DISCRETE MEMORYLESS SOURCE (DMS) Review

- The source output is an unending sequence, $X_1, X_2, X_3, \ldots$, of random letters, each from a finite alphabet $\mathcal{X}$.

- Each source output $X_1, X_2, \ldots$ is selected from $\mathcal{X}$ using a common probability measure with pmf $p_X(x)$.

- Each source output X_k is statistically independent of all other source outputs $X_1, \ldots,$ $X_{k-1}, X_{k+1}, \ldots$.

- Without loss of generality, let $\mathcal{X}$ be $\{1, \ldots, M\}$ and denote $p_X(i)$, $1 \leq i \leq M$ as p_i.

OBJECTIVE: Minimize expected length $\overline{L}$ of prefix codes for a given DMS.

Let $l_1, \ldots, l_M$ be integer codeword lengths.

$$\overline{L}_{\text{min}} = \min_{l_1,\ldots,l_M : \sum 2^{-l_i} \le 1} \left\{ \sum_{i=1}^{M} p_i l_i \right\}$$

Without the integer constraint, $l_i = -\log p_i$ minimizes $\bar{L}_{\text{min}}$, so

$$l_i = -\log p_i \qquad\qquad \text{(desired length)}$$

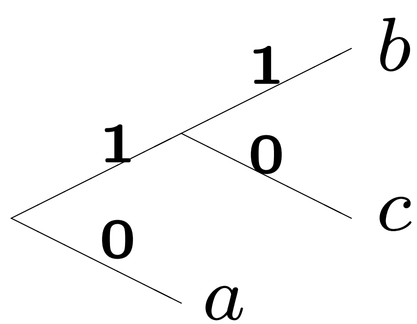
$$\overline{L}_{\text{min}}(\text{non}-\text{int}) = \sum_i -p_i \log p_i \doteq \mathbf{H}(X)$$

$\mathbf{H}(X)$ is the entropy of X. It is the expected value of $-\log p(X)$ and the desired expected length of the binary codeword.

Theorem: Let $\overline{L}_{min}$ be the minimum expected codeword length over all prefix-free codes for X. Then

$$\mathbf{H}(X) \leq \overline{L}_{min} < \mathbf{H}(X) + 1$$

$\overline{L}_{\text{min}} = \mathbf{H}(X)$ iff each p_i is integer power of 2.



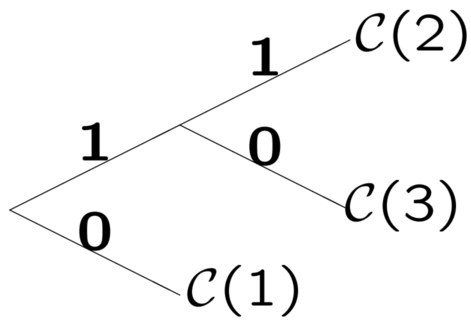
$a \to 0$
$b \to 11$
$c \to 101$

Note that if $p(a) = 1/2$, $p(b) = 1/4$, $p(c) = 1/4$, then each binary digit is IID, 1/2. This is general.

Huffman Coding Algorithm

Above theorem suggested that good codes have $l_i \approx \log(1/p_i)$.

Huffman took a different approach and looked at the tree for a prefix-free code.



$$p_1 = 0.6$$
$$p_2 = 0.3$$
$$p_3 = 0.1$$

Lemma: Optimal prefix-free codes have the property that if $p_i > p_j$ then $l_i \leq l_j$. This means that $p_i > p_j$ and $l_i > l_j$ can't be optimal.

Lemma: Optimal prefix-free codes are full.

The sibling of a codeword is the string formed by changing the last bit of the codeword.

Lemma: For optimality, the sibling of each maximal length codeword is another codeword.

Assume that $p_1 \geq p_2 \geq \cdots \geq p_M$.

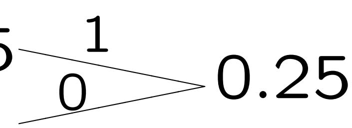
Lemma: There is an optimal prefix-free code in which $\mathcal{C}(M-1)$ and $\mathcal{C}(M)$ are maximal length siblings.

Essentially, the codewords for $M-1$ and M can be interchanged with max length codewords.

