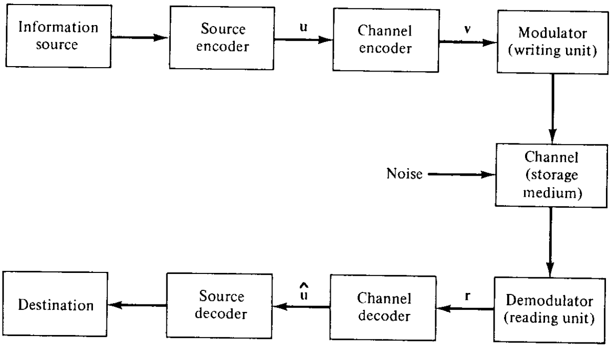
Playing with Signals:
Coding for Compression, Caching, Computing . . .

Nikhil Karamchandani
(filling in for Sibi)

Email: nikhilk@ee.iitb.ac.in

Communication / Storage System



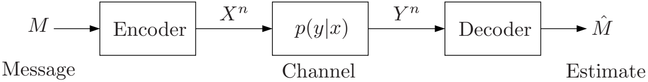
Coding: how, what, why?

- Source coding

$$|M| \approx 2^{nR}, \quad R^\star = H(X)$$

Compression

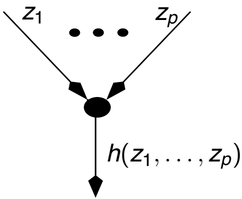
- Channel coding

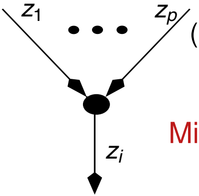


$$|M| \approx 2^{nR}, \quad R^\star = \max_{p(X)} \; I(X;Y)$$

Reliability

- Network coding





(Routing)

Mixing

Source Compression

Source Coding

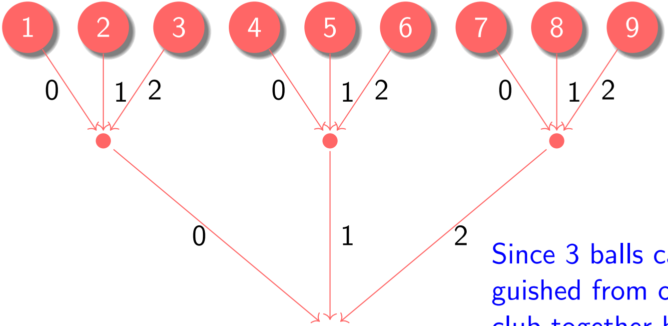
- *Source*: A random variable X, taking values in $\mathcal{X}$ according to probability distribution P_X.
- *Code*: Mapping $\mathcal{C}$ from $\mathcal{X}$ to strings over alphabet $\mathcal{D}$.
  - For $x \in \mathcal{X}$, let $C(x)$ denote the codeword and $l(x)$ denote the length.
- *Metric*: Expected code length $L_{\mathcal{C}}$, given by

$$L_{\mathcal{C}}(X) = \sum_{x \in \mathcal{X}} p(x) l(x).$$

- *Goal*: Find minimum expected code length $L^{\star}(X)$ for source X.

9 Ball Game

Suppose there are 9 balls, look alike, but one of them is heavier than the rest (GOLD!). With two weighings (measurements) on a common balance, can you identify the odd one.

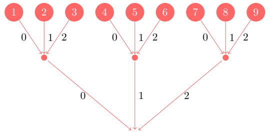


Since 3 balls can be distinguished from one test, we club together blocks of 3

Balls and Sources

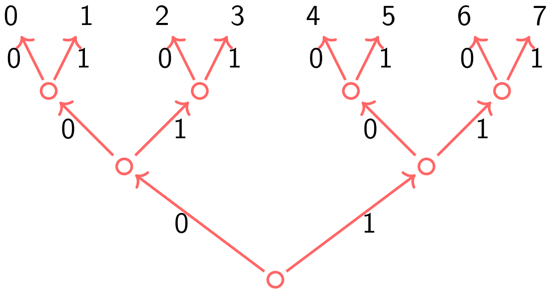
- Suppose we repeatedly perform the first experiment, using a statistical machine that shuffles the golden ball
- The random variable representing the output of the machine is a *source*.
- Every time a source symbol $S_i \in \mathcal{X}$ occurs, we will convey its branch labels.

Question: For a given source and a label-alphabet,
what is the **optimal** tree?



