Lecture 2

1/6/2010

overview of classical information theory
(descriptive fashion)

Data Compression Example

Alice + Bob are connected by a noiseless bit channel $- p(y|x) = \delta_{y,x}$

Alice has four symbols $\{a, b, c, d\}$

Suppose either she or someone else is choosing them at random accding to

$$\Pr\{a\} = 1/2$$
$$\Pr\{b\} = 1/8$$
$$\Pr\{c\} = 1/4$$
$$\Pr\{d\} = 1/8$$

chooses symbols independently

sps. channel accepts only bits.

Naive code is

$a \to 00$, $b \to 01$, $c \to 10$, $d \to 11$

Performance measured by expected length

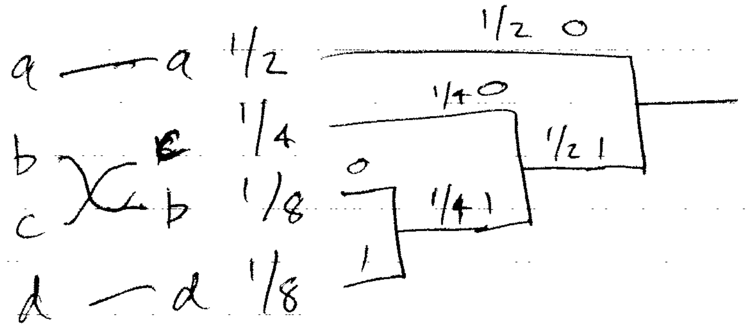
In this case, expected length is

$$\frac{1}{2} \cdot 2 + \frac{1}{8} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 2 = 2$$

1/6/2011

But they can do better b/c dist. is skewed

- use shorter codewords for more likely symbols & longer for less likely
- can do Huffman code



a → 0

b → 110

c → 10

d → 111

Any coded sequence is uniquely decodable

E.g.)

0 0 1 1 0 1 0 1 1 1 0 1 0 1 0 0 0 1 0

a a b c d a c c a a c

Expected length of coding scheme

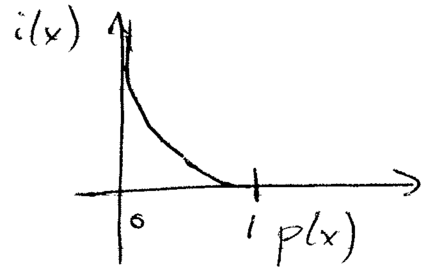
$$\frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 = \frac{7}{4}$$

scheme suggests a measure of information
(want a measure that is higher for less likely events + lower for more likely)

one measure is

$$i(x) \equiv \log_2\left(\frac{1}{p(x)}\right)$$

called "information content"



- it so happens that the length of each symbol in Huffman code is equal to its information content
- scheme is special because we had powers of two

info. content is additive for independent events
our info. source is memoryless

$$\Rightarrow \quad i(x_1, x_2) = -\log(p(x_1, x_2))$$
$$= -\log(p(x_1)) - \log(p(x_2)) = i(x_1) + i(x_2)$$

property is very important

expected information content is

$$\sum_x p(x)\, i(x) = -\sum_x p(x) \log(p(x))$$

this is the <u>entropy</u> of the source

1/6/2011

Shannon's source coding theorem

What is the ultimate limit on the compressibility of information?

- need more general techniques + setting to answer this question in a satisfying way

- given random variable X, the information content of it is itself a random variable

