Information Theory

↳ Information Theory: movement & transformations are constrained by mathematical laws.
- ↳ Complexity of patterns is the length of the shortest program to generate this pattern
- ↳ Ultimate Data Compression = Entropy (H)
- ↳ Ultimate Rate of Reliable Communication = Channel Capacity (C)

Foundations

Random Variables: take on values determined by their probability distributions

Probability has two meanings:

↳ ① RELATIVE FREQUENCY (frequentist/operationalist view): (1) sample random variable multiple time (2) take fraction each outcome occurs

↳ ② DEGREE-OF-BELIEF: plausability of proposition - no experiment - likelihood that a particular state might occur & event outcome can only be determined once.

Less probable an event is, more information is gained by seeing occurence.

PRODUCT RULE:
$p(A,B)$ = joint prob of A and B $= p(A|B)\,p(B) = p(B|A)\,p(A)$
If A and B are independent, $P(A|B) = p(A)$ & $P(B|A) = p(B)$

SUM RULE:
$$p(A) = \sum_B p(A,B) = \sum_B p(A|B)\,p(B)$$

BAYES THEOREM:
$$p(B|A) = \frac{p(A|B)\,p(B)}{p(A)}$$

$\boxed{\text{Prior Prob}}$ = prob of observing something before any data is collected

$\boxed{\text{Posterior Prob}}$ - computed after observation of data.

Entropy

both measured in bits

Information Measure (I) = $\boxed{\log_2 p}$ — p is probability of event occurring

Entropy (H) = $-I$

↳ uncertainty / disorder

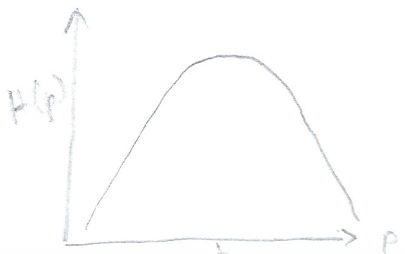
↳ log as we want information to be additive. Joint (independent) probs are multiplicative, so logs makes them additive

$= 1.443 \log_e p = 3.322 \log_{10} p$

Assymptotic Equipartition Theorem

Given uniform probabilities, and question of choosing one of N items, entropy = $\log_2 N$

$$H = -I = -\sum_i p_i \log(p_i)$$



② <u>Notation</u>

X, Y - random variables

↳ $x \in \{a_1, a_2, \ldots, a_i\} = \mathcal{A}$

$\underbrace{y}_{\text{instances}} \in \{b_1, b_2, \ldots, b_k\} = \mathcal{B}$

Ensemble is a random variable. Joint Ensemble is an ensemble whose outcomes are ordered pairs x, y.

↳ XY defines $P(x, y)$ over all outcomes

MARGINAL PROBABILITIES

$$p(x = a_i) = \sum_y p(x = a_i, y)$$

$$p(x) = \sum_y p(x, y)$$

CONDITIONAL PROBABILITIES

$$p(x = a_i \mid y = b_j) = \frac{p(x = a_i, y = b_j)}{p(y = b_j)}$$

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

JOINT ENTROPY

$$\boxed{H(X, Y) = \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)}}$$

If X and Y are independent random variables

$$H(X, Y) = H(X) + H(Y)$$

<u>Conditional Entropy</u>

$$H(X \mid y = b_j) = \sum_x p(x \mid y = b_j) \log \frac{1}{p(x \mid y = b_j)}$$

$$H(X \mid Y) = \sum_y p(y) \left[\sum_x p(x \mid y) \log \frac{1}{p(x \mid y)} \right]$$

<u>Chain Rule for Entropy</u>

JOINT ENTROPY $(H(X, Y))$ = MARGINAL ENTROPY $(H(X))$ + CONDITIONAL ENTROPY $(H(Y \mid X))$

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

~~$H(X_1, X_2, \ldots, X_n) = \sum H(X_i)$~~

$H(X)$ maximised when $p_i = 1/N$

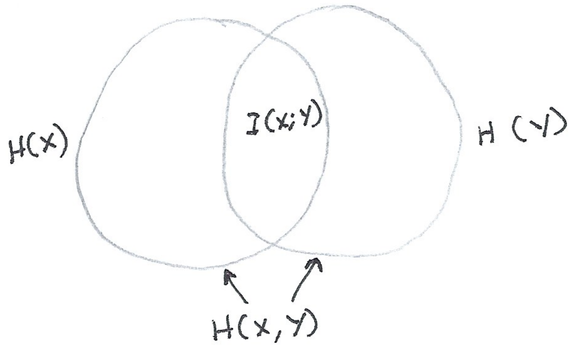
$$H(X) = -\sum_i p_i \log_2 p_i = -\sum_1^N \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N$$

(3)

$I(X;Y) =$ Mutual Information between X and Y $= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \geq 0$

↳ if perfectly correlated, $I(X;Y) = H(X) = H(Y)$

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



if we assume uniform distribution it is called a Flat Prior model.

Cross Entropy

↳ two different distributions ($p(x)$ and $q(x)$) over the same set of outcomes for random variable X.

$H(p,q) = -\sum_x p(x) \log(q(x))$

Used in coding theory to get cost of wrong representation. i.e. if coding scheme designed under assumption $q(x)$, $H(p,q)$ gets length of codewords on average given actual $p(x)$.

(1) Assymetric and (2) minimised if $p(x) = q(x)$

Distance (between two Random Variables)

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

↳ follows axioms → (1) $D(X,Y) \geq 0$
(2) $D(X,X) = 0$
(3) $D(X,Y) = D(Y,X)$
(4) $D(X,Z) \leq D(X,Y) + D(Y,Z)$ — triangle inequality

Kullback-Leibler Distance (Relative Entropy) → measure of inefficiency - information for discrimination

but if prob dist has any val $= 0$ does not work

$D_{KL}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$, $p(x) = q(x) \Rightarrow D_{KL} = 0$
$= H(p,q) - H(p)$

Given optimal code for distribution $p(x)$ (needs $H(p)$ bits) then n (additional bits) needed if described $p(x)$ using optimal code as $q(x)$ $= D_{KL}(p\|q)$

Fano's Inequality

Relates P_e (prob of error) in guessing X from knowledge of Y to conditional entropy $H(X|Y)$ when $|\mathcal{X}|$ is number of possible outcomes
↳ length of symbol alphabet

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}$$

Data Processing Inequality

If X, Y, Z form Markov Chain (Z depends on Y and Y depends on X)
↳ $X \rightarrow Y \rightarrow Z$

$$\Rightarrow I(X;Y) \geq I(X;Z)$$

$\boxed{\text{Source Coding Theorem}}$

Markov Process: source of symbols - letters emitted with known probabilities for each state.

