An introduction to Information Theory

Adrish Banerjee

Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh
India

Aug. 1, 2016

Lecture #5: Variable to block length coding

Outline of the lecture

- Variable to block length coding

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Outline of the lecture

- Variable to block length coding
- Proper message set

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Outline of the lecture

- Variable to block length coding
- Proper message set
- Assigning probabilities to a K-ary rooted tree corresponding to a proper message set.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Outline of the lecture

- Variable to block length coding
- Proper message set
- Assigning probabilities to a K-ary rooted tree corresponding to a proper message set.
- Prefix free coding of a proper message set

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Outline of the lecture

- Variable to block length coding
- Proper message set
- Assigning probabilities to a K-ary rooted tree corresponding to a proper message set.
- Prefix free coding of a proper message set
- Tunstall message set

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Outline of the lecture

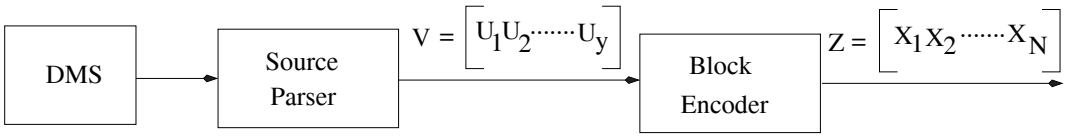
- Variable to block length coding
- Proper message set
- Assigning probabilities to a K-ary rooted tree corresponding to a proper message set.
- Prefix free coding of a proper message set
- Tunstall message set
- Tunstall coding

Outline of the lecture

- Variable to block length coding
- Proper message set
- Assigning probabilities to a K-ary rooted tree corresponding to a proper message set.
- Prefix free coding of a proper message set
- Tunstall message set
- Tunstall coding
- Variable-Length-to-Block Coding Theorem for a DMS

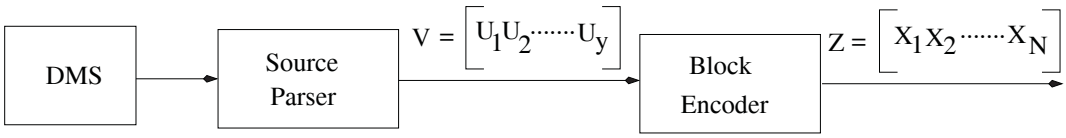
Outline

Variable to block length coding



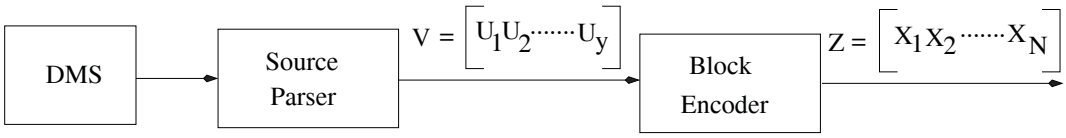
- Source parser does variable length parsing.

Variable to block length coding



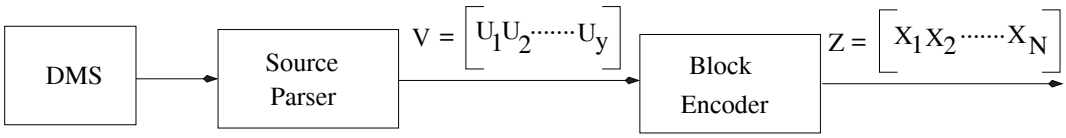
- Source parser does variable length parsing.
- Length Y of the message V is a random variable.

Variable to block length coding



- Source parser does variable length parsing.
- Length Y of the message V is a random variable.
- Average number of D-ary codewords per source letter is given by N/E[Y].

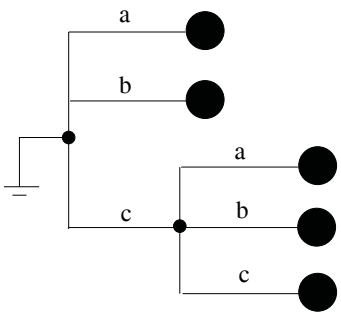
Variable to block length coding



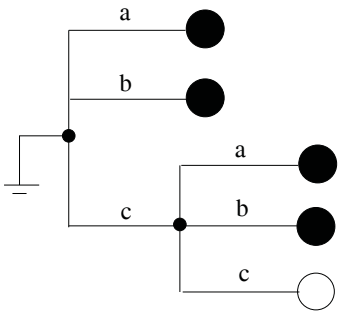
- Source parser does variable length parsing.
- Length Y of the message V is a random variable.
- Average number of D-ary codewords per source letter is given by N/E[Y].
- The average message length E[Y] should be as large as possible.