$$i(X) = -\log(p_X(x)).$$

seemingly self-referential, but OK

on to source coding,

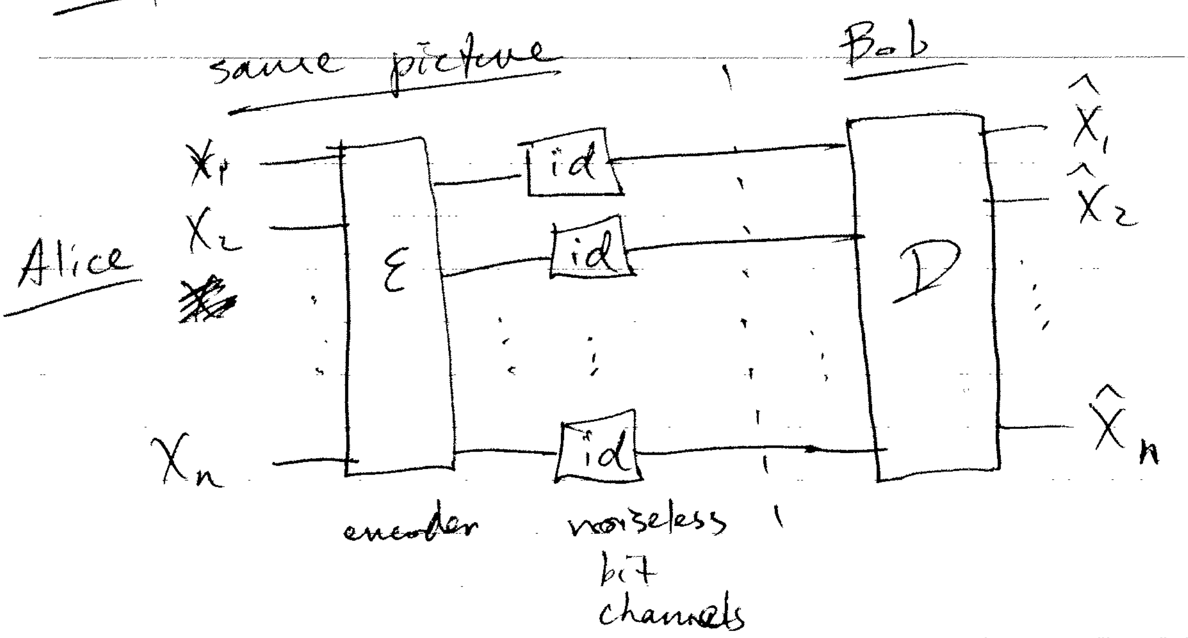
could associate a binary codeword for each x ~~as before, but this scheme may lose~~ some efficiency if alphabet or probabilities are not a power of two.

Shannon's idea: 1) block coding

Source emits large number of realizations + then code data as emitted block

2) allow for slight error, but show that it vanishes asymptotically

<u>1/6/2011</u>



Consider the distribution of the source under the IID assumption

$$
\begin{aligned}
p_{X^n}(x^n) &= p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \\
&= p_{X_1}(x_1)\cdots p_{X_n}(x_n) \\
&= p_X(x_1)\cdots p_X(x_n) \\
&= \prod_{i=1}^{n} p_X(x_i)
\end{aligned}
$$

Let $a_1,\ldots,a_{|\mathcal{X}|}$ denote letters in $\mathcal{X}$

+ Let $N(a_i|x^n)$ denote number of occurrences of a_i in x^n

then

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \underbrace{\prod_{j=1}^{|\mathcal{X}|} p_X(a_j)^{N(a_j|x^n)}}_{\text{probability for a particular sequence}}$$

1/6/2011

Consider information content of ~~the ab~~ a random sequence:

$$\frac{i(X^n)}{n} = -\frac{1}{n}\log\left(p_{X^n}(X^n)\right) \quad \text{(AKA sample entropy)}$$

random variable

we can use formula from before + random quantity $N(a_j|X^n)$

$$-\frac{1}{n}\log\left(p_{X^n}(X^n)\right) = -\frac{1}{n}\log\left(\prod_{j=1}^{|X|} p_X(a_j)^{N(a_j|X^n)}\right)$$

$$= -\frac{1}{n}\sum_{j=1}^{|X|}\log\left(p_X(a_j)^{N(a_j|X^n)}\right)$$

$$= \sum_{j=1}^{|X|} -\left(\frac{1}{n}N(a_j|X^n)\log\left(p_X(a_j)\right)\right)$$

state law of large numbers

$$\lim_{n\to\infty} \Pr\{|\overline{X^n} - \mu| < \epsilon\} = 1$$

where $\overline{X^n}$

implies $\dfrac{N(a_j|X^n)}{n} \to p_X(x)$

$$\Rightarrow \lim_{n\to\infty} \Pr\left\{\left|\frac{i(X^n)}{n} - H(X)\right| < \epsilon\right\} = 1$$

"It is highly likely that the source emits a sequence w/ sample entropy close to the true entropy."

1/6/2011

Leads to the notion of "high probability set"

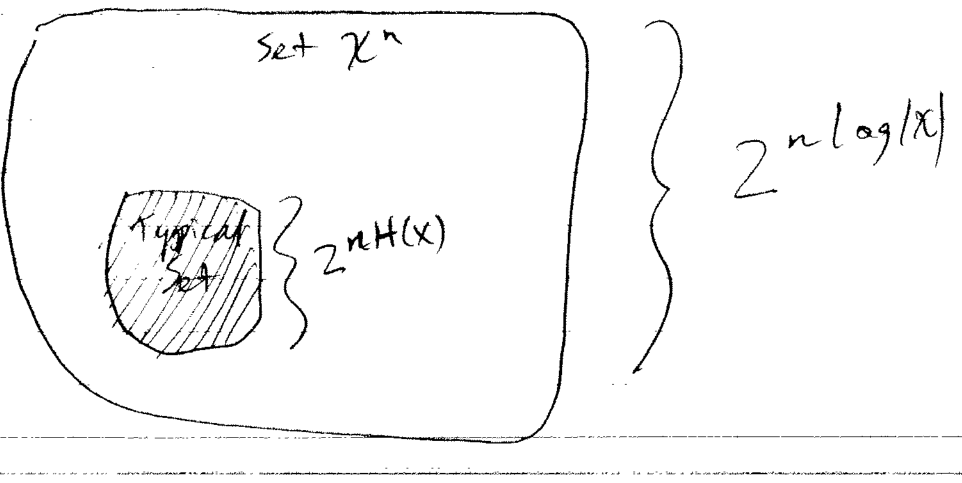
or "typical set"

set of sequences for which the sample entropy is close to the true entropy

asymptotically, this set has all of the probability.

Since we're concerned w/ probability of error in communication, it seems reasonable to give attention only to the high probability set of sequences

- wonderful thing about typical set is that its size is $2^{nH(X)}$ whereas set of all sequences is $|\mathcal{X}|^n = 2^{n \log|\mathcal{X}|}$
- probability is concentrating on an exponentially small set



1/6/2011

Coding Strategy

"Keep the typical sequences + throw away the rest"

(b/c the others' probability is negligible)

sequence x^n $\longrightarrow$ $\boxed{\text{is it typical?}}$ $\xrightarrow{\text{YES}}$ map x^n to element of $\{0,1\}^{nH(x)}$ (size $2^{nH(x)}$)

$\downarrow$ NO

transmit error flag

$\downarrow$

transmit over $nH(x)$ noiseless bit channels

$\downarrow$

Bob decode

coding scheme has vanishing error in asymptotic limit

rate of scheme is

$$\frac{\#\text{ channel bits}}{\#\text{ source symbols}} = \frac{nH(x)}{n} = H(x)$$

any rate $R > H(x)$ is achievable in the sense that $\forall \epsilon > 0$ + sufficiently large n there exists an (n, R, ϵ) compression code

this is called direct part of coding theorem

demonstrate an achievable strategy

If rate $\geq$ entropy, there exists achievable scheme

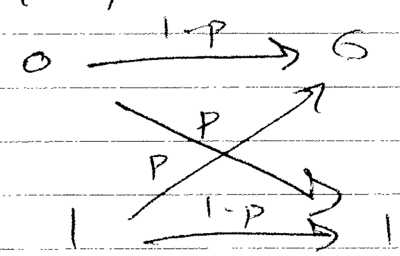
1/6/2011

Converse Part (optimality)

If $\exists$ achievable scheme, then rate $\geq$ entropy.

Simple Example of Error Correction

binary symmetric channel



transmit data as is, then error probability is

$$p(e) = p(e|0)\,p(0) + p(e|1)\,p(1)$$
$$= p$$

would like to supress errors

can try to engineer better channel, but we would like a "systems engineering" solution

use a code

rate is 1/3

$0 \rightarrow 000$

$1 \rightarrow 111$

take ~~a~~ a majority vote at output

1/6/2011

if Alice sends 0, then

output	Prob	Sum	
000	$(1-p)^3$	$(1-p)^3$	$\}$ un correc
001, 010, 100	$p(1-p)^2$	$3p(1-p)^2$	
110, 101, 011	$p^2(1-p)$	$3p^2(1-p)$	$\}$ error
111	p^3	p^3	

$$p(e) = 3p^2(1-p) + p^3$$

$$= 3p^2 - 2p^3$$

coding is helping if

$$3p^2 - 2p^3 < p$$

or equivalently

$$0 < p(2p-1)(p-1)$$

only values of p satisfying are

$$0 < p < 1/2$$

error prob. goes like $O(p^2)$

How to reduce even more?
concatenate...