Binary Number System

- A binary tree representation for numbers.



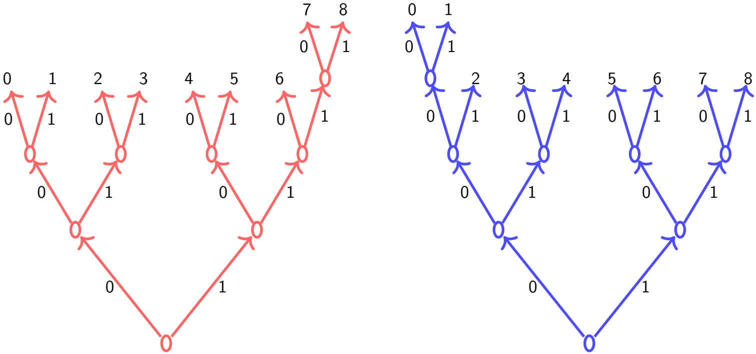
- If the source is fair (uniform distribution), this indeed is the **optimal** tree.

- This also gives the simple bound

$$L^\star(X) \leq \log|\mathcal{X}| + 1$$

Gimme another Drop

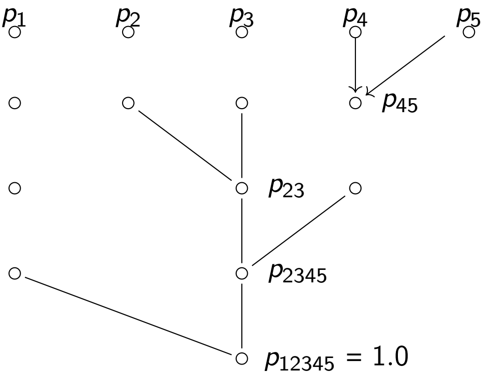
- Suppose $\mathcal{X} = \{0, 1, \cdots, 8\}$ with $p_0 \geq p_1 \geq \cdots \geq p_8$.



- '*Shorter codes to more frequent symbols*' seems to be the key to compression, we seek the best code tree for a given probability distribution on the symbols.

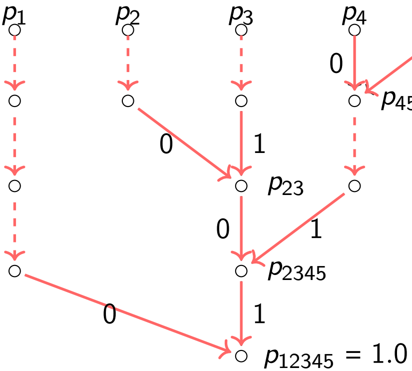
Binary Huffman Example

Let $p_1 = 0.47, p_2 = 0.18, p_3 = 0.15, p_4 = 0.1, p_5 = 0.1$ and $p_{ij} \triangleq p_i + p_j$.



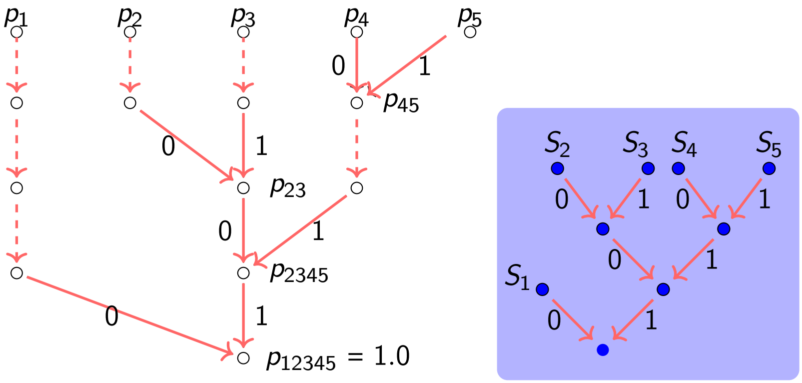
Binary Huffman Example

Let $p_1 = 0.47, p_2 = 0.18, p_3 = 0.15, p_4 = 0.1, p_5 = 0.1$ and $p_{ij} \triangleq p_i + p_j$.



Binary Huffman Example

Let $p_1 = 0.47, p_2 = 0.18, p_3 = 0.15, p_4 = 0.1, p_5 = 0.1$ and $p_{ij} \triangleq p_i + p_j$.



Collapse (delete) the dashed lines to get the highlighted tree.

