Note Title 2/24/2006

TODAY : DATA COMPRESSION

SOME LIMITS TO COMPRESSIBILITY.

First: A Review of the last two lectures.

first few lectures: Studied properties of

- entropy, conditional entropy, mutual information, relative entropy
- One consequence: if variable X has large conditional entropy given Y then any attempt to determine X from Y has large error (Fano's lemma)

Lecture 4: restrict to variables $\bar{X}$, $\bar{Y}$ that are sequence of i.i.d. variables distributed according to X

when $\bar{X} = (X_1, \dots, X_n)$ then

w.p. $\geq 1 - \delta$

(AEP) $\Pr_{\bar{X}}\left[p(\bar{X}) \in \left[2^{-(H(x)+\epsilon)n}, 2^{-(H(x)-\epsilon)n} \right] \right] \geq 1-\delta$

(Typical Set) $A_\epsilon^{(n)} = \{ \bar{X} \in \Omega^n \mid p(\bar{X}) \in [\quad] \}$

$A_\epsilon^{(n)}$ has size $\doteq 2^{H(x) \cdot n}$

with each el't having prob. $2^{-H(x) \cdot n}$.

(flat dist. on small set)

[didn't say but $\delta \approx \exp(-\epsilon^2 n)$]

Leads to formal proof that

- X can be compressed into $\sim H(x) \cdot n$ bits in expectation

- X can't be compressed to much less than $H(x) \cdot n$ bits

LECTURE 5: $\overline{X}$ = stochastic process

(esp. time-invariant,
irreducible,
aperiodic, Markov chain)

Stochastic $\supseteq$ Stationary $\supseteq$ Time Inv. Markov Chain in stat. distribution

Entropy Rate of ( ) : $H(X_2 | X_1)$

$$= \sum_{i,j} \mu_i P_{ij} \log \frac{1}{P_{ij}}$$

where P = transition probability matrix

μ = stationary dist.

(AEP):

$$P_r\left[p(\overline{x}) \in \left[2^{-(H(\mathcal{X})+\epsilon)n}, 2^{-(H(\mathcal{X})-\epsilon)n}\right]\right] \geq 1-\delta$$

$$\delta \approx \exp(-\epsilon^2 n)$$

Finally

Hidden Markov chains

$$\begin{array}{ccccc} X_1 & \rightarrow & X_2 & - \ - \ - \rightarrow & X_n \\ \downarrow & & \downarrow & & \downarrow \\ Y_1 & & Y_2 & - \ - \ - & Y_n \end{array}$$

Entropy rate

$$H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n \mid Y_1 \dots Y_{n-1})$$

Exists, but not easy to compute

Now to today's lecture

Goal : to "compress" X into string of letters over D.

but ....

(1) What is compression?

(2) What is the objective?

E.g. $X \in \{1 \dots N\}$

$X = i$ w.p. p_i

Is $C(i) = 0$ $\forall i$ "compression"?

Trade off between "compression" & loss of information ....

We wish to compress without losing (too much) information.

Terminology

Ω = space of X

D = alphabet for compression

$D^* = \bigcup_{n \geq 0} D^n$ = set of finite strings over D

for $w = (w_1 \ldots w_n) \in D^n$, $|w| = n$.

Compression algorithm

$$C : \Omega \to D^*$$

accompained by Decompressor

$$Dec : D^* \to \Omega$$

Lossy compression : $\Pr_X \left[Dec(C(X)) \neq X \right]$ = small

Lossless compression : $\forall x \quad Dec(C(x)) = x$.

if C s.t. $\exists$ Dec s.t.

(C, Dec) lead to <u>lossless coding</u>

then C is called $\boxed{\text{Non-Singular}}$

We will talk only about <u>Non-singular</u>

(Lossless Coding) but many results

extend to Lossy Coding as well

(er... we mean limitations results of course)

<u>Goal / Measure = ?</u>

Expected Length : $\underset{x}{E}\left[\,|C(x)|\,\right]$

↑

Minimize

__________ x __________

Example

$$\Omega = \left\{ \begin{array}{lll} 1 & \text{w.p.} & \frac{1}{2} \\ 2 & \text{w.p.} & \frac{1}{4} \\ 3 & \text{w.p.} & \frac{1}{8} \\ 4 & \text{w.p.} & \frac{1}{8} \end{array} \right\}$$

$D = \{0, 1\}$

Code 1:

$C_1(1) = 00$

$C_1(2) = 01$

$C_1(3) = 10$

$C_1(4) = 11$

$$E\left[|C_1(x)|\right] = 2$$

Code 2:

$C_2(1) = 0$ $\qquad$ $C_2(3) = 00$

$C_2(2) = 1$ $\qquad$ $C_2(4) = 10$

$$E\left[\, |C_2(x)| \,\right] = \frac{3}{4}\cdot 1 + \frac{1}{4}\cdot 2 = \frac{5}{4}$$

- But something unsatisfactory about second code?

Second code is not <u>extensible</u>.

<u>Real goal</u> to get a code for sequence $(X_1, \ldots X_n)$ where $X_i \sim X$ (say i.i.d.)

Code for X suggests extension code of X^n

$$C(X_1 \ldots X_n) = C(X_1), C(X_2) \ldots C(X_n)$$

Code C_2 above is non-singular but extension is not!

(E.g. $C_2(1\ 1) = C_2(3)$

& so $C_2(11\ 3) = C_2(3\ 11)$)

Motivates another definition

C is <u>uniquely decodable</u>

if $\forall n$, n-extension of C is

non-singular.

<u>Example</u> of uniquely-decodable code

$C_3(1) = 0$ — 1 w.p. $1/2$

$C_3(2) = 10$ — 2 w.p. $1/4$

$C_3(3) = 110$ — 3 " $1/8$

$C_3(4) = 111$ — 4 " $1/8$

$E[\text{length}] = 1.75$

Another example

$C_4(1) = 0$ $C_4(2) = 01$ $C_4(3) = 011$ $C_4(4) = 111$

Proof of unique decodability of C_3.

Property of C_3

- $\forall\ x, y \in \Omega$ $C(x)$ is not a prefix of $C(y)$

Motivates another definition

$$\boxed{C \text{ is a prefix code if } \forall\ x, y \in \Omega \quad C(x) \text{ not prefix of } C(y)}$$

Claim: Prefix code is uniquely decodable

Proof: Let $C(x_1 \ldots x_n) = w_1 \ldots w_m$
$= C(y_1 \ldots y_\ell)$

then $x_1 = y_1$ (since $C(x_1)$ must

be a prefix of $C(y_1)$ or vise versa).

But now by induction we have

$C(x_2 \dots x_n) = w_{k+1} \dots w_m = C(y_2 \dots y_\ell)$

Where $C(x_1) = w_1 \dots w_k = C(y_1)$

& so $x_2 \dots x_n = y_2 \dots y_\ell$ [& $n = \ell$].

Prof of unique decodability of C_4

(well $C_4 = C_3$ in reverse)

Prefix codes (C_3) are nicer than non-prefix codes (C_4) since they can be decoded online..

Hence also called <u>instantaneous</u>

In the rest of the lecture we will study

Limitations on <u>compressibility</u>

- Kraft lower bound
- McMillan lower bound
- Entropy lower bound

<u>Next lecture will match goals</u>

<u>Main target for this lecture</u>

- $$\mathop{E}_{X}\left[\,|C(X)|\,\right] \geq \frac{H(X)}{\log D}$$

for any prefix-free or uniquely decodable code C.

• Actually yields if

$$\mathop{E}_{x}\left[\,|C(x)|\,\right] = K$$

then $K + O(\log K) \geq \dfrac{H(X)}{\log D}$

for any non-singular code [Won't show this but will show $K + 2\sqrt{K} + 2 \geq \dots$]