$0 \to 000 \longrightarrow 000 \quad 000 \quad 000$

$1 \to 111 \longrightarrow 111 \quad 111 \quad 111$

rate 1/9

1/6/2011

error prob. goes like

$$3p^2(e) - 2p^3(e) = O(p^4)$$

we would like error free communication...

keep on concatenating

rate is

$\frac{1}{3^n}$ for error suppression $O(p^{2n})$

rate goes to zero to make error prob go to zero?

Shannon's Channel Coding Theorem

Alice chooses message from a set

$$[M] \equiv \{1, \ldots, M\}$$

(chooses uniformly at random)

set requires $\log_2 M$ bits to represent it

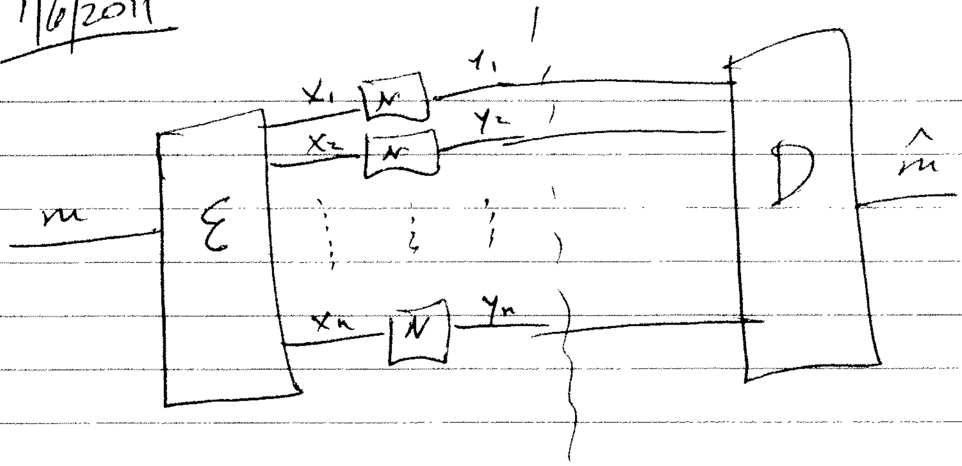
one "layer of randomness"

noisy channel $\mathcal{N} \equiv p_{Y|X}(y|x)$

IID

$$p_{Y^n|X^n}(y^n|x^n) \equiv \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

1/6/2011



for every message m, there is some codeword $x^n(m)$ + Bob makes his best estimate of $x^n(m)$ from received sequence y

$$\text{rate} = \frac{\#\text{ message bits}}{\#\text{ channel uses}} = \frac{\log M}{n}$$

capacity = highest rate of reliable comm.

$$\mathcal{C} \equiv \{x^n(m)\}_{m \in [M]}$$

$p_e(m, \mathcal{C})$ = prob. of error for message m under code $\mathcal{C}$

$$\bar{p}_e(\mathcal{C}) = \frac{1}{M} \sum_{m=1}^{M} p_e(m, \mathcal{C}) = \text{avg. prob. of error}$$

$$p_e^*(\mathcal{C}) = \max_m p_e(m, \mathcal{C}) = \text{maximal prob. of error}$$

1/6/2011

Those error criteria are difficult to analyze.

Shannon's idea: choose the code randomly & analyze the expectation of the average error prob. rather

use some distribution $p_X(x)$

for message 1, choose $x_1(1)$ according to $p(x)$

$x_2(1)$ " " $p(x)$

⋮

$x_n(1)$ " " $p(x)$

$x^n(1) = x_1(1)\, x_2(1) \cdots x_n(1)$

same for $x^n(2)$ (message 2)

$x^n(2) = x_1(2)\, x_2(2) \cdots x_n(2)$

⋮

$x^n(M) = x_1(M)\, x_2(M) \cdots x_n(M)$

every codeword chosen independently of the message m

probability for a particular code c_0 is

$$p_C(c_0) = \prod_{m=1}^{M} \prod_{i=1}^{n} p_X(x_i(m))$$

expectation of the average error probability

$$\mathbb{E}_C\{\bar{p}_e(C)\} = \mathbb{E}_C\{\frac{1}{M}\sum_{m=1}^{M} p_e(m,C)\}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \underbrace{\mathbb{E}_C\{p_e(m,C)\}}_{\text{expectation does not depend on particular message } m}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_C\{p_e(1,C)\}$$

$$= \underbrace{\mathbb{E}_C\{p_e(1,C)\}}_{\text{much easier to analyze}}$$

show that

$$\mathbb{E}_C\{p_e(1,C)\} \le \epsilon$$

$\Rightarrow$ there exists some code C_0 such that

$$\bar{p}_e(C) \le \epsilon$$

can use this to bound $p_e^*(c) \le 2\epsilon$

What about size of code?