The Huffman algorithm first combines $\mathcal{C}(M-1)$ and $\mathcal{C}(M)$ and looks at the reduced tree with $M-1$ nodes.

After combining two least likely codewords as sibliings, we get a "reduced set" of probabilities.

symbol	p_i
1	0.4
2	0.2
3	0.15
4	0.15
5	0.1



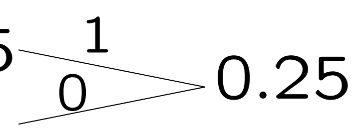
Finding the optimal code for the reduced set results in an optimal code for original set. Why?

Finding the optimal code for the reduced set results in an optimal code for original set. Why?

For any code for the reduced set X', let expected length be $\overline{L}'$.

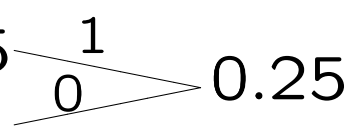
The expected length of the corresponding code for X has $\overline{L} = \overline{L}' + p_{M-1} + p_M$.

symbol	p_i
1	0.4
2	0.2
3	0.15
4	0.15
5	0.1

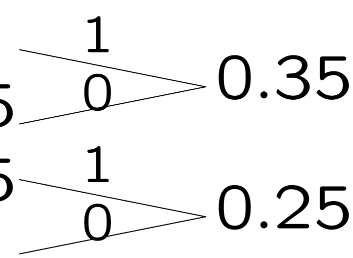


Now we can tie together (siblingify?) the least two probable nodes in the reduced set.

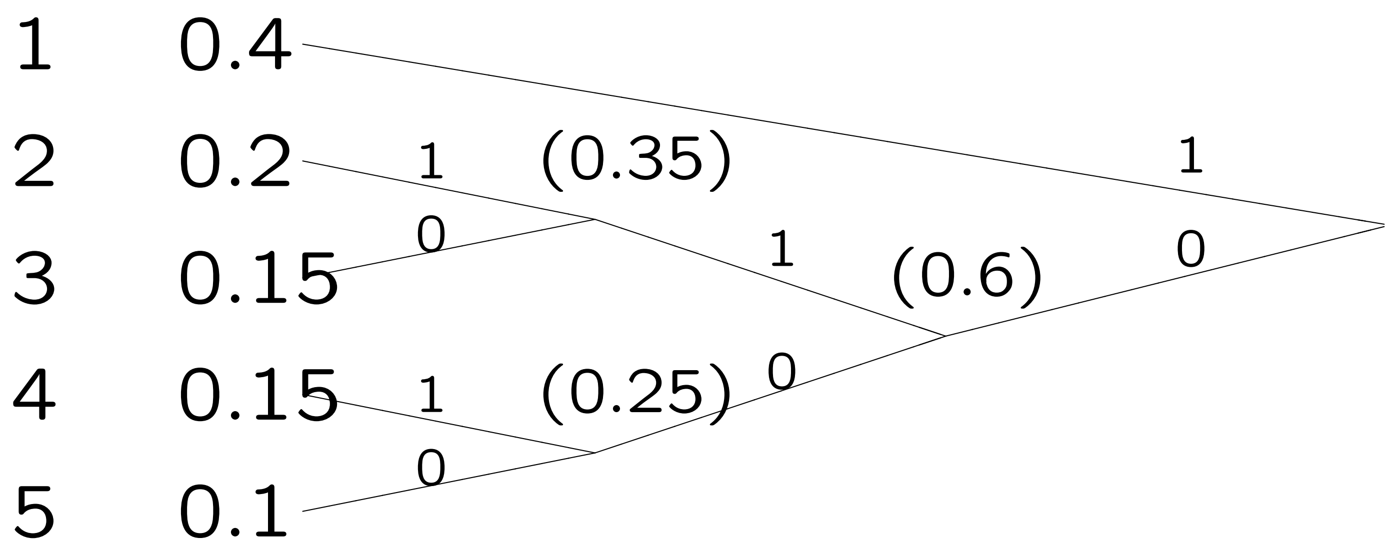
symbol	p_i
1	0.4
2	0.2
3	0.15
4	0.15
5	0.1



1	0.4
2	0.2
3	0.15
4	0.15
5	0.1



Surely the rest is obvious.