Ternary Huffman Code

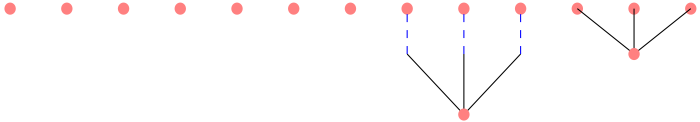
Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$

Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$

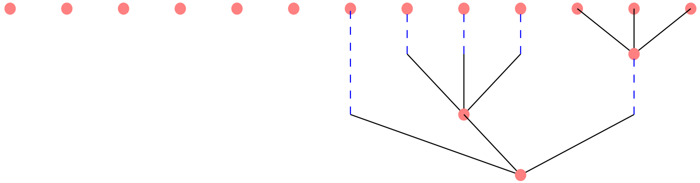
Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$



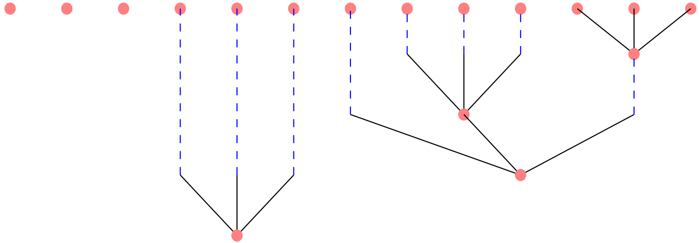
Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$



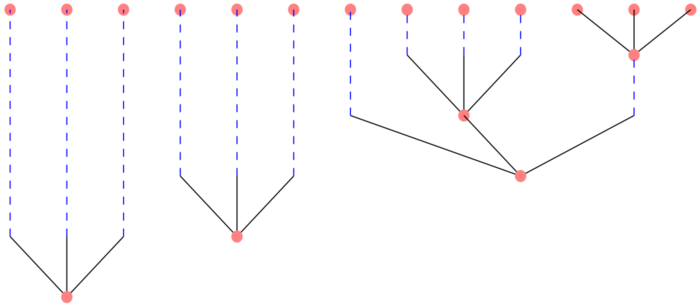
Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$



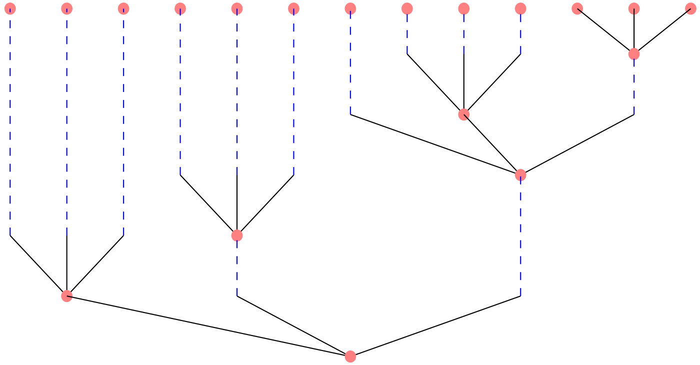
Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$



Ternary Huffman Code

Source	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}
Pr(S)	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$



Huffman Coding

- We will describe **Huffman Coding** when $\mathcal{D} = \{0, 1\}$ (binary).
- Let all labels be empty, and let $m = |\mathcal{X}|$.
  1. Rearrange sources such that $p_1 \geq p_2 \geq \cdots \geq p_m$.
  2. Append labels 0 and 1 respectively to the last two sources.
  3. Merge the last two sources to form a new source X'_{m-1}, having probability $p_{m-1} + p_m$.
  4. Put $m \leftarrow m - 1$ and go to step 1, using the new source set.
- Terminate by assiging 0 and 1 to the two remaining sources.
- Huffman coding is optimal, i.e., it achieves the minimum average codeword length $L^\star(X)$.

Information Entropy

- For $X \sim p_X$, entropy $H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$.
- A quantifiable measure of uncertainty / information content.
- $H(X) \leq \log |\mathcal{X}|$; maximum attained for uniform distribution.
- *Operational meaning*: $H(X) \leq L^{\star}(X) \leq H(X) + 1$.

Information Entropy

- For $X \sim p_X$, entropy $H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$.
- A quantifiable measure of uncertainty / information content.
- $H(X) \leq \log |\mathcal{X}|$; maximum attained for uniform distribution.
- *Operational meaning*: $H(X) \leq L^\star(X) \leq H(X) + 1$.
- Upper bound attained by choosing code with $l(x) = \lceil \log \frac{1}{p(x)} \rceil$.

Claude Shannon: Father of Information Theory

Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons: