Mini-Course L1: Introduction

Sunday, June 5, 2016 1:52 PM

} We think of data as being a sequence of (independent, identically distributed) RVs $X_1, X_2, X_3, \dots$ taking values in a (finite) symbol space $\mathcal{X}$, called the alphabet.

E.g. ① English alphabet

② Morse code

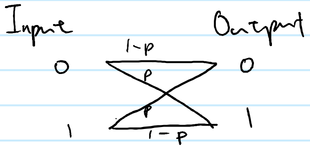
③ Binary encoding of pictures, videos, music, etc.

A message/signal of length n is an instance $(X_1, \dots, X_n)$

Q1. Can we compress the data? i.e. can we encode the data differently to reduce the expected length of a message of length n?

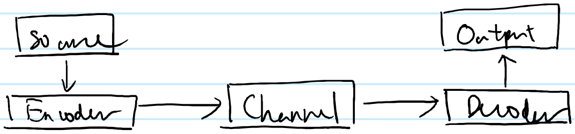
Q2. Can we send messages reliably over unreliable channels?

E.g Binary symmetric channel



Def. A channel is a conditional probability transition matrix $\mathbb{P}(Y=y \mid X=x)$

In order to use a channel more reliably, can insert encoders and decoders



E.g. We can have repetition codes:

$$1101 \mapsto 111\ 111\ 000\ 111 \longrightarrow 111\ 111\ 000\ 110 \overset{\text{majority vote}}{\longmapsto} 1101$$

$\mathbb{P}(\text{Error}) = P(2 \text{ or more flips}) = p^3 + 3p^2(1-p)$

If $p = 0.1$, then $P_e \sim 3 \cdot 0.01 = 0.03$. How good is this?

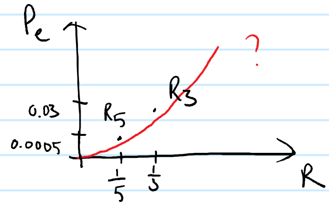
Def. An (M,n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of

① An index set $\{1, \dots, M\}$

② An encoding function $X^n : \{1, \dots, M\} \to \mathcal{X}^n$

The rate R of an (M,n) code is $R = \frac{\log_2 M}{n}$ bits per transmission.

In our case $M=2$, $n=3$ ($X^n: 0 \mapsto 000$, $X^n: 1 \mapsto 111$), so the rate is $R=\frac{1}{3}$. Can do a scatterplot for a channel



What is the best we can hope for?

To answer these questions, we need a notion of the amount of information that is contained in data.

Def Let X be a discrete RV taking values in the set $\mathcal{X}$. The information content of an outcome $x \in \mathcal{X}$ is defined by $h(x) = \log_2 \frac{1}{\mathbb{P}(X=x)}$.

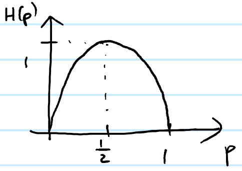
The entropy of X is its expected information content, ie

$$H(X) = \mathbb{E}[h(X)] = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)}$$

How does the entropy function look like? Since $\mathcal{X}$ is finite, we can parametrize all distributions on $\mathcal{X}$ using a simplex Δ. Given $(\theta_1, \dots, \theta_m) \in \Delta$, we then have $H(\theta) = -\sum_{i=1}^{m} \theta_i \log \theta_i$

Note that $t \mapsto t \log t$ is convex, so H is a concave function on Δ.

In particular, for Bernoulli RVs, $m=1$, and denoting $H(p) = H(p, 1-p)$, we have



Let us now see some places where entropy shows up.

① Bent coin lottery

Suppose we have a lottery that works as follows: A bent coin with

$$\mathbb{P}(\text{tails}) = 0.9 \qquad \mathbb{P}(\text{heads}) = 0.1$$

is thrown 1000 times. An outcome of the lottery is thus a binary string of 0s and 1s. Tickets are sold for each possible outcome.

Q. How many tickets must you buy to be 95% sure of winning?

A. Pick the most likely outcomes first. So you want a min n such that

$$\sum^{n} \binom{1000}{k} 0.9^{1000-k} \, 0.1^{k} \geq 0.95$$

A. Pick the most likely outcomes just. So you want a min n such that

$$\sum_{k=1}^{n} \binom{1000}{k} 0.9^{1000-k} \, 0.1^{k} \geq 0.95$$

For simplicity, use normal approximation. Let S = no. of heads. Then $ES = 100$

$Var(S) = np(1-p) = 1000 \cdot 0.1 \cdot 0.9 = 90 \rightsquigarrow sd \sim 9.5$ so $\mathbb{P}(S - ES > 19) \leq 0.05$

Hence, just need to buy tickets with up to 119 heads, that is

$$\binom{1000}{0} \ \binom{1000}{1} + \cdots + \binom{1000}{119} \sim 2^{\log \binom{1000}{119}} \sim 2^{1000 H(\frac{119}{1000})} \sim 2^{530}$$

On the other hand if the coin were fair, would need to buy $0.95 \cdot 2^{1000} \approx 2^{1000}$ tickets

We see that a small fraction of the tickets contain almost all the "information" about the lottery.

<u>Lem</u> $\log \binom{n}{k} \sim nH(\frac{k}{n}) + o(n)$

Pf. $\log n! \sim n \log(\frac{n}{e}) + \frac{1}{2}\log(2\pi n)$ by Stirling, so

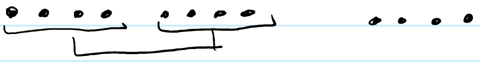
$$\log \binom{n}{k} \sim n \log n - (n-k)\log(n-k) - k \log k + \frac{1}{2}\log(2\pi \cdot \frac{n}{k(n-k)})$$

$$= (n-k)\log \frac{n}{n-k} + k \log \frac{n}{k} + o(n) \quad \square$$

② <u>Weighing problem</u>: We are given 12 balls that look the same, but one of which is heavier or lighter than the rest. We have a weighing balance, and want to come up with a weighing strategy that identifies the special ball with as few weighs as possible.

The best method is to create RVs with the largest entropy

Step 1: $4 \vee 4 \vee 4$



$H(X) = \log_2 3$

Continue this next lecture after we talk about conditional entropy.

Now let us return to the bent coin lottery. We see that a small fraction of the tickets accounted for most of the probable outcomes. This phenomenon is known as the asymptotic equipartition property and holds much more generally.

<u>Thm</u> $X_1, X_2, \ldots \overset{iid}{\sim} p(x)$. Then $-\frac{1}{n}\log p(X_1, \ldots, X_n) \xrightarrow{a.s.} H(X)$

Pf. LLN.

Def. For fixed $\varepsilon > 0$, $n \in \mathbb{N}$, we call the ε-typical set, denoted $A_\varepsilon^{(n)}$ the set of all sequences $(x_1, \dots, x_n)$ s.t. $2^{-n(H(X)+\varepsilon)} \le p(x_1, \dots, x_n) \le 2^{-n(H(X)-\varepsilon)}$

Thm 3.1.2 ① $(x_1, \dots, x_n) \in A_\varepsilon^{(n)} \iff 2^{-n(H(X)+\varepsilon)} \le p(x_1, \dots, x_n) \le 2^{-n(H(X)-\varepsilon)}$

② $\forall \varepsilon$, for n large enough, $\mathbb{P}(A_\varepsilon^{(n)}) > 1-\varepsilon$

③ $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$

④ $|A_\varepsilon^{(n)}| \ge (1-\varepsilon) 2^{n(H(X)-\varepsilon)}$ for n large enough

Pf. ③ $1 \ge \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p(x_1, \dots, x_n)$

$\ge |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)}$

④ For n large enough,

$$(1-\varepsilon) \le \mathbb{P}(A_\varepsilon^{(n)}) = \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p(x_1, \dots, x_n) \le |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)}$$ ■

Thm 3.2.1. Let $X_1, \dots, X_n$ be i.i.d., $\varepsilon > 0$. $\exists$ a code mapping strings of length n into binary strings s.t. the mapping is 1-1, and $\mathbb{E}[\frac{1}{n} l(X_1, \dots, X_n)] \le H(X) + \varepsilon$ for n large enough.

Pf. The idea is to consider the typical set and the non-typical set separately.

We first order all the elements in the typical set. There are $\le 2^{n(H(X)+\varepsilon)}$ sequences, so we can code each of them uniquely using a binary string of length $\le n(H(X)+\varepsilon)+1$.

Put a 1 in front of each code to indicate that the sequence comes from $A_\varepsilon^{(n)}$.

Now order the space of all sequences. There are $|X|^n$ of these, so we can code them uniquely using binary strings of at most $\le n \log|X| + 1$.

As such, code each $(x_1, \dots, x_n) \in A_\varepsilon^{(n)}$ using this code, prefixed with a 0, to indicate that it comes from the non-typical set. We then have

$$\mathbb{E}[l(X_1, \dots, X_n)] = \mathbb{E}[l(X_1, \dots, X_n) \mid (X_1, \dots, X_n) \in A_\varepsilon^{(n)}]\mathbb{P}(A_\varepsilon^{(n)}) + \mathbb{E}[l(X_1, \dots, X_n) \mid (X_1, \dots, X_n) \in A_\varepsilon^{(n)c}]\mathbb{P}(A_\varepsilon^{(n)c})$$

$$\le n(H(X)+\varepsilon) + 2 + \varepsilon(n \log|X| + 1) \quad \text{for } n \text{ large enough}$$

$$= n(H(X) + \varepsilon(\log|X|+1)) + (2+\varepsilon)$$ ■

Def. $\forall n, \delta > 0$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be defined to be the smallest ossible set s.t. $\mathbb{P}(B_\delta^{(n)}) \ge 1-\delta$.

Thm 3.3.1 Let $(X_1, \dots, X_n)$ be i.i.d. For $\delta < \frac{1}{2}$, and any $\delta' > 0$, $\frac{1}{n} \log |B_\delta^{(n)}| > H(X) - \delta'$ for n large enough.

Pf. Since $\mathbb{P}(A_\varepsilon^{(n)}) \to 1$, for large enough n, $\mathbb{P}(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \delta - \delta'$ for any ε

As such, $|A_\varepsilon^{(n)} \cap B_\delta^{(n)}| > (1-\delta)\, 2^{n(H(x) - \delta')}$, and

$$\frac{1}{n} \log |B_\delta^{(n)}| \geq \frac{1}{n} \log |B_\delta^{(n)} \cap A_\varepsilon^{(n)}| > H(X) - \delta' + \underbrace{\frac{\log(1-\delta)}{n}}_{\to 0}$$

Mini-Course L2: Data Compression P1

Tuesday, June 7, 2016 9:40 PM

Let us continue talking about the AEP (asymptotic equipartition property).

Thm $X_1, X_2, \ldots \overset{iid}{\sim} p(x)$. Then $-\frac{1}{n}\log p(X_1,\ldots,X_n) \xrightarrow{a.s.} H(X)$

Pf. LLN.

Def. For fixed $\varepsilon > 0$, $n \in \mathbb{N}$, we call the ε-typical set, denoted $A_\varepsilon^{(n)}$ the set of all sequences $(x_1,\ldots,x_n)$ s.t. $2^{-n(H(X)+\varepsilon)} \le p(x_1,\ldots,x_n) \le 2^{-n(H(X)-\varepsilon)}$

Thm 3.1.2 ① $(x_1,\ldots,x_n) \in A_\varepsilon^{(n)} \iff 2^{-n(H(X)+\varepsilon)} \le p(x_1,\ldots,x_n) \le 2^{-n(H(X)-\varepsilon)}$

② $\forall \varepsilon$, for n large enough, $\mathbb{P}(A_\varepsilon^{(n)}) > 1-\varepsilon$

③ $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$

④ $|A_\varepsilon^{(n)}| \ge (1-\varepsilon)2^{n(H(X)-\varepsilon)}$ for n large enough

Pf. ③ $1 \ge \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} p(x_1,\ldots,x_n)$

$\ge |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)}$

④ For n large enough,

$(1-\varepsilon) \le \mathbb{P}(A_\varepsilon^{(n)}) = \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} p(x_1,\ldots,x_n) \le |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)}$ ∎

Thm 3.2.1. Let $X_1,\ldots,X_n$ be i.i.d., $\varepsilon > 0$. $\exists$ a code mapping strings of length n into binary strings s.t. the mapping is 1-1, and $\mathbb{E}[\frac{1}{n} l(X_1,\ldots,X_n)] \le H(X)+\varepsilon$ for n large enough.

Pf. The idea is to consider the typical set and the non-typical set separately.

We first order all the elements in the typical set. There are $\le 2^{n(H(X)+\varepsilon)}$ sequences, so we can code each of them uniquely using a binary string of length $\le n(H(X)+\varepsilon)+1$.

Put a 1 in front of each code to indicate that the sequence comes from $A_\varepsilon^{(n)}$.

Now order the space of all sequences. There are $|X|^n$ of these, so we can code them uniquely using binary strings of at most $\le n\log|X|+1$.

As such, code each $(x_1,\ldots,x_n) \in A_\varepsilon^{(n)}$ using this code, prefixed with a 0, to indicate that it comes from the non-typical set. We then have

$$\mathbb{E}[l(X_1,\ldots,X_n)] = \mathbb{E}[l(X_1,\ldots,X_n) \mid (X_1,\ldots,X_n)\in A_\varepsilon^{(n)}]\mathbb{P}(A_\varepsilon^{(n)}) + \mathbb{E}[l(X_1,\ldots,X_n) \mid (X_1,\ldots,X_n)\in A_\varepsilon^{(n)c}]\mathbb{P}(A_\varepsilon^{(n)c})$$

$$\le n(H(X)+\varepsilon)+2 + \varepsilon(n\log|X|+1) \quad \text{for } n \text{ large enough}$$

$$= n(H(X)+\varepsilon(1+\log|X|)) + (2+\varepsilon)$$ ∎

$$\leq n(H(X)+\varepsilon) + 2 + \varepsilon(n \log|\mathcal{X}| + 1) \quad \text{for } n \text{ large enough}$$

$$= n(H(X) + \varepsilon(\log|\mathcal{X}|+1)) + (2+\varepsilon) \quad \blacksquare$$

Def. $\forall n, \delta > 0$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be defined to be the smallest ossible set s.t. $\mathbb{P}(B_\delta^{(n)}) \geq 1-\delta$.

Thm 3.3.1 Let $(X_1, \dots, X_n)$ be i.i.d. For $\delta < \frac{1}{2}$, and any $\delta' > 0$, $\frac{1}{n} \log |B_\delta^{(n)}| > H(X) - \delta'$ for n large enough.

Pf. Since $\mathbb{P}(A_\varepsilon^{(n)}) \longrightarrow 1$, for large enough n, $\mathbb{P}(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \geq 1-\delta-\delta'$ for any ε

As such, $|A_\varepsilon^{(n)} \cap B_\delta^{(n)}| > (1-\delta) 2^{n(H(X)-\delta')}$, and

$$\frac{1}{n} \log |B_\delta^{(n)}| \geq \frac{1}{n} \log |B_\delta^{(n)} \cap A_\varepsilon^{(n)}| > H(X) - \delta' + \underbrace{\frac{\log(1-\delta)}{n}}_{\to 0}$$

Q. Have we answered the data compression question definitively?

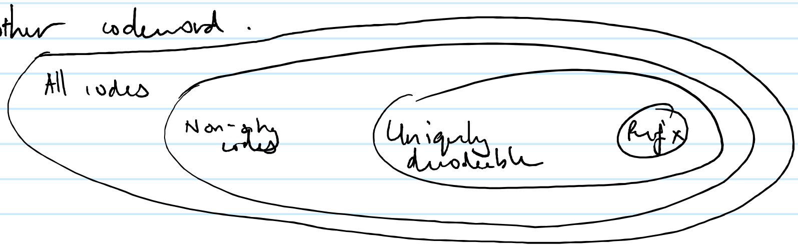
No. How do we know $nH(X)$ is the optimal length? Also, result only holds asymptotically

Def A source code for a RV X is a mapping from $\mathcal{X}$, the range of X to D^*, the set of finite length strings of symbols from a D-ary alphabet.

Denote $C(x)$ = codeword for outcome $x \in \mathcal{X}$, $l(x)$ = length of $C(x)$.

Every code C extends to finite-length strings from $\mathcal{X}$ via semigroup homomorphism.

Def A code is nonsingular if it is injective on $\mathcal{X}$. It is called uniquely decodable if the extension to strings is nonsingular. It is called a prefix code if no codeword is a prefix of another codeword.



E.g. X	Singular	Nonsingular	UD	Prefix
0	0	0	10	0
1	0	010	00	10
2	0	01	11	110
3	0	10	110	111

Thm 5.2.1 (Kraft inequality) For any instantaneous code, $\sum D^{-l_i} \leq 1$

Pf. Consider a D-ary tree. Each codeword of length l_i corresponds to a leaf at depth l_i.

Thm 5.2.2 (Extended Kraft inequality) Also holds for countably infinite set of codewords.

Thm 5.3.1 Also holds for any uniquely decodable code

Pf. First assume that $\mathcal{X}$ is finite. Then $\forall k$,

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{\vec{x}\in\mathcal{X}^k} D^{-l(\vec{x})}$$

$$= \sum_{m=1}^{k l_{max}} a(m) D^{-m}$$, where m $a(m)$ = no. of codewords of length m

$$\le \sum_{m=1}^{k l_{max}} D^{m} D^{-m} = k l_{max}$$

$$\Rightarrow \sum_{x\in\mathcal{X}} D^{-l(x)} \le k^{\frac{1}{k}} l_{max}^{\frac{1}{k}} \longrightarrow 1$$

When $\mathcal{X}$ is infinite, take a limit. □

Let us now try to obtain lower bounds. To do this, we first need to define relative entropy

Def. Let p, q be two distributions on $\mathcal{X}$. Then the relative entropy / KL divergence of q from p is $D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$

Lem $D(p\|q)\ge 0 \ \forall p, q$, $D(p\|q) = 0 \iff p = q$

Pf. $-D(p\|q) = \sum p(x)\log\frac{q(x)}{p(x)} \overset{\text{Jensen's}}{\le} \log\left(\sum p(x)\frac{q(x)}{p(x)}\right) \le \log(1) = 0$ □

Rem $D(p\|q)$ is not symmetric. Nonetheless, it is the right information-theoretic "metric" on the space of distributions.

Next time, we will explain how relative entropy can be used to give a lower bound.

Mini-Course L3: Data Compression P2

Tuesday, June 7, 2016 12:21 PM

Recall:

Def. Entropy of a RV X is defined by $H(X) = \sum p(x) \log_2 \frac{1}{p(x)}$

Relative entropy of a RV X is defined by $D(p\|q) = \sum p(x) \log_2 \frac{p(x)}{q(x)}$

Notn: Length of a codeword for x: $\ell(x)$

Fact $D(p\|q) \geq 0$ w/ equality ⟺ $p = q$

Thm Given a RV X taking values in a finite alphabet $\mathcal{X}$, $\forall \varepsilon > 0$, for n large enough, we can construct a code for the sequence $(X_1, \dots, X_n)$ such that

$$E\left[\frac{1}{n} \ell(X_1, \dots, X_n)\right] \leq H(X) + \varepsilon$$

Hence, $H(X)$ is the average bit per symbol needed to encode data from X.

Thm (Kraft inequality) For any uniquely decodable D-ary source code, have

$$\sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1$$

Thm 5.3.1 The expected length L of any uniquely decodable code satisfies $L \geq H_D(X)$ with equality iff $p_i = D^{-\ell_i}$

Pf. Set $r_i = \frac{D^{-\ell_i}}{\sum_j D^{-\ell_j}}$, $c = \sum D^{-\ell_j}$

Then
$$\begin{aligned} L - H_D(X) &= \sum p_i \ell_i + \sum p_i \log_D p_i \\ &= -\sum p_i \log_D D^{-\ell_i} + \sum p_i \log_D p_i \\ &= -\sum p_i \log_D \frac{D^{-\ell_i}}{\sum_j D^{-\ell_j}} + \sum p_i \log_D p_i - \log_D c \\ &= \sum p_i \log_D \frac{p_i}{r_i} - \log_D c \\ &= D_p(p\|r) - \log_D c \geq 0 \end{aligned}$$

If the distribution p is D-ary, then we can achieve the bound, otherwise, the best we can do is to choose $\ell_i^* = \lceil \log_D \frac{1}{p_i} \rceil$. Letting L^* be the expected length of such a code, get

$$\boxed{H_D(X) \leq L^* < H_D(X) + 1} \quad \leftarrow$$ Entropy is average number of binary questions needed to find out the outcome of a RV.

To reduce the 1 D-it overhead, we can code strings of n symbols instead of individual symbols.

This gives an optimal per symbol length of $H(X) \le L < H(X) + \frac{1}{n}$

Thm 5.4.3 (Wrong code) The expected length under $p(x)$ of the code assignment $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ satisfies

$$H(p) + D(p\|q) \le \mathbb{E}_p l(X) < H(p) + D(p\|q) + 1$$

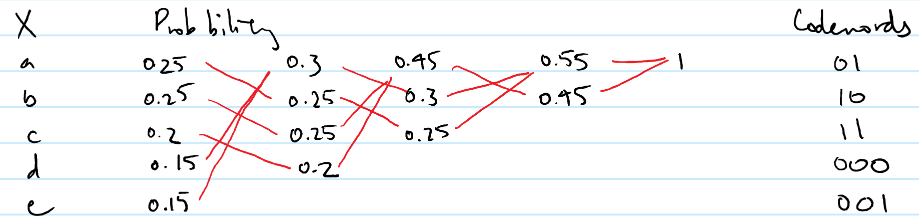
Pf. $\mathbb{E}_p l(X) = \sum p(x) \lceil \log \frac{1}{q(x)} \rceil$

$$= \sum p(x) \log \frac{1}{p(x)} + \sum p(x) \log p(x) + \sum p(x) \lceil \log \frac{1}{q(x)} \rceil \quad \square$$

This implies that $D(p\|q)$ is the "cost" of assuming the distribution is $q(x)$ when the true distribution is $p(x)$.

How do we find optimal codes? One way is to use the Huffman algorithm.

E.g.



X	Probability					Codewords
a	0.25	0.3	0.45	0.55	1	01
b	0.25	0.25	0.3	0.45		10
c	0.2	0.25	0.25			11
d	0.15	0.2				000
e	0.15					001

At each step, we combine the 2 least probable symbols

Lem 5.8.1 For any distribution, ∃ an optimal instantaneous code s.t.

① Lengths are ordered inversely with probabilities

② Two longest codewords have same length

③ The two most unlikely symbols have codewords that are siblings

Thm 5.8.1 Huffman coding is optimal, i.e. if C^* is a Huffman code, C' any other uniquely decodable code, then $L(C^*) \le L(C)$

Pf. We call codes satisfying the lemma canonical codes.

Given a prob distribution $(p_1, \dots, p_n)$, define the Huffman reduction of p to be $p' = (p_1, \dots, p_{n-2}, p_{n-1} + p_n)$, where we assume $p_1 \ge \dots \ge p_n$.

Suppose we have an optimal code C_{n-1} for p'. We can extend it to a code C_n^* for p by making the codewords for p_{n-1} and p_n siblings with the codeword in C_{n-1} for $p_{n-1} + p_n$ as their prefix.

Conversely, given an optimal canonical code C_n for p, can shrink it to a code C_{n-1}^* for p' by combining the codewords for p_{n-1} and p_n.

We have $L(C_n^*) = L(C_{n-1}) + p_{n-1} + p_n$, $L(C_{n-1}^*) = L(C_n) - p_{n-1} - p_n$, so we get $L(C_n^*) \le L(C_{n-1}^*) + p_{n-1} + p_n = L(C_n)$, which implies that C_n^* is optimal.

If C_{n-1} is a Huffman code, so is C_n^*, so we are done by induction.

We have shown that Huffman codes are optimal prefix codes. To extend to uniquely decodable codes, use the Kraft inequality. ■

Rem Huffman codes are used commercially in almost all compression schemes.

Rem (Optimality of Huffman) Huffman codes are optimal for coding symbol by symbol when the RVs are i.i.d. and the distribution is known. Whenever any of these assumptions fail, there may be other codes that do better.

① Arithmetic coding codes each string as an interval in $[0,1]$. It handles the non-i.i.d case better and also removes the 1-bit overhead. However, it is has higher complexity

② There are algorithms for adaptive Huffman coding when the distribution is not known.

Mini-Course L4: Channel Coding P1

Wednesday, June 8, 2016 10:02 PM

Recall:

Def: A discrete channel is a system comprising an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a probability transition matrix $\{p(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$. The channel is said to be memoryless if $p(y|x)$ is independent o prior inputs or outputs.

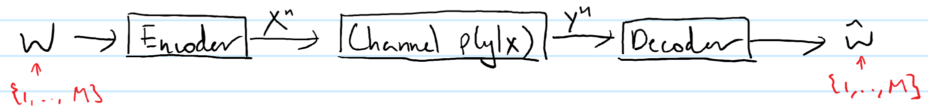
Def An (M,n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of

① An index set $\{1,..,M\}$

② An encoding function $X^n : \{1,..,M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), .., x^n(M)$. The set of all codewords is called a codebook.

③ A decoding function $g : \mathcal{Y}^n \to \{1,..,M\}$.

Diagram for how a channel is used



Def The rate R of an (M,n) code is $R = \frac{\log M}{n}$ bits per transmission.

Def Let $\lambda_i = \mathbb{P}(\hat{W} \neq W \mid W = i)$ ↝ conditional probability of error

$\lambda^{(n)} = \max_i \lambda_i$ ↝ max probability of error

$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{n} \lambda_i$ ↝ average probability of error

A rate is said to be achievable if $\exists$ a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the max prob of error $\lambda^{(n)} \to 0$ as $n \to \infty$.

Q. What is the max achievable rate? (is this even positive since there is a tradeoff between rate and error probability.

To answer this question, we need new notions of information. For data compression, we were dealing with just one RV. For channels, we have two: the input and the ouput (recall that the output is a random function of the input. As such, we need notions of information/entropy that take into account 2 or more RVs.

Recall:

① Entropy $H(X)$ is the average no of binary questions needed to determine the outcome of X.

② $D(p\|q)$ is the inefficiency of assuming the distribution is q when the true distribution is p.

Def. The joint entropy of two RVs X and Y, denoted $H(X,Y)$ is the entropy of their joint distribution

The conditional entropy of X given Y is defined as

$$H(X|Y) = \mathbb{E}\log_2\frac{1}{p(X|Y)} = \sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log_2\frac{1}{p(x|y)} = \sum p(x,y)\log_2\frac{1}{p(x|y)}$$

i.e it is the average amount of uncertainty about X, given knowledge of Y.

Claim: $H(X,Y) = H(X|Y) + H(Y)$

Pf. $H(X,Y) = -\sum p(x,y)\log_2 p(x,y) = -\sum p(x,y)\log_2 p(x|y) - \sum p(x,y)\log_2 p(y)$

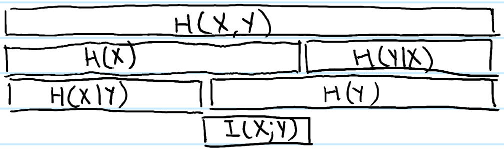
$= H(X|Y) + H(Y)$ □

Def. The mutual information of X and Y, denoted $I(X;Y)$ is one of the following equivalent quantities

① $D(p(x,y)\|p(x)p(y))$ ② $H(X) - H(X|Y)$ ③ $H(Y) - H(Y|X)$

We can think of mutual information as the amount by which the uncertainty about X is revealed on average by knowing Y.

Here's the relationships amongst the above quantities



Special cases:

① If X, Y are independent, then

- $H(X|Y) = -\sum p(x,y)\log_2 p(x|y) = -\sum p(x,y)\log_2 p(x) = H(X)$
- $I(X;Y) = 0$, $H(X,Y) = H(X) + H(Y)$ ~ no information

② If $X=Y$, then

- $H(X|Y) = -\sum p(x,y)\log_2 1 = 0$
- $I(X;Y) = H(X) = H(Y)$. ~ full information.

③ If $Y=f(X)$, then $H(Y|X) = 0 \Rightarrow I(X;Y) = H(Y) \le H(X)$

Def. The channel capacity of a discrete memoryless channel is defined as $C = \max_{p(x)} I(X;Y)$

The maximum exists because

Thm 2.7.4 Let $(X,Y) \sim p(x,y) = p(x)p(y|x)$. The mutual information is a concave function in $p(x)$ with $p(y|x)$ fixed

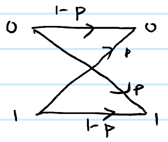
Pf. Write $I(X;Y) = H(Y) - \sum p(x) H(Y|X=x)$.

Have $p(y) = \sum_{x\in\mathcal{X}} p(x,y) = \sum_{x\in\mathcal{X}} p(y|x)p(x)$ $\left(\vec{p}_y = \vec{p}_X \cdot [p(y|x)]\right)$,

so $\vec{p}_y$ is a linear function of $\vec{p}_X$. □

E.g. ① Binary symmetric channel

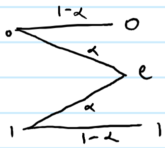
$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum p(x) H(Y|X=x) \\ &= H(Y) - H(p) \\ &\le 1 - H(p) \end{aligned}$$



with equality ⟺ Y is uniform. This happens if X is uniform.

② Binary erasure channel

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\alpha) \end{aligned}$$

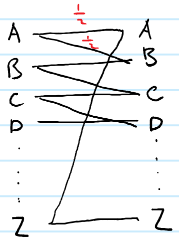


Letting $P(X=1) = p$, Let E be the event that $Y = e$. Then

$$\begin{aligned} H(Y) = H(Y,E) &= H(E) + H(Y|E) \\ &= H(\alpha) + (1-\alpha)H(X) \end{aligned}$$

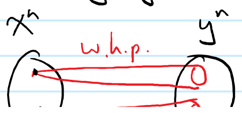
$\Rightarrow I(X;Y) \le 1-\alpha$ with equality ⟺ X is uniform

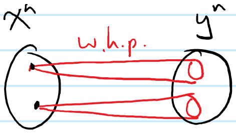
③ Noisy typewriter



$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \log_2 2 \le \log_2 26 - \log_2 2 = \log_2 13$$

The idea of the channel coding theorem is that for large enough block lengths, every channel looks like the noisy typewriter




X^n
Y^n
w.h.p.

Mini-Course L5: Channel Coding P2

Thursday, June 9, 2016 11:33 AM

Recall: Conditional entropy, mutual information, capacity. Notation λ_i, $\lambda^{(n)}$

Def The set $A_\varepsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x,y)$ is the set of sequences such that

① $\left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \varepsilon$ ② $\left|-\frac{1}{n}\log p(x) - H(X)\right| < \varepsilon$ ③ $\left|-\frac{1}{n}\log p(y) - H(Y)\right| < \varepsilon$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$

Thm 7.6.1. (Joint AEP) Let (X^n, Y^n) be sequences of length n drawn according to $\prod_{i=1}^n p(x_i, y_i)$. Then

① $P((X^n, Y^n) \in A_\varepsilon^{(n)}) \to 1$ as $n \to \infty$

② $|A_\varepsilon^{(n)}| \le 2^{n(H(X,Y)+\varepsilon)}$

③ If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then $P((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \le 2^{-n(I(X;Y)-3\varepsilon)}$, and for n large enough, $P((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \ge (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}$

Pf. ① follows from LLN

② Have $1 \ge \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n, y^n) \ge |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X,Y)+\varepsilon)}$

③ $P((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) = \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n)$

$\le 2^{n(H(X,Y)+\varepsilon)} \cdot 2^{-n(H(X)-\varepsilon)}\, 2^{-n(H(Y)-\varepsilon)}$

$= 2^{-n(I(X;Y)-3\varepsilon)}$ □

Thm 7.7.1 For every rate $R < C$, $\exists$ a sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$

Pf. Fix n, and for $i = 1, .., 2^{nR}$, choose a codeword $X^n(i)$ according to $p(x^n) = \prod_{i=1}^n p(x_i)$ to form a codebook $\mathcal{C}$, where $p(x)$ achieves $C = \max_p I(X;Y)$

Next, choose a message $W \in [2^{nR}]$ uniformly, and send it over the channel. The receiver receives a sequence Y^n according to $p(y^n | x^n(w))$, and declares that the message $\hat{W}$ has been sent if ① $(X^n(\hat{W}), Y^n)$ are jointly typical

② No other codeword $X^n(i)$ is jointly typical with Y^n.