DISCRETE SOURCE CODING: REVIEW

The Kraft inequality, $\sum_i 2^{-l_i} \le 1$, is a necessary and sufficient condition on prefix-free codeword lengths.

Given a pmf, $p_1, \dots, p_M$ on a set of symbols, the Huffman algorithm constructs a prefix-free code of minimum expected length, $\overline{L}_{\min} = \sum_i p_i l_i$.

A discrete memoryless source (DMS) is a sequence of iid discrete chance variables $X_1, X_2, \dots$. The entropy of a DMS is $\mathbf{H}(X) = \sum_i -p_i \log(p_i)$.

Theorem: $\mathbf{H}(X) \le \overline{L}_{\min} < \mathbf{H}(X) + 1$.

ENTROPY OF X, $|\mathcal{X}| = M$, $\Pr(X{=}i) = p_i$

$$\mathsf{H}(X) = \sum_i -p_i \log p_i = \mathsf{E}[-\log p_X(X)]$$

$-\log p_X(X)$ **is a rv, called the log pmf.**

$\mathsf{H}(X) \geq 0$**; Equality if** X **deterministic.**

$\mathsf{H}(X) \leq \log M$**; Equality if** X **equiprobable.**

For independent rv's X, Y, XY **is also a chance variable taking on the sample value** xy **with probability** $p_{XY}(xy) = p_X(x)p_Y(y)$.

$$\begin{aligned}\mathsf{H}(XY) &= \mathsf{E}[-\log p(XY)] = \mathsf{E}[-\log p(X)p(Y)] \\ &= \mathsf{E}[-\log p(X) - \log p(Y)] = \mathsf{H}(X) + \mathsf{H}(Y)\end{aligned}$$

For a discrete memoryless source, a block of n random symbols, $X_1, \ldots, X_n$, can be viewed as a single random symbol X^n taking on the sample value $\mathrm{x}^n = x_1 x_2 \ldots x_n$ with probability

$$p_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} p_X(x_i)$$

The random symbol $\mathrm{X^n}$ has the entropy

$$\begin{aligned} \mathsf{H}(\mathbf{X^n}) &= \mathsf{E}[-\log p(\mathbf{X^n})] = \mathsf{E}[-\log \prod_{i=1}^{n} p_X(X_i)] \\ &= \mathsf{E}\left[\sum_{i=1}^{n} -\log p_X(X_i)\right] = n\mathsf{H}(X) \end{aligned}$$

Fixed-to-variable prefix-free codes

Segment input into n-blocks $\mathbf{X}^n = X_1 X_2 \ldots X_n$.

Form min-length prefix-free code for $\mathbf{X^n}$.

This is called an n-to-variable-length code

$$\mathbf{H}(\mathbf{X^n}) = \mathbf{n}\mathbf{H}(\mathbf{X})$$

$$\mathbf{H}(\mathbf{X^n}) \leq \mathbf{E}[\mathbf{L}(\mathbf{X^n})]_{\mathsf{min}} < \mathbf{H}(\mathbf{X^n}) + 1$$

$$\overline{L}_{\mathsf{min},n} = \frac{\mathbf{E}[L(X^n)]_{\mathsf{min}}}{n} \qquad \text{bpss}$$

$$\mathbf{H}(X) \leq \overline{L}_{\mathsf{min},n} < \mathbf{H}(X) + 1/n$$

$$\bar{L}_{\mathsf{min},n} \to \mathbf{H}(X)$$

WEAK LAW OF LARGE NUMBERS (WLLN)

Let $Y_1, Y_2, \ldots$ **be sequence of rv's with mean** $\overline{Y}$ **and variance** σ_Y^2**.**

The sum $A = Y_1 + \cdots + Y_n$ **has mean** $n\overline{Y}$ **and variance** $n\sigma_Y^2$

