



## Shannon Information Theory


"The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point."
On the transmission of information over a noisy channel:

- An information source that produces a message
- A transmitter that operates on the message to create a signal which can be sent through a channel
- A channel, which is the medium over which the signal, carrying the information that composes the message, is sent
- A receiver, which transforms the signal back into the message intended for delivery
- A destination, which can be a person or a machine, for whom or which the message is intended


## Information theory is...

- It all starts with Probability Theory!


1. (Wikipedia) Information theory is a branch of applied mathematics, electrical engineering, and computer science involving the quantification of information.
2. Information theory is probability theory where you take logs to base 2 .

## Entropy <br> Entropy! ©



## Entropy - a measure of randomness

- Entropy

The entropy of a random variable $X$ with a probability mass function $p(x)$ is defined by

$$
\begin{equation*}
H(X)=-\sum_{x} p(x) \log _{2} p(x) \tag{1.1}
\end{equation*}
$$

Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$. We can calculate the entropy of the horse race as

$$
\begin{align*}
H(X) & =-\frac{1}{2} \log \frac{1}{2}-\frac{1}{4} \log \frac{1}{4}-\frac{1}{8} \log \frac{1}{8}-\frac{1}{16} \log \frac{1}{16}-4 \frac{1}{64} \log \frac{1}{64} \\
& =2 \text { bits. } \tag{1.3}
\end{align*}
$$

## Entropy

- Consider the Random Variable $X$, the outcome of each consecutive horse race.
- Assigning a binary number to each horse in the race would require $8=2^{k} \rightarrow k=3$ bits/symbol
- Shannon states that some encoding scheme on the sequence $X_{1}, X_{2} \ldots X_{n}$ can reduce the average number of bits -> $\mathrm{H}(X)=$ 2 bits/symbol
- We will show you how soon


## Connection to Physics

- A macrostate is a description of a system by large-scale quantities such as pressure, temperature, volume.
- A macrostate could correspond to many different microstates $i$, with probability $p_{i}$.
- Entropy of a macrostate is
- $S=-k_{B} \sum_{i} p_{i} \ln p_{i}$


## Entropy: The bent coin

Example 2.1.1 Let

$$
X= \begin{cases}1 & \text { with probability } p,  \tag{2.4}\\ 0 & \text { with probability } 1-p .\end{cases}
$$

Then

$$
\begin{equation*}
H(X)=-p \log p-(1-p) \log (1-p) \xlongequal{\text { def }} H(p) . \tag{2.5}
\end{equation*}
$$

- Consider $X$ to be the result of a coin toss.
- This coin has been modified to land a Head (1) with probability p and Tails (0) with probability 1-p.
- What is the entropy of $X$ as we vary p ?

Plot a graph of $\mathbf{H}(\mathrm{p})$ against p .

## Entropy



FIGURE 2.1. $H(p)$ vs. $p$.

## Bent/Unfair Coin

- The most unpredictable coin is the fair coin.
- More generally for a variable of k states, a uniform distribution across all k states exhibits maximum entropy
- The observation of a coin with two tails is not random at all.


## Example: Mystery Text

I. Emma Woodh*use, hands*me, clever* and rich,*with a comiortab*e home an* happy di*position,*seemed to*unite som* of the b*st bless*ngs of e*istence;*and had *ived nea*ly twenty *ne year* in the*world w*th very*little *o distr*ss or vex*her. *he was*the yo*ngest *f the *wo dau*hters *f a most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc* of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a ver* ea*ly

$t \star a n * a n * i n * i s * i n * t$ *em*mb*an*e *f *er*ca*es*es* $a * d * h * r * p * a * e * h * d * b * e *$ *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a* $g * * e * * e * *, * * h * * h * *$
**l**n**i**l**s**r**o**a**o**e**i*
$a * * * \mathrm{C} * * * \mathrm{n} * * * \mathrm{~S} * * * \mathrm{e} * * * \mathrm{y} * * * \mathrm{~S} * * * \mathrm{~d} * * * \mathrm{~s} * * * \mathrm{a} * * * \mathrm{r} * * * \mathrm{e} * * * \mathrm{n} * * *$

II. Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty one years in the world with very little to distress or Vex her. She was the youngest of the two daughters of a most affectionate, indolent father; and had , in consequence of her sister's marriage , been mistress of his house from a very early period. Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses; and her place had been supplied by an excellent woman as governess, who had fallen little short of a mother in affection. Sixteen years had Miss Taylor been in Mr Woodhouse's family, less as a eoverness than a friend , very

Entropy of English - 2.6 Bits/letter || Raw encoding 27+4 = 31sym = 5.1 Bits/letter
Source: MacKay VideoLectures 02, Slide 5
(13) ELEC 3004: Systems

## Source coding

## Data Compression/Source coding

- Removes redundancy to reduce bit length of messages
- Save bits on common symbols/sequences
- Spend extra bits on the 'surprises', as they very rarely occur
- Lossless (Huffman, Algorithmic, LZ, DEFLATE)
- Lossy (JPEG, MP3, H. 265 etc)
- Lossy techniques exploit perceptual dynamic range to expend bits where a human is sensitive, and save where they're not

Source CODING


## Huffman Coding

- Huffman is the simplest entropy coding scheme
- It achieves average code lengths no more than $1 \mathrm{bit} /$ symbol of the entropy
- A binary tree is built by combining the two symbols with lowest probability into a dummy node
- The code length for each symbol is the number of branches between the root and respective leaf



## Ex: Bent/Unfair Coin

- Form symbols TT, HT, TH, HH
- $\mathrm{P}(\mathrm{TT})=0.81 ; \mathrm{P}(\mathrm{HT})=\mathrm{P}(\mathrm{TH})=0.09 ; \mathrm{P}(\mathrm{HH})=0.01$
- $X_{T T}=0\left|p=0.81 \quad X_{T H}=10\right| p=0.09$
$X_{H T}=110\left|p=0.09 \quad X_{H H}=111\right| p=0.01$
- Transmission cost $0.81 \cdot 1+0.09 \cdot 2+0.09 \cdot 3+0.01 \cdot 3=$ $1.29 \frac{\text { bits }}{\text { symbol }}$
- $H(X)=0.4690 \cdot 2=0.938 \frac{\text { Bits }}{\text { symbol }}-$ non-optimal code
- Huffman transmission cost $\mathrm{H}(\mathrm{X}) \leq \mathrm{C}<H(X)+1 \frac{\text { bit }}{\text { symbol }}$
- longer code words $=$ better performance



## Lossless compression and entropy

- The lower the entropy of the source, the more we can gain via compression.
- The fair coin exhibits no exploitable means of compressing a sequence of observations from their direct description, without losing some information in the process
- All forms of compression exploit structure to reduce the average number of bits to describe a sequence.


## Channel Coding



## Channel coding

- All physical transmission is susceptible to noise.
- Noise results in errors when a transmission is converted back to the source space
- Channel coding introduces redundancy to reduce the probability an error is made in conversion back to the source.


## Noisy channels: Binary Symmetric Channel

$$
\begin{array}{ll}
x \vec{Z}_{0}^{0} y & \begin{array}{l}
P(y=0 \mid x=0)=1-f ;
\end{array} \quad \begin{array}{l}
P(y=0 \mid x=1)=f ; \\
\\
P(y=1 \mid x=0)=f ;
\end{array} \quad P(y=1 \mid x=1)=1-f .
\end{array}
$$

- A 1 bit digital channel where each bit experiences a probability, $f$, of being misinterpreted (flipped)
- Ex: A File of $\mathrm{N}=10,000$ bits is transmitted with $\boldsymbol{f}=\mathbf{0 . 1}$



## How many bits are flipped?

Assume a Binomial distribution

- Mean: $\mu=N p$
- Variance: $\sigma^{2}=N p q$

Then:

$$
\begin{gathered}
\mu=(0.1) 10,000 \\
\sigma^{2}=(0.1)(0.9) 10,000=900
\end{gathered}
$$



Thus:
$1000 \pm 30$




## $P_{\text {error }}$ of the $R_{3}$ code

- $P\left(r_{000}, s_{1}\right)=P\left(r_{111}, s_{0}\right)=0.1^{3}=0.001$
- $P\left(r_{100}, s_{1}\right)=P\left(r_{010}, s_{1}\right) \ldots=0.1^{2}=0.01$
- $P_{\text {err }}=(0.001+0.01 \cdot 3) \cdot p\left(s_{1}\right)+(0.001+0.01 \cdot 3) \cdot p\left(s_{0}\right)$
- $P_{\text {err }}=0.031$
- The number of errors $\mu=310 \pm 17.3$
- Approximately 3 x reduction in errors, but 3 x as many bits required to send the message


## Performance of Repetition Codes



- Want to maximise rate, minimise Error
- Software (e.g. Checksums) can identify errors and correct to higher stringency
- Can we do better?


## The Noisy-channel coding theorem

- The channel capacity for the general channel:

- Is the mutual information between X and Y :

$$
\text { Capacity }=I(X ; Y)
$$

- We will show Capacity of the BSC is:

$$
\text { Capacity }_{B S C}=I\left(X_{B S C} ; Y_{B S C}\right)=1-H\left(P_{B S C}\right)
$$

- Where H is the Shannon entropy of the bent coin with $P_{B S C}$


## Conditional Entropy

- Suppose Alice wants to tell Bob the value of X
- And they both know the value of a second variable Y.
- Now the optimal code depends on the conditional distribution $p(X \mid Y)$
- Code length for $X=i$ has length $-\log _{2} p(X=i \mid Y)$
- Conditional entropy measures average code length when they know Y

$$
H(X \mid Y)=-\sum_{X, Y} p(X, Y) \log _{2} p(X \mid Y)
$$

## Mutual information

- How many bits do Alice and Bob save when they both know Y?

$$
\begin{gathered}
I(X ; Y)=H(X)-\mathrm{H}(\mathrm{X} \mid \mathrm{Y}) \\
=\sum_{X, Y} p(X, Y)\left(-\log _{2} p(X)+\log _{2} p(X \mid Y)\right) \\
=\sum_{X, Y} p(X, Y) \log _{2}\left(\frac{p(X, Y)}{p(X) p(Y)}\right)
\end{gathered}
$$

- Symmetrical in X and Y !
- Amount saved in transmitting X if you know Y equals amount saved transmitting Y if you know X .


## Properties of Mutual Information

$I(X ; Y)=H(X)-H(X \mid Y)$
$=H(Y)-H(Y \mid X)$
$=H(X)+H(Y)-H(X, Y)$

- If $X=Y, I(X ; Y)=H(X)=H(Y)$

- If $X$ and $Y$ are independent, $I(X ; Y)=0$
$H(X, Y)=H(X)+H(Y \mid X)=H(Y)+H(X \mid Y) H(X \mid Y) \leq H(X)$
- In straightforward terms, information is never negative, we are always at least as certain about $\mathrm{X} \mid \mathrm{Y}$ vs X


## Conditional Entropy \& Mutual Information

- Note that the distributions $p(X), p(Y)$ and $p(X \mid Y)$ may be complex.
- computing $p(X \mid Y)$ may be intractable
- Mutual Information and Conditional entropy hold for all distributions, but quantifiying them may be intractable



## Mutual information of the $B S C_{p=0.1}$

$$
\begin{gathered}
P\left(X_{0}, Y_{0}\right)=P\left(X_{1}, Y_{1}\right)=0.5 * 0.9=0.45 \\
P\left(X_{0}, Y_{1}\right)=P\left(X_{1}, Y_{0}\right)=0.5 * 0.1=0.05 \\
P\left(X_{0}\right)=P\left(X_{1}\right)=P\left(Y_{0}\right)=P\left(Y_{1}\right)=0.5 \\
I(X ; Y)=\sum_{X, Y} p(X, Y) \log _{2}\left(\frac{p(X, Y)}{p(X) p(Y)}\right) \\
I(Y ; Y)=2 \cdot 0.45 \log _{2}\left(\frac{0.45}{0.5^{2}}\right)+2 \cdot 0.05 \log _{2}\left(\frac{0.05}{0.5^{2}}\right) \\
I(Y ; Y)=0.7632-0.2322=0.531
\end{gathered}
$$