Entropy of Markov Process expressed as bits per symbol - if emitted at known rate can get into bits per second. (baud rate)
↳ calculate entropy for each state then take weighted average given by occupancy probabilities P_i

$$H = \sum_i P_i H_i = -\sum_i P_i \sum_j p_i(j) \log(p_i(j))$$

Fixed Length Codes

Given set of N symbols, with entropy H, need fixed length block, $R = \lfloor \log_2(N) \rfloor + 1$. Code rate is R bits per symbol. $(H \leq R)$
↳ aim is $H = R$

Efficiency $\eta = \frac{H}{R}$

Variable Length Codes

Can achieve more compressed using variable-length codes.

Features of codes:

↳ ① Uniquely Decodable: codes cannot be multiple combinations
↳ ② Prefix: No codeword is the prefix of a longer codeword.

noiseless coding theorem

Shannon Source Coding Theorem: $R \geq H$

"For a discrete source with entropy H, for any $\varepsilon > 0$, possible to encode symbols into uniquely decodable code at average rate R st: $R = H + \varepsilon$ as an asymptotic limit"

5

Huffman Codes

Algorithm to get ~~any given~~ an optimal prefix code for given probability distribution of symbols → illustration of Shannon Source-Coding Theorem

- Assign bits in reverse sequence corresponding to increasing symbol prob
- More probable symbols encoded with shorter codewords.

1. Find two symbols with lowest probs + assign bit to distinguish them
   ↳ defines branch in binary tree
2. Combine these two into virtual node with prob as sum
3. Start again and go until one symbol node

N.B. no unique Huffman code for a symbol alphabet ⇒ Huffman code is as efficient as possible

Kraft-McMillan Inequality

Any instantaneous code (with prefix property) must have condition: necessary but not sufficient

↳ if N codewords have length $c_1 \leq c_2 \leq \dots \leq c_n$ then:

$$\sum_{i=1}^{N} \frac{1}{2^{c_i}} \leq 1$$

DISCRETE CHANNEL CAPACITY

Input alphabet: $X = \{x_1, \dots, x_n\}$

Output alphabet: $Y = \{y_1, \dots, y_m\}$

$$\sum_{k=1}^{K} p(y_k | x_j) = 1$$

Channel represented as transition probabilities $p(y_k | x_j)$ — these can form the channel matrix

$$p(x_j, y_k) = p(y_k | x_j)\, p(x_j)$$] joint prob. dist.

$$p(y_k) = \sum_{j=1}^{J} p(x_j, y_k) = \sum_{j=1}^{J} p(y_k | x_j)\, p(x_j)$$] marginal prob for output symbol y_k

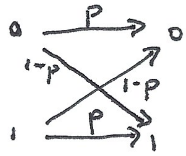
Av prob of symbol error [$$P_e = \sum_{j=1}^{J} \sum_{k=1, k \neq j}^{K} p(y_k | x_j)\, p(x_j)$$

↳ $1 - P_e$ is prob of correction reception

6

Binary Symmetric Channel

Two input and output symbols $\{0,1\}$ with channel matrix $\begin{pmatrix}1-p & p\\ p & 1-p\end{pmatrix}$



$H(X) = 1$ bit , $H(Y) = 1$ bit

$$H(X|Y) = -\sum_{x,y} p(x,y)\log p(x|y)$$

$$= -p\log(p) - (1-p)\log(1-p)$$

$$I(X;Y) = H(X) - H(X|Y)$$

$$= 1 + p\log p + (1-p)\log(1-p)$$

(binary) assuming symmetric channel
↳ flips equally likely

Channel Capacity (C) = $\max_{\{p(x_i)\}} I(X;Y)$ → equiprobable symbols maximise $I(X;Y)$

$I(X;Y)$ optimal with $p=0$ or 1 ⇒ channel capacity is $\underbrace{1\text{ bit}}_{\text{mutual information}}$ per transmitted bit

Maximal ~~for~~ equiprobable case for probability distribution.

Overcoming noise

① Repetition Codes

↳ if we transmit every symbol $N=2m+1$ times + do majority voting.

↳ require if $m+1$ bits in error

$$P_e = \sum_{i=m+1}^{2m+1} \binom{2m+1}{i} p^i (1-p)^{2m+1-i}$$

Shannon Second Theorem - Channel Coding Theorem

"For channel of capacity C and symbol source of entropy H provided $H \le C$ $\exists$ coding scheme st source is reliably transmitted through channel with residual error rate $< \epsilon$ (arbitrarily small)" - existence proof, no algorithm

Ex Given 7 bit blocks, no flip and one flip in each is equiprobable

$(b_1, b_2, \dots = \bar{b}_1, b_2, \dots = b_1, \bar{b}_2, \dots)$

$$C_s = \max_{p(x_j)} I(X;Y) = \max(H(Y) - \cancel{H(X|Y)}\ H(Y|X))$$

$H(Y) = 7$ because 2^7 equiprobable symbols

$$p(y_k|x_j) = \frac{1}{8}$$

~~p(x_j) = 2~~ 8x

channel capacity per bit

$$C_b = \frac{1}{7}\left(7 - \sum_j \sum_k p(y_k|x_j)\log\frac{1}{p(y_k|x_j)}\, p(x_j)\right)$$

$$= \frac{1}{7}\left(7 + \sum_j 8\left(\frac{1}{8}\log\frac{1}{8}\right)\frac{1}{N}\right)$$

$= \frac{4}{7}$ bits per bit

7

Syndromes

Given source entropy $H \leq C = \frac{4}{7}$ bit per bit. Construct new 7-bit codewords with 4 bits of symbol encoding and 3 error correction bits.