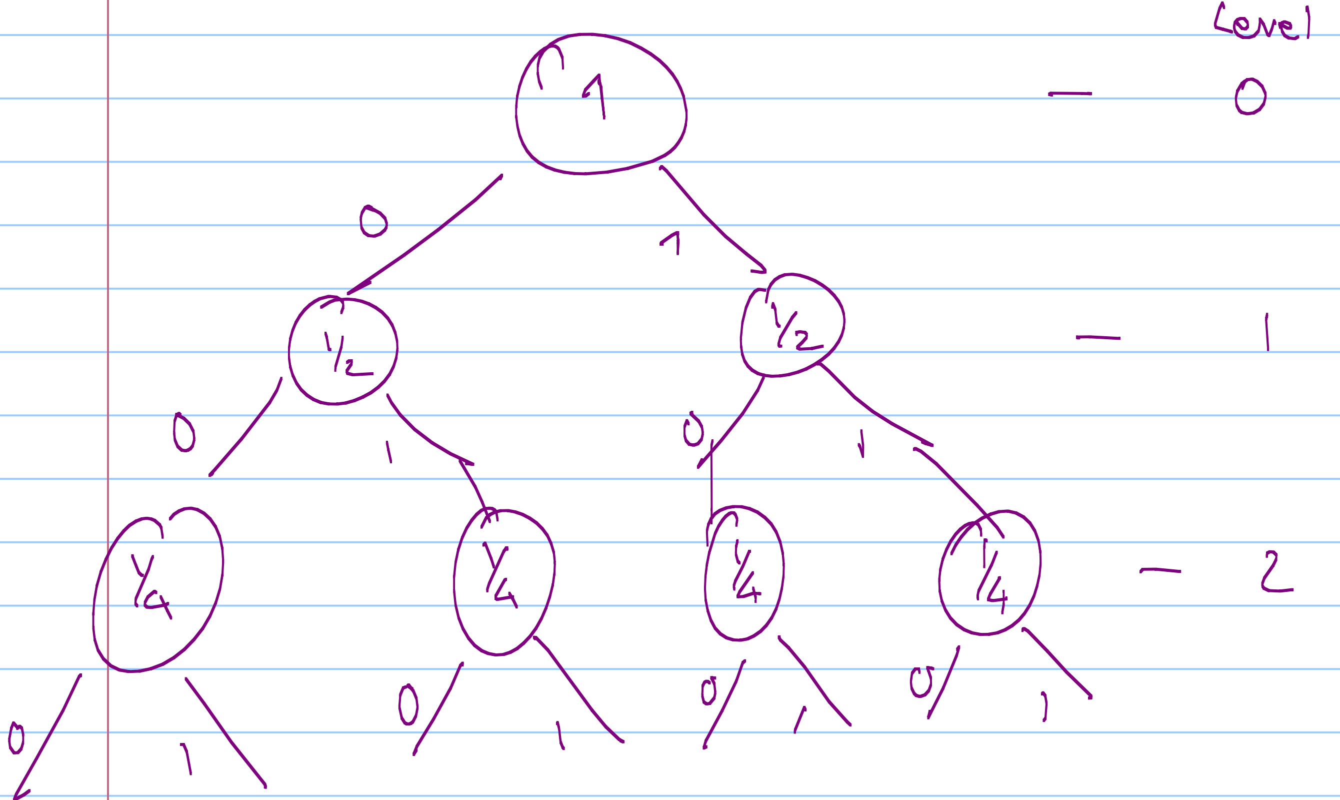
Main tool towards Entropy bound

1. Kraft's Inequality

Theorem (Kraft): if $C(i)$ has length l_i for $i = 1 \dots N$ over D-ary alphabet

then $\sum D^{-l_i} \leq 1$, if C prefix code.

Proof : Consider D-ary tree with branches denoting various letters of D. example with $D=2$



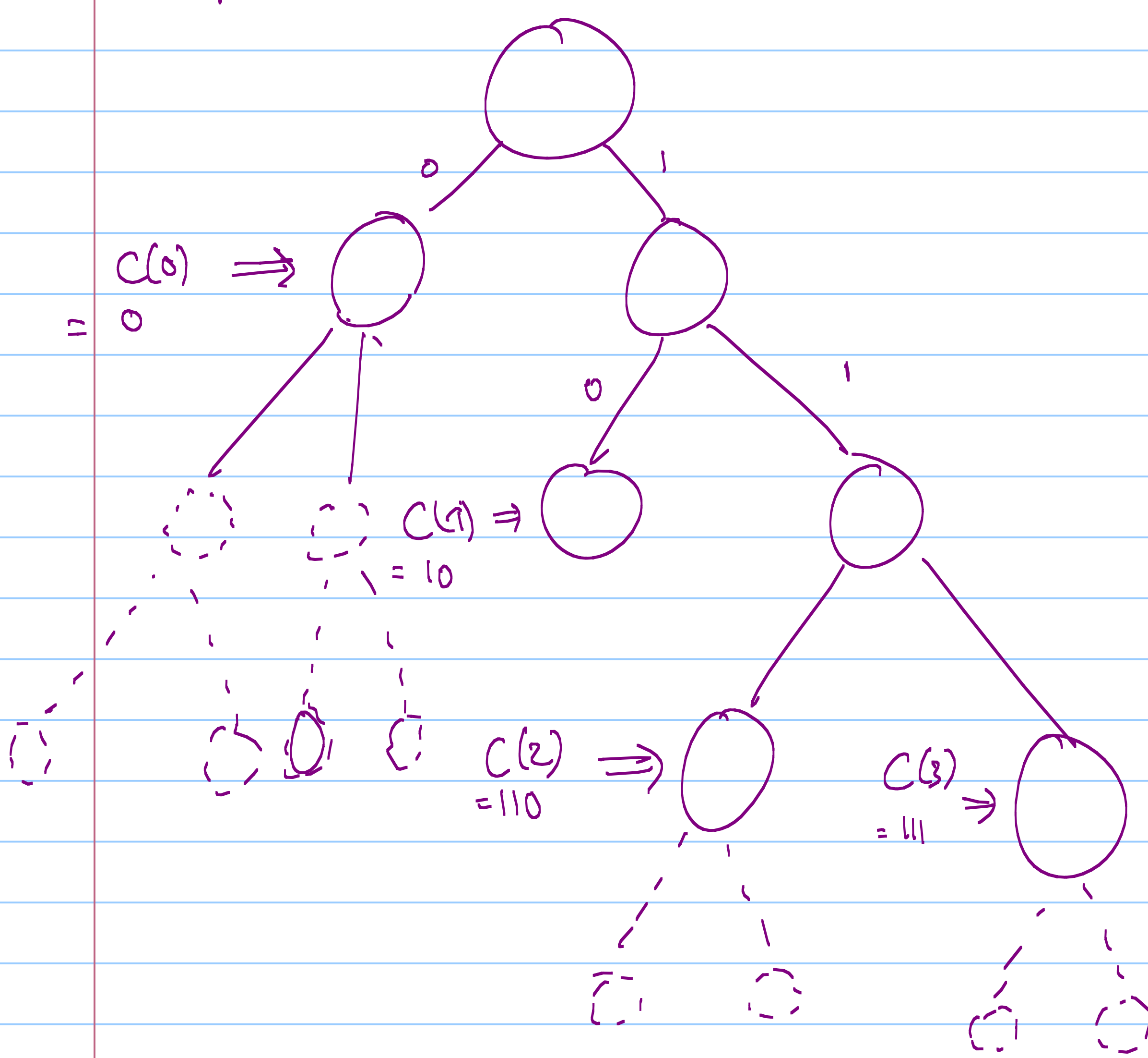
for node at level i associate weight D^{-i}. Note that the weight of the root $= 1$;

Weight of every node = Sum of wts. of children

<u>Now</u> lets understand "C(i)"'s.

for every i mark node C(i).

example



Retain only the part of the tree that lies on the path from some $C(i)$ to the root.

Reassign weights, leaving weights of $C(i)$ as they were & putting in for every node weight = sum of weight of its children.

- On the one hand weights of nodes don't go up
- On the other hand weight of root

$$= \sum_{i=1}^{N} D^{-l_i}$$

Thus $\sum_{i=1}^{N} D^{-l_i} \leq 1$

McMillan's Bound

Theorem: if C is uniquely decodable & $|C(i)| = l_i$ then

$$\sum D^{-l_i} \leq 1$$

Proof: (Optional. See Cover & Thomas)

Entropy lower bound

Note: Kraft says nothing about probabilities

To relate to Expected Decoding length...

$$\mathop{E}_{X}\left[\,|C(x)|\,\right] = \sum p_i l_i$$

But now lets write $l_i = \dfrac{-\log D^{-l_i}}{\log D}$

$$E_x\left[\,|C(x)|\,\right] = -\sum_{i=1}^{N} p_i \frac{\log D^{-l_i}}{\log D}$$

But $\sum D^{-l_i} \leq 1$ [Kraft]

Let $q_i = D^{-l_i}$

& let $\sum D^{-l_i} = 1 - q_0$

$$E_x\left[\,|C(x)|\,\right] = -\sum_{i=1}^{N} \frac{p_i \log q_i}{\log D}$$

$$= \frac{H(x)}{\log D} + \frac{D(p\|q)}{\log D} \geq H(x)$$

Conclude

Theorem: if C is prefix free or uniquely decodable then for $X \sim p$

$$\mathop{E}_{X}\left[\,|C(X)|\,\right] \geq H(X)$$

Addendum What about non uniquely-decodable codes?

Answer 1: Not a fair question. Really need unique decodability.

Answer 2: Doesn't make much difference.

ξ_1

Lemma: if C is non-singular code of exp. length K, then $\exists$ C' prefix free of expected length $K + 2\lceil\sqrt{K}\rceil$.

(Can probably do better....)

Proof:

Given C first produce C_1 s.t. $\forall i$ $|C_1(i)|$ is a multiple of $\lceil\sqrt{K}\rceil$. We have

$$\text{(*)} \quad \underset{x}{E}\left[|C_1(x)|\right] \leq E\left[|C(x)|\right] + \lceil\sqrt{K}\rceil$$

(since no string extends by more than $\lceil\sqrt{K}\rceil$).

Now produce C_2 where $|C_2(i)|$ is a multiple of $\lceil\sqrt{K}\rceil + 1$

as follows.

Suppose $C_1(i) = \boxed{w_1}\boxed{w_2} \cdots \boxed{w_c}$

where $|w_j| = \sqrt{K}$

then $C_2(i) = \boxed{w_1}\boxed{0}\boxed{w_2}\boxed{0} \cdots \boxed{w_c}\boxed{1}$

all zeroes ONE

Claim. C_2 is prefix free (why?)

$$\Rightarrow E\left[|C_2(x)|\right] \le \frac{\lceil\sqrt{K}\rceil + 1}{\lceil\sqrt{K}\rceil} \cdot E\left[|C_1(x)|\right]$$

$$= (\sqrt{K}+1)(\sqrt{K}+1) \le K + O(\sqrt{K}).$$

Conclude

Essentially for any reasonable encoding

$$E\left[\,|C(x)|\,\right] \geq \frac{H(x)}{\log D}$$

But is this tight?

Will see in next lecture.