of integer $\{l_i\}$, does there exist a uniquely decodable code with $\{l_i\}$?
- "Market model": total budget 1; cost per codeword of length l is 2^{-l}.

Kraft inequality: For any uniquely decodeable code C over the binary alphabet {0,1}, the codeword lengths must satisfy:

$$\sum_i 2^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists a uniquely decodable prefix code with these codelengths.

Limits of unique decodeability



What can we hope for?

Lower bound on expected length: The expected length L(C,X) of a uniquely decodable code is bounded below by H(X).

Compression limit of symbol codes: For an ensemble X there exists a prefix code

$$H(X) \leq L(C,X) < H(X) + 1.$$

“Proof-map” of the lower bound

Define $q_i \equiv 2^{-l_i / z}$, where $z = \sum_{i'} 2^{-l_{i'}}$

By the definition of log

Thus $l_i = \log 1/q_i - \log z$

Substitution

$$L(C,X) = \sum_i p_i l_i = \sum_i p_i \log 1/q_i - \log z$$

Kraft inequality

≥ 0

$$\geq \sum_i p_i \log 1/p_i \boxed{- \log z}$$

Gibbs inequality

$$\geq H(X)$$

(What happens if we use the "wrong" code?)

Assume the "true probability distribution" is $\{p_i\}$. If we use a complete code with lengths l_i, they define a probabilistic model $q_i = 2^{-l_i}$. The average length is

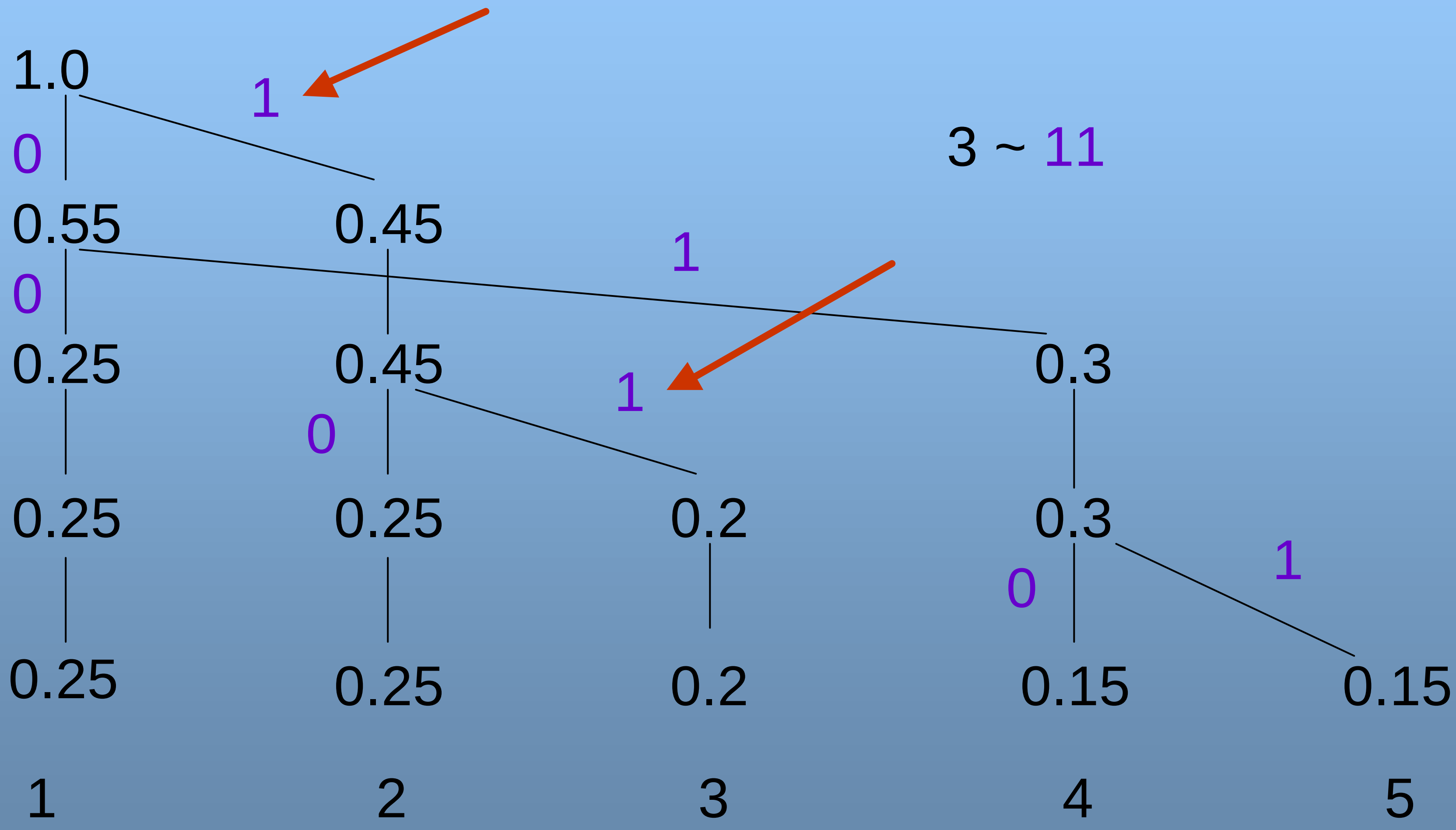
$$L(C,X) = H(X) + \boxed{\sum_i p_i \log p_i / q_i}$$

Kullback-Leibler divergence $D_{KL}(p||q)$

"Optimal" symbol code: Huffman coding

- Take two least probable symbols in the alphabet as defined by $\{p_i\}$.
- Combine these symbols into a single symbol, $p_{new} = p_1 + p_2$. Repeat (until one symbol)

Huffman in practice



Huffman for the Linux manual

L(C,X) = 4.15 bits

H(X) = 4.11 bits

a_i	p_i	$\log_2 \frac{1}{p_i}$	l_i	$c(a_i)$
a	0.0575	4.1	4	0000
b	0.0128	6.3	6	001000
c	0.0263	5.2	5	00101
d	0.0285	5.1	5	10000
e	0.0913	3.5	4	1100
f	0.0173	5.9	6	111000
g	0.0133	6.2	6	001001
h	0.0313	5.0	5	10001
i	0.0599	4.1	4	1001
j	0.0006	10.7	10	1101000000
k	0.0084	6.9	7	1010000
l	0.0335	4.9	5	11101
m	0.0235	5.4	6	110101
n	0.0596	4.1	4	0001
o	0.0689	3.9	4	1011
p	0.0192	5.7	6	111001
q	0.0008	10.3	9	110100001
r	0.0508	4.3	5	11011
s	0.0567	4.1	4	0011
t	0.0706	3.8	4	1111
u	0.0334	4.9	5	10101
v	0.0069	7.2	8	11010001
w	0.0119	6.4	7	1101001
x	0.0073	7.1	7	1010001
y	0.0164	5.9	6	101001
z	0.0007	10.4	10	1101000001
-	0.1928	2.4	2	01

Figure 3.3. Huffman code for the English language ensemble introduced in figure 1.16.

Why is this not the end of the story?

- Adaptation: what if the ensemble X changes? (as it does…)
  - ✓calculate probabilities in one pass
  - ✓communicate code + the Huffman-coded message
- "The extra bit": what if H(X) ~1 bit?
  - ✓Group symbols to blocks and design a "Huffman block code"