## Ultimate error free performance

- $\mathrm{H}(0.1)=0.4690$
- Capacity $=0.531$ Bits per Symbol
- Theoretically can obtain a $p_{b}=0$ at that rate
- Does not tell us how to obtain that rate




## What's Practically Achievable?



-
Interative forward error correction (FEC) codes

- Shannon told us the limit, reaching it is not so easy


## Source: MacKay VideoLectures 01, Slide 54



## Break ()

## Communications

## Physical layer - Modulation Schemes

- The final piece to transmitting over real communication channels
- Analog modulation (AM (SSB) ,FM , PM, QAM)
- Digital modulation(ASK,PSK,FSK,QAM)
- Both encode the signal on the amplitude (A) or phase (F,P,Q) of a carrier sine frequency
- Will only cover digital Quadrature Amplitude Modulation (QAM)


## QAM

- A pulse of the carrier frequency encompasses a symbol.
- The symbol can take on a number of states, encoded as finite levels of amplitude and phase (quadrature)
- Here we need to window our sine carrier to form a pulse
- A rectangular window would produce spectral leakage, and heavy interference - use a raised cosine window
- A RCW results in no inter-symbol-interference



## Constellations

- For QAM 16-16 discrete locations on the inphase / quadrature plane
- Red lines denote decision boundary



## Doubling data rate - QAM4 to QAMI 6

- Bits per symbol of QAM4 $=2$, QAM16 $=4$
- Constellation diagrams
- Assuming receive power of $\frac{E b}{N o}=10 d B$, what is the probability that a message sent crosses the descision boundary? $-\frac{E b}{N o}$ is the energy per bit received, over the average noise energy at the receiver.

| QAM4 v QAMI6 |  |  |
| :---: | :---: | :---: |
|  | $\begin{aligned} & \text { Perr }_{4}=8 \cdot 10^{-6} \\ & \text { BER }=4 \cdot 10^{-6} \\ & \text { Rate }=2 \frac{\text { Bits }}{\text { Symbol }} \end{aligned}$ |  |
| 匀 | camenmams |  |

## QAM4 v QAMI 6

- Double the data rate
- $500 \times$ more likely to flip a bit!
- Generally BER is a specification of the communication protocol
- For example Gigabit Ethernet adjusts the constellation to maintain a raw BER below $10^{-10}$ (one per 10 gigabits)
- In other words, your Ethernet connection will produce $\sim 7-8$ one bit errors per hour of 4 k Netflix streaming ( $\sim 7 \mathrm{~GB} / \mathrm{Hr}$ )
- Source coding reduces this to $10^{-14}$ (one per 100 terabits)


## QAM vs Theoretical maximum



- Using a smaller constellation restricts the channel capacity
- Using too high a constellation requires complex channel coding to exploit


## Shannon Capacity of the AWGN channel

- $C=B \log _{2}\left(1+\frac{S}{N}\right)$
- B is the bandwidth S the signal power in real terms, N noise in real terms, and C the capacity in bits.
- Upper bound on error free transmission
- By intelligently selecting QAM constellation and Channel coding schema in tandem capacity can be maximised.
- Requires some coordination between transmitter and sender


## Shannon and Weaver: Models of Communication

Three "problems" in Communication:

- The technical problem: how accurately can the message be transmitted?
- The semantic problem: how precisely is the meaning "conveyed"?

- The effectiveness problem: how effectively does the received meaning affect behaviour?



## What is the "optimal code"?

- X is a random variable
- Alice wants to tell Bob the value of X (repeatedly)
- What is the best binary code to use?
- How many bits does it take (on average) to transmit the value of X ?