If either of these fails, we declare an error. We let E denote this event.

We have

$$P(E) = \sum_{i=1}^{2^{nR}} P(E|W=i)P(W=i) = \frac{1}{2^{nR}}\sum_{i=1}^{2^{nR}} P(E|W=i) = P(E|W=1)$$

$$\mathbb{P}(E) = \sum_{i=1}^{2^{nR}} \mathbb{P}(E|W=i)\,\mathbb{P}(W=i) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \mathbb{P}(E|W=i) = \mathbb{P}(E|W=1)$$

For $i=1,\dots,2^{nR}$, define $E_i = \{(X^n(i), Y^n) \in A_\varepsilon^{(n)}\}$. Then

$$\mathbb{P}(E|W=1) = \mathbb{P}(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}})$$

$$\leq \mathbb{P}(E_1^c) + \sum_{i=2}^{2^{nR}} \mathbb{P}(E_i)$$

Have $\mathbb{P}(E_1^c) \leq \varepsilon$ for n large, while $\mathbb{P}(E_i) \leq 2^{-n(I(X;Y)-3\varepsilon)}$, so

$$\mathbb{P}(E|W=1) \leq \varepsilon + 2^{n(\underbrace{R-I(X;Y)}_{<0}-3\varepsilon)} \leq 2\varepsilon \quad \text{for } n \text{ large enough}$$

As such $\mathbb{P}(E) \leq 2\varepsilon$. Recall that this is the average prob of error, averaged over all codebooks.

We may hence find a fixed codebook c s.t. $\mathbb{P}(E|\mathcal{C}=c) \leq 2\varepsilon$. Furthermore, we have

$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}) \leq 2\varepsilon$, so at least half of the codewords satisfy $\lambda_i(\mathcal{C}) \leq 4\varepsilon$ by Markov. Restricting our codebook to these codewords, we get a $(2^{nR-1}, n)$ code that has max prob of error $\lambda^{(n)} \leq 4\varepsilon$ with rate $R' = R - \frac{1}{n}$ □

We can also prove the converse statement that rates above C are not feasible. In order to do that, we need some inequalities

<u>Def.</u> We say that X, Y, Z form a Markov chain, denoted $X \to Y \to Z$ if

$$p(z|x,y) = p(z|y)$$

<u>Rem</u> We have $W \to X^n \to Y^n \to \hat{W}$

<u>Thm</u> (Data-processing inequality) If $X \to Y \to Z$, then $I(Z;X) \leq I(Z;Y)$

Pf. $I(Z;Y) = H(Z) - H(Z|Y)$

$$= H(Z) - H(Z|X,Y)$$

$$\geq H(Z) - H(Z|X)$$

$$= I(Z;X) \quad □$$

<u>Thm</u> (Fano's inequality) $H(W|\hat{W}) \leq P_e^{(n)} nR + 1$

Pf. Let $E = \begin{cases} 0 & \text{if } W=\hat{W} \\ 1 & \text{o/w} \end{cases}$

Then $H(E,W|\hat{W}) = \overset{0}{\cancel{H(E|W,\hat{W})}} + H(W|\hat{W})$

$$= H(W|E,\hat{W}) + H(E|\hat{W})$$

$$\leq H(W|E) + H(E)$$

$$\leq P_e^{(n)} \log 2^{nR} + 1 \quad □$$

$\le H(W|E) + H(E)$

$\le P_e^{(n)} \log_2 2^{nR} + 1$ □

Thm For any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, must have $R \le C$.

Pf. Consider a sequence of $(2^{nR}, n)$-codes with $\lambda^{(n)} \to 0$.

For fixed n, let W be drawn uniformly. Then

$nR = H(W)$

$= H(W|\hat{W}) + I(W;\hat{W})$

$\le nRP_e^{(n)} + 1 + I(W;\hat{W})$

$\le nRP_e^{(n)} + I(X^n;Y^n) + 1$

$\le nRP_e^{(n)} + nC + 1$ (to be proved)

$\Rightarrow R \le C$ by taking $n \to \infty$, or $P_e^{(n)} \ge 1 - \frac{C}{R} - \frac{1}{nR}$ □

Lem (Chain rule for entropy) $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1})$

Lem 7.9.2 Let Y^n be the result of passing X^n through a discrete memoryless channel.

Then $I(X^n;Y^n) \le nC \quad \forall\, p(x^n)$.

Pf. $I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n)$

$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | X^n, Y_1, \ldots, Y_{i-1})$

$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$

$\le \sum_{i=1}^{n} H(Y_i) - H(Y_i|X)$

$= \sum_{i=1}^{n} I(X_i;Y_i) \le nC$ □