rate is $R = \frac{\log M}{n}$

so think of message set size as

$$M = 2^{nR}$$

1/6/2011

Alice is using distribution $p(x)$ to generate sequences x^n.

Highly likely that x^n is typical

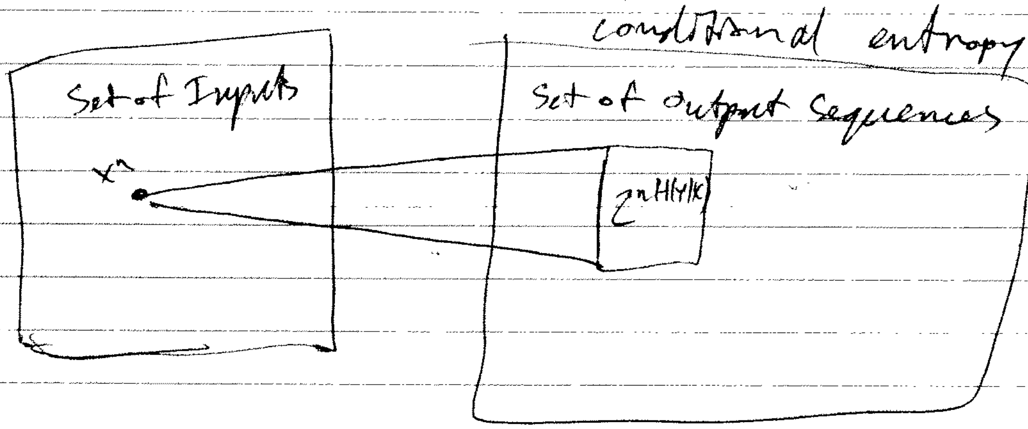
$$x^n \longrightarrow p_{Y^n|X^n}(y^n|x^n) \longrightarrow Y^n$$

number of output sequences likely to correspond to x^n

tool: conditional typicality

set of size $2^{nH(Y|X)}$

conditional entropy



From Bob's perspective, distribution he "sees" is $p(y) = \sum_x p(y|x)\, p(x)$

holds that $H(Y) \geq H(Y|X)$

1/6/2011

Alice transmits codeword $x^n(m)$

Bob is ignorant of m so from his perspective, sequences generated according to $p(y)$

1) Determine if y^n lies in typical set for y of size $2^{nH(Y)}$

If not, declare error. If so, proceed

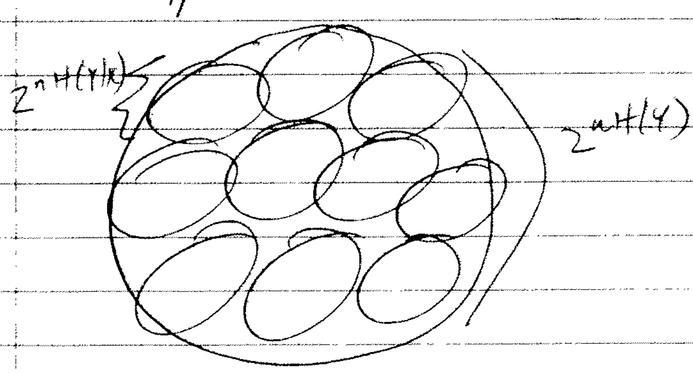
2) Determine (w/ knowledge of code) the conditionally typical set from which y^n would come.

If y^n lies in wrong set, error occurs.

If not, done.

How many message can we decode?

Same question as how many ~~message~~ conditionally typical sets we can fit into typical set



$$2^{nR} \approx \frac{2^{nH(Y)}}{2^{nH(Y|X)}}$$

$$= 2^{n(H(Y)-H(Y|X))}$$

$$= 2^{nI(X;Y)}$$

choose best dist. for generating code

$R = \max I(X;Y)$