## Optimal code lengths

- In the optimal code, the word for $X=i$ has length
- $\log _{2} \frac{1}{p(X=i)}=-\log _{2} p(X=i)$
- For example:

- ABAACACBAACB coded as 010001101110001110
- If code length is not an integer, transmit many letters together


## Kullback-Leibler divergence

- Measures the difference between two probability distributions
- (Mutual information was between two random variables)
- Suppose you use the wrong code for $X$. How many bits do you waste?
- $D_{K L}(p \| q)=\sum_{x} p(x)\left[\log _{2} \frac{1}{q(x)}-\log _{2} \frac{1}{p(x)}\right]$
$=\sum_{x} p(x) \log _{2} \frac{p(x)}{q(x)}$
- $D_{K L}(p \| q) \geq 0$, with equality when p and q are the same.
- $I(X ; Y)=D_{K L}(p(x, y) \| p(x) p(y))$


## Continuous variables

- $X$ uniformly distributed between 0 and 1 .
- How many bits required to encode X to given accuracy?

| Decimal places | Entropy |
| :--- | :--- |
| 1 | 3.3219 |
| 2 | 6.6439 |
| 3 | 9.9658 |
| 4 | 13.2877 |
| 5 | 16.6096 |
| Infinity | Infinity |

- Can we make any use of information theory for continuous variables?


## K-L divergence for continuous variables

- Even though entropy is infinite, K-L divergence is usually finite.
- Message lengths using optimal and non-optimal codes both tend to infinity as you have more accuracy. But their difference converges to a fixed number.

$$
\sum_{x} p(x) \log _{2} \frac{p(x)}{q(x)} \rightarrow \int p(x) \log _{2} \frac{p(x)}{q(x)} d x
$$

## Calculating the Entropy of an Image



The entropy of lena is $=7.57$ bits/pixel approx

ELEC 3004: Systems

## Huffman Coding of Lenna



Average Code Word Length $=\sum_{k=0}^{255} p_{k} l_{k}=7.59$ bits $/$ pixel

So the code length is not much greater than the entropy

## But this is not very good

- Why?
- Entropy is not the minimum average codeword length for a source with memory
- If the other pixel values are known we can predict the unknown pixel with much greater certainty and hence the effective (ie. conditional) entropy is much less.
- Entropy Rate
- The minimum average codeword length for any source.
- It is defined as

$$
H(\chi)=\lim _{n \rightarrow \infty} \frac{1}{N} H\left(X_{1}, X_{2}, \ldots, X_{n}\right)
$$

## Coding Sources with Memory

- It is very difficult to achieve codeword lengths close to the entropy rate
- In fact it is difficult to calculate the entropy rate itself $P\left(X_{1} \mid X_{2} \ldots X_{n}\right)$ is described in $\mathrm{R}^{\mathrm{n}}$ Space - for lenna $\mathrm{n}=65536$
- We looked at LZW as a practical coding algorithm
- Average codeword length tends to the entropy rate if the file is large enough
- Efficiency is improved if we use Huffman to encode the output of LZW
- LZ algorithms used in lossless compression formats (eg. .tiff, .png, .gif, .zip, .gz, .rar... )



## Differential Coding

- Key idea - code the differences in intensity.


$$
\mathrm{G}(\mathrm{x}, \mathrm{y})=\mathrm{I}(\mathrm{x}, \mathrm{y})-\mathrm{I}(\mathrm{x}-1, \mathrm{y})
$$

## Differential Coding



- The entropy is now $5.60 \mathrm{bits} /$ pixel which is much less than 7.57 bits/pixel we had before (despite having twice as many symbols)


## Entropy In General

- Entropy of a source is maximised when all signals are equiprobable and is less when a few symbols are much more probable than the others.



## Lossy Compression

- But this is still not enough compression
- Trick is to throw away data that has the least perceptual significance


Effective bit rate $=8$ bits/pixel


Effective bit rate $=1$ bit/pixel (approx)

## 6. alec 3004 systems



- Review:
- Chapter 6 of FPW
- Chapter 13 of Lathi
- Deeper Pondering??


## Final Exam Reviews

1. Wed, June 5, 2019

- EBESS - Loc: TBA
- 3-7pm???

2. Thursday June 6, 2019

- $12 \mathrm{n}-2 \mathrm{p}$
- TBA
- Some Review Notes (from Course Textbooks)
$\rightarrow$ http://robotics.itee.uq.edu.au/ ~elec3004/tutes.html