Calculate 3 bits as XOR of 3 of the four bits

b_1, b_2, b_4 are error correction bits: $b_4 = b_5 \oplus b_6 \oplus b_7$

$b_2 = b_3 \oplus b_6 \oplus b_7$

$b_1 = b_3 \oplus b_5 \oplus b_6$

On receipt, calculate syndromes: $s_4 = b_4 \oplus b_5 \oplus b_6 \oplus b_7$

$s_2 = b_2 \oplus b_3 \oplus b_6 \oplus b_7$

$s_1 = b_1 \oplus b_3 \oplus b_5 \oplus b_7$

if all 0, no error

$H =$

Hence, source entropy goes to $\frac{4}{7}$ bit per bit of data ⇒ satisfies requirement that $H \leq C$

Hamming codes perfect as use m bits to correct $2^m - 1$ error patterns and transmit $2^m - 1 - m$ useful bits.

$$P_e = \sum_{i=2}^{7} \binom{7}{i} p^i (1-p)^{7-i}$$

Information in Vector Spaces

Example $\underline{v}$ is vector of data or samples and $\{e_1, \ldots, e_n\}$ are basis vectors of orthonormal system. Projection of $\underline{v}$ into that space gets coefficients a_i as inner product of v with each basis vector.

inner product

$$v = \sum_{i=1}^{n} a_i e_i = \sum_{i=1}^{n} \langle v, e_i \rangle e_i$$

Example $f(x)$ represented as linear combination of functions $\varphi_i(x)$ s.t.:

$$f(x) = \sum_{i=1}^{n} a_i \varphi_i(x)$$

where: $a_i = \langle f, \varphi_i \rangle = \int_{-\infty}^{\infty} f(x) \varphi_i(x)\, dx$

↳ could be Fourier coefficient

↳ inner product

Frequency is the Fourier Transform

Propagation Function: inner product between vector of inputs and vector of learned weights.

⑧

Definitions

Vector Space V is a linear combination of vectors in V if $\exists$ coefficients to make that be true

$\boxed{v_1, \dots, v_n}$ only if linearly independent.

vectors are ~~a spac~~ basis for ~~vectors~~ space n is dimension of the space.

$\boxed{\text{span}}\{v_1, \dots v_n\} = \{u \in V : u \text{ is linear comb of } v_1, \dots, v_n\}$

↳ everything that can be represented by linear combination of span vectors

Subset $W \subset V$ is linear subspace of $V \Rightarrow$ involves projecting onto that subspace of vectors - dimensionality reduction

For V, $v_1, \dots, v_n \in V$ are $\boxed{\text{linearly independent}}$ if for scalars $a_1, a_2, \dots$

$$a_1 v_1 + a_2 v_2 + \dots + a_n v_n = 0 \text{ ; then } a_1 = a_2 = \dots = a_n = 0$$

Inner Product

$\langle u, v\rangle$ is scalar value satisfying

↳ ① $\forall v \in V$, $\langle v, v\rangle \in \mathbb{R} \geq 0$ ~~$\Leftrightarrow \langle u, v\rangle$~~

↳ ② $\forall \underline{v} \in V$, $\langle \underline{v}, \underline{v}\rangle = 0 \Leftrightarrow \underline{v} = \underline{0}$

↳ ③ $\forall \underline{u}, \underline{v}, \underline{w} \in V$, scalars a, b; $\langle a\underline{u} + b\underline{v}, \underline{w}\rangle = a\langle \underline{u}, \underline{w}\rangle + b\langle \underline{v}, \underline{w}\rangle$

↳ ④ $\forall \underline{u}, \underline{v} \in V \Rightarrow \langle \underline{u}, \underline{v}\rangle = \overline{\langle \underline{v}, \underline{u}\rangle}$

↳ if vectors are real,

$\langle u, v\rangle = \langle v, u\rangle$

$\overline{a}$ is complex conjugate

↳ Properties

↳ ① $\forall v \in V$ and scalar $a \Rightarrow \langle av, av\rangle = |a^2| \langle v, v\rangle$

↳ ② $\forall v \in V$ $\langle \underline{0}, \underline{v}\rangle = 0$

↳ ③ $\forall v \in V$ $u_1, u_2, \dots, u_n \in V$ scalars $a_1, \dots, a_n \Rightarrow$

$$\left\langle \sum_{i=1}^{n} a_i u_i, v \right\rangle = \sum_{i=1}^{n} a_i \langle u_i, v\rangle$$

$$\left\langle v, \sum_{i=1}^{n} a_i u_i \right\rangle = \sum_{i=1}^{n} \overline{a_i} \langle v, u_i\rangle$$

↳ Examples

↳ Euclidean Space $\mathbb{R}^n$: $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$

↳ Complex Vectors $\mathbb{C}^n$: $\langle x, y\rangle = \sum_{i=1}^{n} x_i \overline{y_i}$

Norm

Norm on Vector Space V: $V \to \mathbb{R}_+$ has properties

↳ ① $\forall v \in V$, $\|v\| \geq 0$

↳ ② $\|v\| = 0 \Leftrightarrow v = \underline{0}$

↳ ③ $\forall v \in V$. scalar a, $\|av\| = |a| \, \|v\|$

↳ ④ $\forall u, v \in V$ $\|u + v\| = \|u\| + \|v\|$

generalisation of the notion of the distance between two vectors

(9)

Orthogonal & Orthonormal Systems

$u, v \in V$ are **orthogonal** $(u \perp v)$ if $\langle u, v \rangle = 0$. Sequence of vectors (finite or infinite) $\{v_i\}$ is orthogonal system if:

↳ ① $v_i \neq \bar{0} \quad \forall v_i$

↳ ② $v_i \perp v_j \quad \forall i \neq j$

Orthogonal system is **orthonormal** if $\|v_i\| = 1 \quad \forall i$

↳ all unit vectors (e_i)

If $\{e_1, e_2, \dots, e_n\}$ is an orthonormal system in V.

$$u = \sum_{i=1}^{n} a_i e_i \Rightarrow a_i = \langle u, e_i \rangle$$

/ in orthonormal system, expansion coefficients are same as projection coefficients

$$u = \sum_{i=1}^{n} a_i e_i = \sum_{i=1}^{n} \langle u, e_i \rangle e_i$$

Infinite Orthonormal Systems : V with $\dim(V) = \infty$

Let $\{v_1, v_2, \dots\}$ be infinite sequence of vectors in normed vector space V.
Let $\{a_1, a_2, \dots\}$ be sequence of scalars.