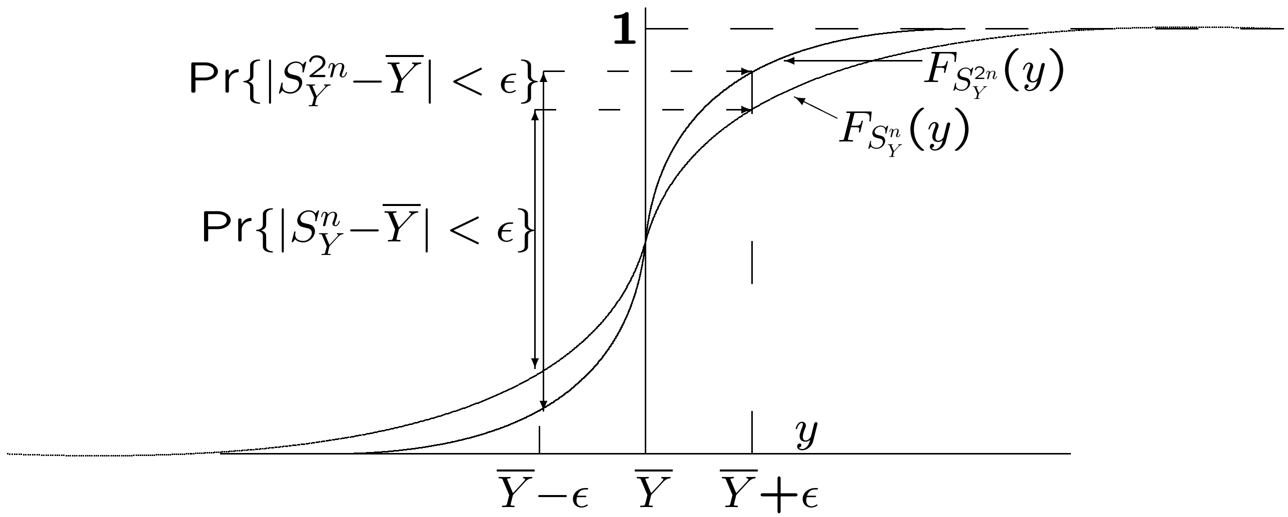
The sample average of $Y_1, \ldots, Y_n$ **is**

$$S_Y^n = \frac{A}{n} = \frac{Y_1 + \cdots + Y_n}{n}$$

It has mean and variance

$$\mathsf{E}[S_y^n] = \overline{Y}; \qquad \mathsf{VAR}[S_Y^n] = \frac{\sigma_Y^n}{n}$$

Note: $\lim_{n\to\infty} \mathsf{VAR}[A] = \infty \qquad \lim_{n\to\infty} \mathsf{VAR}[S_Y^n] = 0$**.**



The distribution of S_Y^n clusters around $\overline{Y}$, clustering more closely as $n \to \infty$.

Chebyshev: for $\epsilon > 0$, $\Pr\{|S_Y^n - \overline{Y}| \geq \epsilon\} \leq \frac{\sigma_Y^2}{n\epsilon^2}$

For any $\epsilon, \delta > 0$, **large enough** n,

$$\Pr\{|S_Y^n - \overline{Y}| \geq \epsilon\} \leq \delta$$

ASYMPTOTIC EQUIPARTITION PROPERTY (AEP)

Let $X_1, X_2, \ldots ,$ **be output from DMS.**

Define log pmf as $w(x) = -\log p_X(x)$**.**

$w(x)$ **maps source symbols into real numbers.**

For each j**,** $W(X_j)$ **is a rv; takes value** $w(x)$ **for** $X_j = x$**. Note that**

$$\mathsf{E}[W(X_j)] = \sum_x p_X(x)[-\log p_X(x)] = H(X)$$

$W(X_1), W(X_2), \ldots$ **sequence of iid rv's.**

For $X_1 = x_1, X_2 = x_2$, the outcome for $W(X_1) + W(X_2)$ is

$$\begin{aligned} w(x_1) + w(x_2) &= -\log p_X(x_1) - \log p_X(x_2) \\ &= -\log\{p_{X_1}(x_1) p_{X_2}(x_2)\} \\ &= -\log\{p_{X_1 X_2}(x_1 x_2)\} = w(x_1 x_2) \end{aligned}$$

where $w(x_1 x_2)$ is -log pmf of event $X_1 X_2 = x_1 x_2$

$$W(X_1 X_2) = W(X_1) + W(X_2)$$

$X_1 X_2$ is a random symbol in its own right (takes values $x_1 x_2$). $W(X_1 X_2)$ is -log pmf of random symbol $X_1 X_2$.