Outline

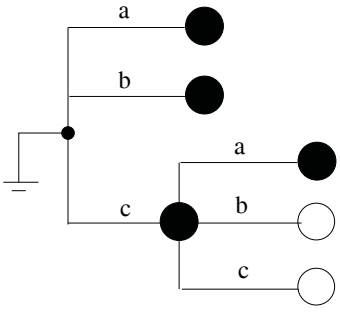
Proper message set



A proper message set



An improper message set

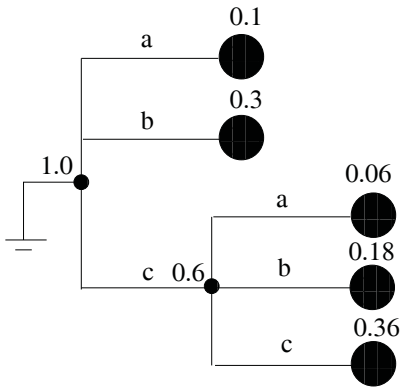


An improper message set

- A proper message set for a K-ary source is a set of messages that form a complete set of leaves for a rooted K-ary tree.

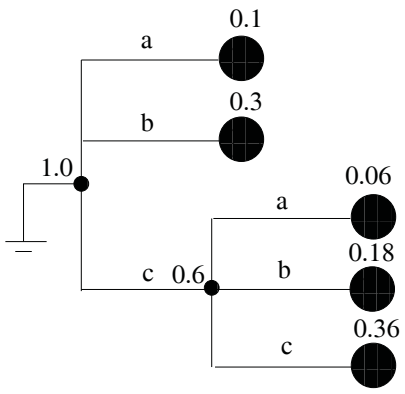
Outline

Assigning probabilities to a K-ary rooted tree corresponding to a proper message set



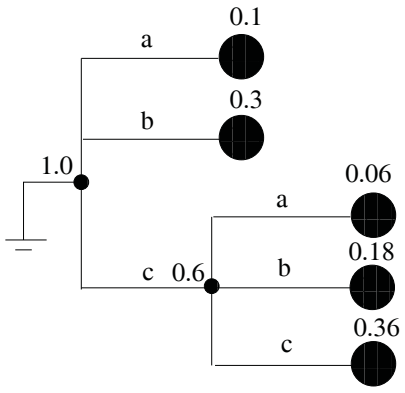
- The root is assigned a probability 1.

Assigning probabilities to a K-ary rooted tree corresponding to a proper message set



- The root is assigned a probability 1.
- Each subsequent vertex is assigned a probability equal to the probability of the node from which it stems multiplied by the probability of the letter emitted by DMS on the branch connecting that node to the vertex.

Assigning probabilities to a K-ary rooted tree corresponding to a proper message set



- The root is assigned a probability 1.
- Each subsequent vertex is assigned a probability equal to the probability of the node from which it stems multiplied by the probability of the letter emitted by DMS on the branch connecting that node to the vertex.
- Leaf entropy, $H_{\text{leaf}} = H(V)$

Proper message set

- The uncertainty, $H(V)$, of a proper message set for a K-ary DMS with output uncertainty $H(U)$ satisfies

$$H(V) = E[Y]H(U)$$

where E[Y] is the average message length.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Proper message set

- The uncertainty, $H(V)$, of a proper message set for a K-ary DMS with output uncertainty $H(U)$ satisfies

$$H(V) = E[Y]H(U)$$

where E[Y] is the average message length.

- All branching entropy $H_i = H(U)$.

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Proper message set

- The uncertainty, $H(V)$, of a proper message set for a K-ary DMS with output uncertainty $H(U)$ satisfies

$$H(V) = E[Y]H(U)$$

  where E[Y] is the average message length.
- All branching entropy $H_i = H(U)$.
- Leaf entropy is given by

$$\begin{aligned} H_{\text{leaf}} &= \sum_i P_i H_i && \text{where } P_i \text{ is the } i\text{-th node probability.} \\ \implies H(V) &= H(U)\sum_i P_i \\ H(V) &= H(U)E[Y] && \text{From path length lemma.} \end{aligned}$$

Outline

Prefix free coding of a proper message set

- For any D-ary prefix-free encoding of any proper message set for a DMS, the ratio of the average codeword length, E[W], to the average message length E[Y], satisfies

$$\frac{E[W]}{E[Y]} \geq \frac{H(U)}{\log D}$$

Proof:

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Prefix free coding of a proper message set

- For any D-ary prefix-free encoding of any proper message set for a DMS, the ratio of the average codeword length, E[W], to the average message length E[Y], satisfies

$$\frac{E[W]}{E[Y]} \geq \frac{H(U)}{\log D}$$

Proof:

- We know that for prefix free coding

$$E[W] \geq \frac{H(V)}{\log D}$$

Prefix free coding of a proper message set

- For any D-ary prefix-free encoding of any proper message set for a DMS, the ratio of the average codeword length, E[W], to the average message length E[Y], satisfies

$$\frac{E[W]}{E[Y]} \geq \frac{H(U)}{\log D}$$

Proof:

- We know that for prefix free coding

$$E[W] \geq \frac{H(V)}{\log D}$$

- Also, we know that for proper message set

$$H(V) = H(U)E[Y]$$

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Prefix free coding of a proper message set

- For any D-ary prefix-free encoding of any proper message set for a DMS, the ratio of the average codeword length, E[W], to the average message length E[Y], satisfies

$$\frac{E[W]}{E[Y]} \geq \frac{H(U)}{\log D}$$

Proof:

- We know that for prefix free coding

$$E[W] \geq \frac{H(V)}{\log D}$$

- Also, we know that for proper message set

$$H(V) = H(U)E[Y]$$

- Combining the above two equations, we get the desired inequality.

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Outline

Tunstall message set

- A message set with $M = K + q(K-1)$ messages is a *Tunstall message* set for a K-ary DMS if the K-ary rooted tree can be formed, beginning with the extended root, by q applications of the rule: extend the most likely leaf.

Tunstall message set

- A message set with $M = K + q(K-1)$ messages is a *Tunstall message* set for a K-ary DMS if the K-ary rooted tree can be formed, beginning with the extended root, by q applications of the rule: extend the most likely leaf.
- A proper message set for a K-ary DMS is a Tunstall message set if and only if, in its K-ary rooted tree every node is at least as probable as every leaf.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Tunstall message set

- A message set with $M = K + q(K-1)$ messages is a *Tunstall message* set for a K-ary DMS if the K-ary rooted tree can be formed, beginning with the extended root, by q applications of the rule: extend the most likely leaf.
- A proper message set for a K-ary DMS is a Tunstall message set if and only if, in its K-ary rooted tree every node is at least as probable as every leaf.
- A proper message set with M messages for a DMS maximizes the average message length E[Y], over all such proper message sets if and only if it is a Tunstall message set.

Outline

Tunstall coding

- Tunstall algorithm for optimum D-ary block encoding with blocklength N of a proper message set for a K-ary DMS with output variable U.

Tunstall coding

- Tunstall algorithm for optimum D-ary block encoding with blocklength N of a proper message set for a K-ary DMS with output variable U.

Step 0: Check to see if $D^N \geq K$. If not, abort, else calculate the quotient q when $D^N - K$ is divided by $K - 1$.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Tunstall coding

- Tunstall algorithm for optimum D-ary block encoding with blocklength N of a proper message set for a K-ary DMS with output variable U.

Step 0: Check to see if $D^N \geq K$. If not, abort, else calculate the quotient q when $D^N - K$ is divided by $K - 1$.

Step 1: Construct the Tunstall message set of size $M = K + q(K - 1)$ for the DMS by beginning from the extended root and making q extensions of the most likely leaf at each step.

Tunstall coding

- Tunstall algorithm for optimum D-ary block encoding with blocklength N of a proper message set for a K-ary DMS with output variable U.

Step 0: Check to see if $D^N \geq K$. If not, abort, else calculate the quotient q when $D^N - K$ is divided by $K - 1$.

Step 1: Construct the Tunstall message set of size $M = K + q(K - 1)$ for the DMS by beginning from the extended root and making q extensions of the most likely leaf at each step.

Step 2: Assign a distinct D-ary codeword of length N to each message in the Tunstall message set.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Tunstall coding

- For a binary memoryless source with $P_U(0) = 0.6$, construct binary Tunstall coding with block length, $N = 3$.

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

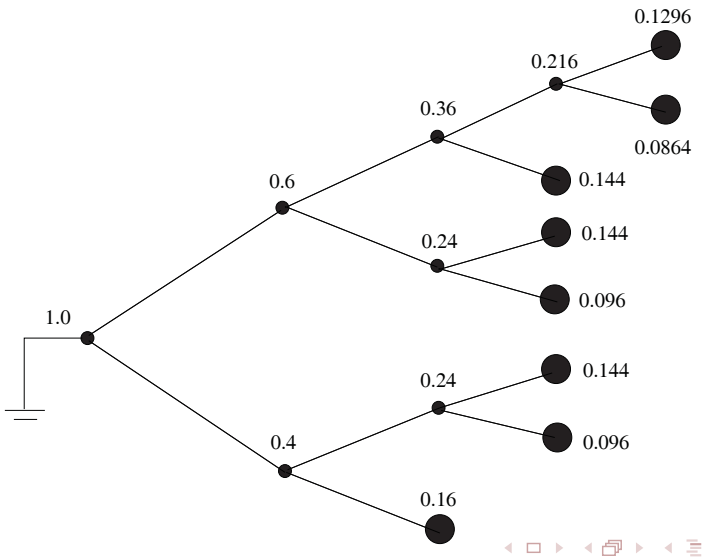
Tunstall coding

- For a binary memoryless source with $P_U(0) = 0.6$, construct binary Tunstall coding with block length, $N = 3$.
- Step 0 results in $q = 6$, thus $M = 8$.

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Tunstall coding

- For a binary memoryless source with $P_U(0) = 0.6$, construct binary Tunstall coding with block length, $N = 3$.
- Step 0 results in $q = 6$, thus $M = 8$.
- Step 1 results in following Tunstall message set



Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Tunstall coding

- Step 2 involves assigning codewords (can be done in more than one way) to the Tunstall message set.

Message	Codeword
0 0 0 0	0 0 0
0 0 0 1	1 1 0
0 0 1	0 0 1
0 1 0	0 1 0
0 1 1	0 1 1
1 0 0	1 0 0
1 0 1	1 0 1
1 1	1 1 1

Outline

Variable-Length-to-Block Coding Theorem for a DMS

- The ratio, $E[Y]/N$ of average message length to block length for an optimum D-ary block length N encoding of a proper message set for a K-ary DMS satisfies

$$\frac{\log D}{H(U)} - \frac{\log(2/p_{\min})}{NH(U)} < \frac{E[Y]}{N} < \frac{\log D}{H(U)}$$

  where $H(U)$ is the uncertainty of a single source letter and where $p_{\min} = \min_u P_U(u)$ is the probability of the least likely source letter.

Proof:

- The probability of least likely message in the message set is given by $Pp_{\min}$ where P is the probability of the node from which it stems.
- Since there are M messages, the probablity of least likely message can atmost be 1/M. Therefore

$$Pp_{\min} \leq \frac{1}{M}$$

Variable-Length-to-Block Coding Theorem for a DMS

- By Tunstall lemma, no leaf (message) has probability more than P. Therefore

$$P_V(v) \leq \frac{1}{Mp_{\min}} \quad \text{for all v}$$

Variable-Length-to-Block Coding Theorem for a DMS

- By Tunstall lemma, no leaf (message) has probability more than P. Therefore

$$P_V(v) \leq \frac{1}{Mp_{\min}} \quad \text{for all v}$$

- This implies that

$$\begin{aligned} -\log P_V(v) &\geq \log M - \log\left(\frac{1}{p_{\min}}\right) \\ or \quad H(V) &\geq \log M - \log\left(\frac{1}{p_{\min}}\right) \end{aligned}$$

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Variable-Length-to-Block Coding Theorem for a DMS

- By Tunstall lemma, no leaf (message) has probability more than P. Therefore

$$P_V(v) \leq \frac{1}{Mp_{\min}} \quad \text{for all v}$$

- This implies that

$$\begin{aligned} -\log P_V(v) &\geq \log M - \log\left(\frac{1}{p_{\min}}\right) \\ or \quad H(V) &\geq \log M - \log\left(\frac{1}{p_{\min}}\right) \end{aligned}$$

- Since $M + (K-1) \geq D^N$ and $M > K$, we get

$$\begin{aligned} 2M &\geq D^N \\ or \quad M &\geq \frac{D^N}{2} \end{aligned}$$

Variable-Length-to-Block Coding Theorem for a DMS

- Substituting this lower bound on M in the above expression, we get

$$H(V) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Variable-Length-to-Block Coding Theorem for a DMS

- Substituting this lower bound on M in the above expression, we get

$$H(V) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Substituting $H(V) = E[Y]H(U)$ in the above expression we get

$$E[Y]H(U) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

Variable-Length-to-Block Coding Theorem for a DMS

- Substituting this lower bound on M in the above expression, we get

$$H(V) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Substituting $H(V) = E[Y]H(U)$ in the above expression we get

$$E[Y]H(U) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Also $H(V) \leq \log D^N$. Hence we get,

$$N \log D \geq E[Y]H(U) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

Variable to block length coding Proper message set Assigning probabilities to a K-ary rooted tree corresponding to a proper message set Prefix free coding o

Variable-Length-to-Block Coding Theorem for a DMS

- Substituting this lower bound on M in the above expression, we get

$$H(V) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Substituting $H(V) = E[Y]H(U)$ in the above expression we get

$$E[Y]H(U) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Also $H(V) \leq \log D^N$. Hence we get,

$$N \log D \geq E[Y]H(U) \geq N \log D - \log\left(\frac{2}{p_{\min}}\right)$$

- Dividing the above expression by N H(U) we get

$$\frac{\log D}{H(U)} - \frac{\log(2/p_{\min})}{NH(U)} < \frac{E[Y]}{N} < \frac{\log D}{H(U)}$$