$\sum_{n=1}^{\infty} a_n v_n$ converges in norm to $w \in V$ if

$$\lim_{m \to \infty} \left\| w - \sum_{n=1}^{m} a_n v_n \right\| = 0$$

w would be exactly represented by a linear combination of vectors $\{v_i\}$ in space V in limit that we could use all of them. This property of infinite orthonormal system in inner product space is a closure

Karhunen-Loeve Transform: numerical method for constructing orthonormal system such that any set of vectors can be represented with best possible accuracy using any specified finite number of terms

also Hotelling Transform, Dimensionality Reduction or Principal Components Analysis

FOURIER REPRESENTATIONS

Decomposition of functions into superpositions of elementary sinusoidal functions – consider information in piecewise continuous functions f for convenience over interval $[-\pi, \pi]$ by projecting to vector space of:

$$\{ \underbrace{\tfrac{1}{\sqrt{2}}}_{\text{average value}}, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots \}$$

↳ Fourier components

(10)

$$F(f) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx$$

if even symmetry, only cosine terms, if odd symmetry, then only sine terms

for 2π-periodic function, ie $g(x+2\pi) = g(x)$

If f even, $a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos(nx)\,dx$

$b_n = 0$

even: $f(-x) = f(x)$

odd: $f(-x) = -f(x)$

If f odd, $a_n = 0$

$b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\sin(nx)\,dx$

1. if f, g are even then fg even
2. if f, g are odd then fg even
3. if f even and g odd then fg odd
4. If g odd, for any $h>0$, $\int_{-h}^{h} g(x)\,dx = 0$
5. If g even, for any $h>0$, $\int_{-h}^{h} g(x)\,dx = 2\int_0^{h} g(x)\,dx$

N.B. If derivative is infinite, Fourier coefficient dies very quickly and convergence is faster

Complex Fourier Series

$\{1, e^{ix}, e^{-ix}, \dots\}$

N.B. $\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx$

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$ where $c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-inx}\,dx$

$\hookrightarrow$ $c_n = \frac{a_n - ib_n}{2}$

$c_{-n} = \frac{a_n + ib_n}{2}$

$\left(c_0 = \frac{a_0}{2}\right)$

Spectral Properties Of Continuous-Time Channels

Information bands assigned spectral band - carrier signal modulated inside in a parameter (frequency (1) of sine wave, amplitude (2), phase (3)). Results in a complex $f(t)$

$$f(t) = \sum_n c_n e^{i\omega_n t}$$ (by Fourier Analysis)

Channels are linearly time-invariant systems whose eigenfunctions are complex exponentials

Linear time-invariant systems obey superposition and can be described by linear $h(t)$

(11)

Complex exp is never changed, only multiplied by complex constant α

Can use amplitude and phase changing α_n for ω_n can incorporate into $f(t)$ to understand how $f(t)$ is affected by transmission through channel, hence:

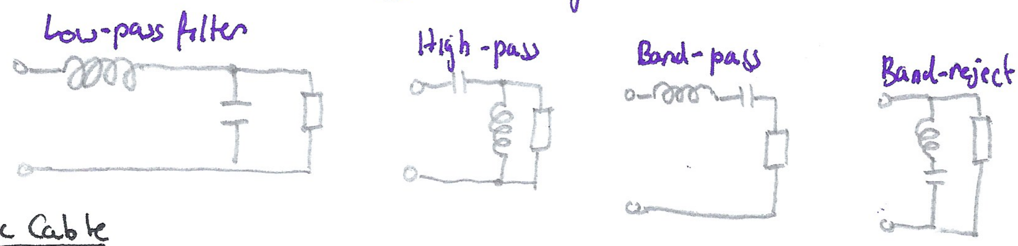
$$f(t) = \sum_n \alpha_n c_n e^{i\omega_n t}$$

Filters

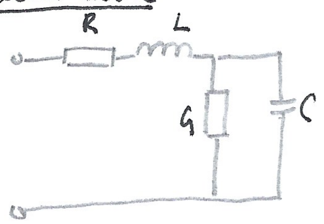
↳ Resistors - just have constant impedance : $Z = R$

↳ Capacitors - low impedance at high frequencies and high impedance at low frequencies. $Z(\omega) = \frac{1}{i\omega C}$

↳ Inductors : $Z(\omega) = i\omega L$ — in Henrys



Coax Cable



↳ non-zero series resistance R

↳ non-zero series inductance L

↳ non-zero shunt conductance G

↳ non-zero capacitance C

Effectively like a low pass filter - restricted to finite bandwidth W. Frequency components $> W \approx 1/RC$ attenuated by signal pathway

But add wideband noise to signal during transmission ⇒ normally shot noise

↳ if uniform, called white noise

CONTINUOUS INFORMATION

Given value X has probability density $p(x)$ with $\int_{-\infty}^{\infty} p(x)\,dx = 1$, differential entropy:

$$h(X) = \int_{-\infty}^{\infty} p(x) \log_2\left(\frac{1}{p(x)}\right) dx$$

$$h(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log_2\left(\frac{1}{p(x,y)}\right) dx\, dy$$

$$h(X|Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log_2\left(\frac{p(y)}{p(x,y)}\right) dx\,dy$$

$$h(X,Y) \leq h(X) + h(Y)$$

$$i(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y) \log_2\left(\frac{p(x|y)}{p(x)}\right) dx\,dy = h(Y) - h(Y|X)$$

$$C = \max_{p(x)} i(X;Y)$$

(12) Entropy maximised with equiprobable symbols - in continuous case if single value x is limited to range v, prob density $p(x) = 1/v$

Variance of continuous random variable X = power of sound signal = differential entropy $h(X)$. Can be proven that $p(x)$ of excursions around mean μ:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)$$

$$h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$$ (where $p(x)$ is maximised)
↳ white noise since power spectrum is flat

Channel with injected Gaussian noise

Limited Spectral Bandwidth W ⇒ signal + noise are lowpass: no frequency components higher than W
Power Spectral Density N_0 is white noise
↳ noise power = $N_0 W$
↳ noise variance = $N_0 W$

Given noise N independent of input signal X, $h(Y|X) = h(N)$

$$h(Y|X) = h(N) = \frac{1}{2}\log_2(2\pi e\sigma^2) = \frac{1}{2}\log_2(2\pi e N_0 W)$$

↳ channel output $Y = X + N$ has variance $P + N_0 W$

$$i(X;Y) = \frac{1}{2}\log_2\left(1 + \frac{P}{N_0 W}\right)$$

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{N_0 W}\right)$$ → in bits per channel symbol

Shannon's Third Theorem

$$\boxed{C = W\log_2\left(1 + \frac{P}{N_0 W}\right) \text{ bits/second}}$$

↓

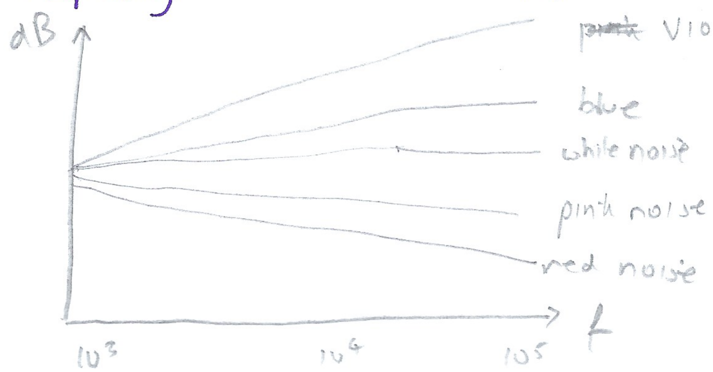
Shannon-Hartley Theorem

↓

Noisy Channel Coding Theorem

(1 + Signal to Noise ratio (SNR) ⇒ decibel
↳ $10 \times \log_{10}(SNR)$ - if ratio of power
↳ $20 \times \log_{10}(SNR)$ - if ratio of amplitudes

Increasing W yields monotonic but asymptotically limited improvement in capacity: as $W \to \infty$, $C \to \frac{P}{N_0}\log_2 e$



straight line only in loglog plot

(flicker noise - $\frac{1}{f}$) has power spectral density $\propto \frac{1}{f}$

$$C = \int_{\omega_1}^{\omega_2} \log_2(1+SNR(\omega))\,d\omega \text{ bits/sec}$$

ENCODINGS BASED ON FOURIER TRANSFORM PROPERTIES

Fourier transform extended to aperiodic function by making range $(b-a) \to \infty$. Interval between frequency components becomes infinitesimal $\to$ acquires density of reals not integers. Fourier Transform not Fourier Series

[$f(x)$ must be piecewise continuous and absolutely integrable]

$F(\omega)$ of $f(x)$

↳ ① $F(\omega)$ defined $\forall \omega \in \mathbb{R}$

↳ ② $F(\omega)$ is continuous

↳ ③ $\lim_{\omega \to \pm\infty} F(\omega) = 0$

For $f: \mathbb{R} \to \mathbb{C}$, $F: \mathbb{R} \to \mathbb{C}$ given by:

↳ $F(\omega) = \mathcal{F}_{[f]}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x) e^{-i\omega x}\,dx$

local-to-global and global-to-local

Properties

① $\mathcal{F}_{[af+bg]}(\omega) = a\mathcal{F}_{[f]}(\omega) + b\mathcal{F}_{[g]}(\omega)$

② Hermitian Symmetry: $F(-\omega) = \overline{F(\omega)}$

③ $f(x)$ is even function $\Rightarrow F(\omega)$ is even and purely even valued

④ $f(x)$ odd $\Rightarrow F(\omega)$ is odd and purely imaginary

frequency shifting

⑤ $\mathcal{F}_{[\exp(icx)f(x)]}(\omega) = \mathcal{F}_{[f]}(\omega - c)$

↳ shift in $f(x)$ by b causes $\mathcal{F}_{[f]}$ to be multiplied by $e^{i\omega b}$

multiply $f(x)$ by e^{icx} causes shift by c in $\mathcal{F}_{[f]}(\omega)$

⑥ Modulation Property

$$\mathcal{F}_{[f(x)\cos(cx)]}(\omega) = \frac{1}{2}\left(\mathcal{F}_{[f]}(\omega-c) + \mathcal{F}_{[f]}(\omega+c)\right)$$

$$\mathcal{F}_{[f(x)\sin(cx)]}(\omega) = \frac{1}{2i}\left(\mathcal{F}_{[f]}(\omega-c) - \mathcal{F}_{[f]}(\omega+c)\right)$$

Derivatives

Fourier transforms can be used to solve differential equations

f has derivative $f' \Rightarrow \mathcal{F}_{[f']}(\omega) = i\omega\mathcal{F}_{[f]}(\omega)$

$$\mathcal{F}_{[f^{(n)}]}(\omega) = (i\omega)^n \mathcal{F}_{[f]}(\omega)$$

can be used in the opposite direction for integration

Hence is effectively a filtering operation $\Rightarrow$ high-pass filtering

↳ can also generalise derivatives to non integers

Inverse Fourier Transform

$$f(x) = \int_{-\infty}^{\infty} \mathcal{F}_{[f]}(\omega) e^{i\omega x} \, dx$$

[again local-to-global and global-to-local property]

Convolution (commutative)

N.B. convolution of two Gaussians is a Gaussian.

$f, g : \mathbb{R} \to \mathbb{C}$

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y) g(y) \, dy$$

$$\mathcal{F}_{[f*g]}(\omega) = 2\pi \mathcal{F}_{[f]}(\omega) \cdot \mathcal{F}_{[g]}(\omega)$$

Passband (modulated carrier) -- demodulated --> baseband (encoded information)

Amplitude Modulation

Single sideband

$f(t)$ is baseband message

$\hookrightarrow F(\omega) = \mathcal{FT}\{f(t)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt$

instead firstly modulate by e^{ict}:

$$\mathcal{FT}\{e^{ict} f(t)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ict} f(t) e^{-i\omega t} \, dt$$

$$= F(\omega - c)$$

Demodulation:

$$f(t) = [e^{ict} f(t)] e^{-ict}$$

Double Sideband

Just multiply $f(t)$ by one cosine wave

$$\mathcal{FT}\{\cos(ct) f(t)\} = \frac{1}{2}(F(\omega - c) + F(\omega + c))$$

NB FDMA and phase modulation can be used for multiple applications

QUANTISED DEGREES OF FREEDOM IN CONTINUOUS SIGNAL

Bandlimiting continuous function means it has a finite, countable number of degrees-of-freedom ⇒ quantised

$\boxed{\text{Nyquist's Sampling Theorem}}$: if signal strictly bandlimited to some highest frequency W, then sampling at rate $2W$ specifies it completely everywhere. ⇒

$\boxed{\text{Logan's Theorem}}$: if bandlimited to one octave, simply listing zero-crossings determines it

(15)

Nyquist's Theorem

Sampling Function $comb(t) = \delta_x(t)$ - endless sequence of regularly spaced ~~times~~ tines separated by some sampling interval X.

Each tine is a Dirac δ-function: limit of a Gaussian who's width shrinks to 0. Multiplying signal with $\delta(t)$ samples value at $t=0$. Portrayed sequence of tines spaced by X is sum of shifted δ-functions, making sampling comb.

$$\delta_X(t) = \sum_n \delta(t - nX)$$

Sampling function $\delta_X(t)$ is self-Fourier → $\frac{1}{X}$ in frequency

$$FT(\delta_X(t)) = \Delta_X(\omega) = \frac{1}{X}\sum_m (\omega X - 2\pi m)$$

$$\delta(t) = \begin{cases} \infty & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}, \quad \int_{-\infty}^{\infty} \delta(t) = 1$$

Multiplying $f(t)$ by $\delta(t-c)$ picks out value of $f(t)$ at $t=c$

$$\int_{-\infty}^{\infty} f(t)\,\delta(t-c)\,dt = f(c)$$

Given signals which have been bandlimited to W:

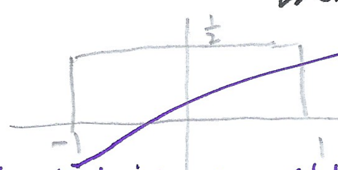
$$F(\omega) = 0 \quad \text{for } |\omega| > W$$

If not already true for $f(t)$, first run through lowpass filter - means Fourier Transform $F(\omega)$ becomes truncated.

If sampling $f(t)$ multiplying by $\delta_X(t)$ + use sampling rate $2W$ so sampling interval $X \leq 1/(2W)$ we know fourier transform of resulting sequence of samples = convolution of $F(\omega)$ with $\Delta_X(\omega)$ - Fourier transform of $\delta_X(t)$

Means $F(\omega)$ becomes reproduced at every tine of $\Delta_X(\omega)$. But given they did not overlap or superimpose get original by strict lowpass filtering.

↳ since sampling rate is double W



ideal lowpass filter described as zero-centred pulse function $F(\omega)$ in frequency domain ω.

if not, get aliasing, not separatable by lowpass filters

get weird effects - eg backwards moving wheels.

inverse Fourier transform of $F(\omega)$ is $f(x)$

$$f(x) = \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x}\, dx = \frac{1}{2}\int_{-1}^{1} e^{i\omega x}\, dx$$

$$= \left[\frac{1}{2}\frac{e^{i\omega x}}{ix}\right]_{\omega=-1}^{\omega=1} = \frac{\sin(x)}{x} = \text{sinc}(x)$$

For lowpass filter at W: $\frac{\sin(2\pi W t)}{2\pi W t}$

Effectively, sinc fills in space between sample points giving us back the same continuous function $f(t)$

Logan's Theorem : if signal is bandpass limited to one octave : $N_H \leq 2N_L$, then listing zero-crossings suffice for exact reconstruction (up to scale factor in amplitude)

↳ explains cartoons

Does not work with Amplitude Modulated signals

Gabor's Information Diagram : Uncertainty principle limits localisibility of a signal in time and frequency

de Broglie principle

Shorter duration = broader bandwidth

Diagrams constructed with axes: time & frequency - no function with area $> \frac{1}{4\pi}$ $\Rightarrow$ blobs can be any shape

↳ logons

GABOR - HEISENBERG - WEYL PRINCIPLE

Effective width (Δx) of complex-valued functions $f(x)$ in terms of normalised variance :

normalized first-moment of $\|f(x)\|$

$$\mu = \frac{\int_{-\infty}^{\infty} x f(x) f^*(x) dx}{\int_{-\infty}^{\infty} f(x) f^*(x) dx}$$

$$(\Delta x)^2 = \frac{\int_{-\infty}^{\infty} f(x) f^*(x) (x-\mu)^2 dx}{\int_{-\infty}^{\infty} f(x) f^*(x) dx}$$

Effective width of (Δw) $F(w) = FT\{f(x)\}$

$$(\Delta w)^2 = \frac{\int_{-\infty}^{\infty} F(w) F^*(w) (w-v)^2 dw}{\int_{-\infty}^{\infty} F(w) F^*(w) dw}$$

where v is the mean value of $\|F(w)\|$

$$v = \frac{\int_{-\infty}^{\infty} w F(w) F^*(w) dw}{\int_{-\infty}^{\infty} F(w) F^*(w) dw}$$

using Schwartz Inequality, can show : $(\Delta x)(\Delta w) \geq \frac{1}{4\pi}$

↳ property of all functions and their Fourier Transforms

Gabor Wavelets : family of functions achieving lower bound in uncertainty - complex exponentials multiplied by Gaussians

Logons

$$f(x) = \exp(-(x-x_0)^2 \alpha^2) \boxed{\exp(i w_0 (x - x_0))}$$

Gaussian's space parameter

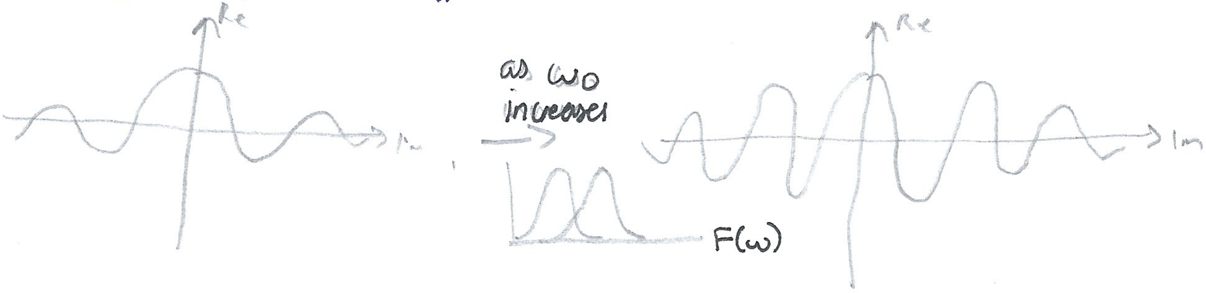
phasor

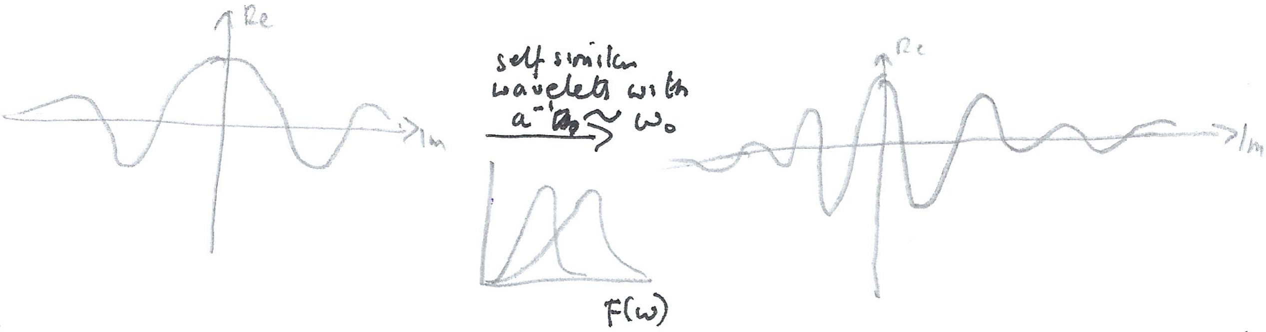
Wavelets are self-Fourier: $F(\omega) = \exp(-(\omega-\omega_0)^2\alpha^2)\exp(-ix_0(\omega-\omega_0))$
↳ this is still an Gaussian

Helical functions of x, localised at epoch x_0, modulated with frequency ω_0 and size or spread constant α

For wavelet with epoch at $x_0 = 0$, Fourier Transform is Gaussian at modulation frequency ω_0 and size $1/\alpha$

Gabor proposed them as expansion basis - but since non-orthogonal, difficult to compute expansion coefficients





2D Gabor Wavelets are used in Computer Vision - form complete basis image structure with vocabulary of:

↳ ① Location
↳ ② Scale
↳ ③ Spatial Frequency
↳ ④ Orientation
↳ ⑤ Phase

Particular example is iris recognition systems which is particularly used for phase structure
↳ works because of high entropy, hence collision avoidance

DATA COMPRESSION CODES

Redundancy = potential for compression

① <u>Run Length Encoding</u>: summarise repetition

② <u>Predictive Coding</u>: deviations from predictions encoded rather than just the information

③ <u>Dictionary Based Methods</u>: Lempel and Ziv exploit fact that strings of symbols have probabilities that very much more than probabilities of symbols individually - sparseness can be exploited
↳ first construct dictionary then scan - most common words have shortest indices to record.
↳ This can be done adaptively.

18

Vector Quantization

Also uses dictionary lookup - exploits sparsely populated combinations of sample symbols - can be generalized to image structures for example by creating codebook of pixel combinations. However, size of codebook memory for ~~memory size~~ is an issue.

Discrete And Fast Fourier Transforms

Describe functions of discrete values - $f[n]: f[0], f[1], \dots \Rightarrow$ effectively a vector of data points. Discrete Transforms of $f[n]$ is matrix multiplication $(N \times N)$ matrix

$$(F[1], \dots, F[N]) = (f[1], \dots f[N]) \begin{pmatrix} e_1[1] & \cdots & e_n[1] \\ \vdots & & \vdots \\ e_1[N] & \cdots & e_n[N] \end{pmatrix}$$

Fast Fourier Transform takes this from $O(N^2)$ to $O(N \log_2 N)$

N^{th} roots of unity: $e_k = \exp(-2\pi i n k/N)$

can represent any data sequence $f = (f[0], \dots f[N-1]) \in \mathbb{C}^n$ by vector sum:

$$f = \frac{1}{N} \sum_{k=0}^{N-1} \langle f, e_k \rangle e_k$$

Discrete Fourier Transform

periodic with period N.

$$\boxed{F[k]} = \langle f, e_k \rangle = \sum_{n=0}^{N-1} f[n] \exp(-2\pi i n k/N)$$ and have inverse:

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] \exp(2\pi i n k/N)$$

Complete DFT requires as many Fourier coefficients as number of values in sequence of $f[n]$ that we're doing DFT for.

Properties

↳ ① Cyclical Convolution: CC of $f[n]$ and $g[n]$ of period N:

$$(f * g)[n] = \sum_{m=0}^{N-1} f[m]\, g[n-m]$$

↳ if negative, taken mod N

Hence, DFT of $f * g$ is product $F[k]G[k]$ where F and G are the DFTs of f and g respectively

Fast Fourier Transform

Exploits ~~inefficienc~~ inefficiencies in computing DFT

$$F[k] = \sum_{n=0}^{N-1} f[n] \exp(-2\pi i n k/N)$$

$$= f[0] + f[1]\exp(-2\pi i k/N) + \dots + f[N-1]\exp(-2\pi i k(N-1)/N)$$

(19)

Hence, need to do N multiplications and N additions. To do this for $k=0, 1, \dots, (N-1)$, requires $2N^2$ operations

$W = \exp(2\pi i / N)$

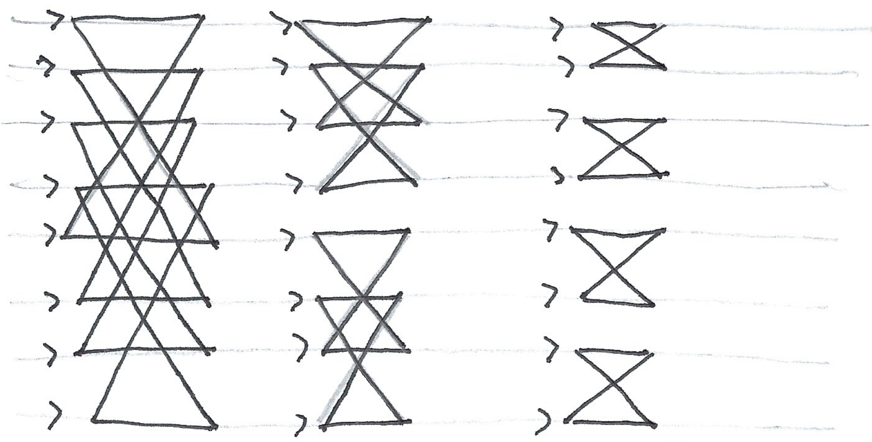
Hence, $\exp(-2\pi i n k / N) = W^{-nk}$

connects points halfway apart

Can rewrite Fourier coefficients in terms of powers of W. => can split into two halves - each requiring only a quarter of the multiplication