Probabilities multiply, log pmf's add.

For $\mathbf{X}^n = \mathbf{x}^n$; $\mathbf{x}^n = (x_1, \ldots, x_n)$, the outcome for $W(X_1) + \cdots + W(X_n)$ is

$$\sum_{j=1}^{n} w(x_j) = -\sum_{j=1}^{n} \log p_X(x_j) = -\log p_{\mathbf{X}^n}(\mathbf{x}^n)$$

Sample average of log pmf's is

$$S_W^n = \frac{W(X_1) + \cdots W(X_n)}{n} = \frac{-\log p_{\mathbf{X}^n}(\mathbf{X}^n)}{n}$$

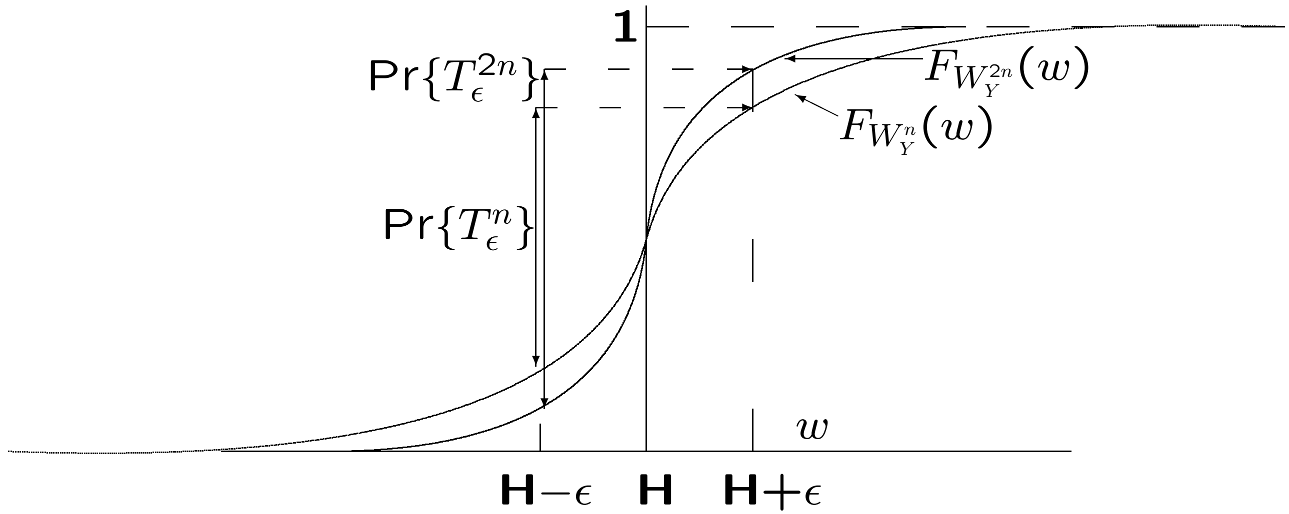
WLLN applies and is

$$\Pr\left(\left| S_W^n - \mathsf{E}[W(X)] \right| \geq \epsilon \right) \leq \frac{\sigma_W^2}{n\epsilon^2}$$

$$\Pr\left(\left| \frac{-\log p_{\mathbf{X}^n}(\mathbf{X}^n)}{n} - \mathsf{H}(X) \right| \geq \epsilon \right) \leq \frac{\sigma_W^2}{n\epsilon^2}.$$

Define typical set as

$$T_\epsilon^n = \left\{ \mathbf{x}^n : \left| \frac{-\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} - \mathbf{H}(X) \right| < \epsilon \right\}$$



As $n \to \infty$, typical set approaches probability 1:

$$\Pr(\mathbf{X}^n \in T_\epsilon^n) \geq 1 - \frac{\sigma_W^2}{n\epsilon^2}$$

We can also express T_ϵ^n as

$$T_\epsilon^n = \left\{ \mathbf{x}^n : n(\mathsf{H}(X) - \epsilon) < -\log p(\mathbf{x}^n) < n(\mathsf{H}(X) + \epsilon) \right\}$$

$$T_\epsilon^n = \left\{ \mathbf{x}^n : \ 2^{-n(\mathsf{H}(X)+\epsilon)} < p_{\mathbf{X}^n}(\mathbf{x}^n) < 2^{-n(\mathsf{H}(X)-\epsilon)} \right\}.$$

Typical elements are approximately equiprobable in the strange sense above.

The complementary, atypical set of strings, satisfy

$$\Pr[(T_\epsilon^n)^c] \le \frac{\sigma_W^2}{n\epsilon^2}$$

For any $\epsilon, \delta > 0$, large enough n, $\Pr[(T_\epsilon^n)^c] < \delta$.

For all $\mathbf{X^n} \in \mathbf{T^n_\epsilon}$**,** $p_{\mathbf{X}^n}(\mathbf{X^n}) > \mathbf{2^{-n[H(X)+\epsilon]}}$**.**

$$1 \geq \sum_{\mathbf{X}^n \in T^n_\epsilon} p_{\mathbf{X}^n}(\mathbf{X^n}) > |\mathbf{T^n_\epsilon}|\, \mathbf{2^{-n[H(X)+\epsilon]}}$$

$$|T^n_\epsilon| < 2^{n[\mathbf{H}(X)+\epsilon]}$$

$$1-\delta \leq \sum_{\mathbf{X}^n \in T^n_\epsilon} p_{\mathbf{X^n}}(\mathbf{X^n}) < |\mathbf{T^n_\epsilon}|\mathbf{2^{-n[H(X)-\epsilon]}}$$

$$|T^n_\epsilon| > (1-\delta)2^{n[\mathbf{H}(X)-\epsilon]}$$

Summary: $\Pr[(T^n_\epsilon)^c] \approx 0, \quad |T^n_\epsilon| \approx 2^{n\mathbf{H}(X)},$

$$p_{\mathbf{X^n}}(\mathbf{X^n}) \approx \mathbf{2^{-nH(X)}} \quad \textbf{for } \mathbf{X^n} \in \mathbf{T^n_\epsilon}.$$

EXAMPLE

Consider binary DMS with $\Pr[X{=}1] = p < 1/2$**.**

$$\mathsf{H}(X) = -p\log p - (1{-}p)\log(1{-}p)$$

The typical set T_ϵ^n **is the set of strings with about** pn **ones and** $(1{-}p)n$ **zeros.**

The probability of a typical string is about $p^{pn}(1{-}p)^{(1-p)n} = 2^{-n\mathsf{H}(X)}$**.**

The number of n**-strings with** pn **ones is** $\frac{n!}{(pn)!(n-pn)!}$

Note that there are 2^n **binary strings. Most of them are collectively very improbable.**

The most probable strings have almost all zeros, but there aren't enough of them to matter.

Fixed-to-fixed-length source codes

For any $\epsilon, \delta > 0$, and any large enough n, assign fixed length code word to each $\mathrm{X^n} \in \mathrm{T}_\epsilon$.

Since $|T_\epsilon| < 2^{n[\mathbf{H}(X)+\epsilon]}$, $\overline{L} \leq \mathbf{H}(X)+\epsilon+1/n$.

$$\Pr\{\text{failure}\} \leq \delta.$$

Conversely, take $\overline{L} \leq \mathbf{H}(X) - 2\epsilon$, and n large.

Since $|T_\epsilon^n| > (1-\delta)2^{n[\mathbf{H}(X)-\epsilon]}$, most of typical set can not be assigned codewords.

$$\Pr\{\text{failure}\} > 1 - \delta - 2^{-\epsilon\epsilon n} \rightarrow 1$$

Kraft inequality for unique decodability

Suppose $\{l_i\}$ are lengths of a uniquely decodable code and $\sum_i 2^{-l_i} = b$. We show that $b > 1$ leads to contradiction. Choose DMS with $p_i = (1/b)2^{-l_i}$, i.e., $l_i = -\log(bp_i)$.

$$\overline{L} = \sum_i p_i l_i = \mathbf{H}(X) - \log b$$

Consider string of n source letters. Concatenation of code words has length less than $n[\mathbf{H}(X) - b/2]$ with high probability. Thus fixed length code of this length has low failure probability.

Contradiction.

6.450 Principles of Digital Communication I
Fall 2009