↳ This (Danielson-Lanczos Lemma) can be done recursively

↳ Hence, $O(N \log_2 N)$



Also called the Butterfly → only $O(N)$ in space terms

Fourier Methods can be extended to higher dimensions of functions. Fourier components of images (of $N \times M$) are 2D complex exponentials $f[n,m] = \exp(2\pi i (kn/N + jm/M))$

↳ spatial frequency $= \sqrt{(k^2 + j^2)}$

↳ phase (orientation) $= \arctan(j/k)$

Wavelet Representations Of Information

Wavelets are size-specific local undulations - acting as bases for representing information. Hence, used as basis for JPEG-2000. Image compression works since:

↳ neighbouring pixels are highly correlated

↳ Projecting this onto Fourier basis leads to highly decorrelated coefficients => many are 0 or small so don't need to be encoded

↳ done by Direct Cosine Transform on 8x8 tiles - coefficients are quantised => RLE is used

↳ more coarsely for higher frequencies and fewer bits used for this (defined by a quantisation table)

Dyadic Wavelets

Dyadic transformations for generating wavelet $\psi(x)$ spawn orthonormal wavelet basis $\psi_{jk}(x)$ for expansions of $f(x)$

$$f(x) = \underbrace{\sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty}}_{\text{also include } \theta \text{ for rotation if 2D}} c_{jk} \boxed{\psi_{jk}(x)}$$

→ generating by shifting and scaling operations applied to a mother wavelet → $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$

20

The compression works well but can suffer from block quantisation artifacts. 2000: using encoders like Daubechies 9/7 wavelet. Across multiple scales and lattice of positions, wavelet inner products with the image yield coefficients that constitute the Discrete Wavelet Transform (DWT). Implemented by ~~[illegible]~~ recursively filtering and downsampling image vertically and horizontally in scale pyramid

Kolmogorov Complexity

Any set of data can be created by a program - algorithmic complexity is the minimum length of this program ⇒ the Kolmogorov Complexity is the data string's minimal description length.

- ↳ this is approx equal to entropy of distribution from which data string drawn.
- ↳ Most sequences length n have Kolmogorov Complexity K close to n.
  - ↳ if $> n$, sequence is algorithmically random
- ↳ Not computable - cannot know you've found shortest possible description

Infinite string is K-incompressible iff:

$$\lim_{n \to \infty} \frac{K(x_1 x_2 \ldots x_n | n)}{n} = 1$$

Strong Law of Large Numbers for Incompressible Sequences - proportions of 0s and 1s in any incompressible string must nearly be equal. Must also pass all statistical test for randomness.

Scientific Applications

① Astrophysics

Pulsars are collapsed neutron stars - spinning causes emission of an electromagnetic radio beam. Signals faint and hidden in noise. Since coherent, can use auto-correlation integral $\rho_f(t)$ extracts this component from $f(t)$ as noise cancels itself out.

$$\rho_f(t) = \int_{-\infty}^{\infty} f(\tau) f(\tau + t)\, d\tau$$

Inverse Fourier transform of Power modulus.

If $f(t)$ has $FT\{f(t)\} = F(\omega)$, then $FT\{\rho_f(t)\}$ is power spectral density of $f(t)$

$$FT\{\rho_f(t)\} = \boxed{F(\omega) F^*(\omega)}$$ — power function

$$\rho_f(t) = FT^{-1}\{F(\omega) F^*(\omega)\}$$

② Genomics: translating genetic code into Amino Acids regarded as error-prone information channel → Entropy of gene inheritance is 1 bit

- ↳ $H = N$ bits per gene where N is number of generations gone back
- ↳ Every person appears multiple times in family tree → Genetic Isopoint is time in history when everyone is ancestor of everyone else: $N \approx 1.77 \log_2(m)$ generations ago.

Information transmission rate by blowfly photoreceptors $\propto SNR(\omega)$ analysis at synapses using $C = \int \log_2(1+SNR(\omega))\,d\omega$ is upto 1.65 kB/s

③ Biometric Pattern Recognition

↳ IrisKids
↳ Face Recognition
↳ Forensics

Discriminating power of a biometric reflects its entropy; surviving large database searches without false matches requires high entropy

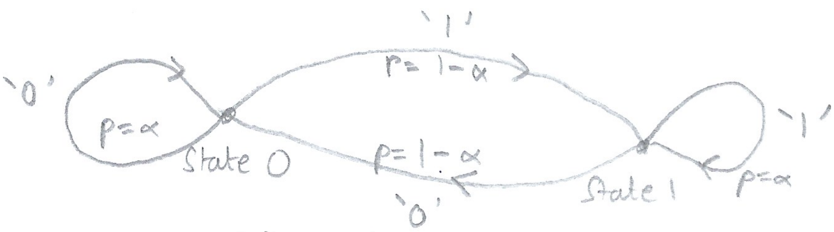
Why phase?

↳ Achieves structural information, independent of amplitude. Hence achieves some invariances
↳ Higher entropy than amplitude - hence coarsely quantised
↳ Phase classification ≡ clustering algorithm

Can use Gabor wavelets (which encode phase naturally, but in frequency-specific way) or in total way: ⇒ ① Analytic Functions: $f(x)$ - Hilbert Transforms $i f_{HT}$ cousin
↳ $f(x) - i f_{HT}(x)$

Iris Code

Regard Iris Code as channel. Codes computed from natural and from white noise iris patterns are well-modelled by bits emitted from two-state Markov process, with differing values of α:



$H(\alpha) = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$ bits per bit emitted
Statistical distributions well computed with $\alpha = 0.867 \Rightarrow C = 0.566$ bits per bit.
↳ Correlations lead to this reduction in entropy

Lots of mutual information between adjacent rings of Iris Codes (0.311 bits per bit pair max)
Mask used to remove eyelashes, specular reflections from glasses, etc.
↳ eg. A and B are data words of two Iris Codes
↳ C, D are respective mask words
result = $(A \oplus B)$ & C & D
↳ each executes in one clock tick