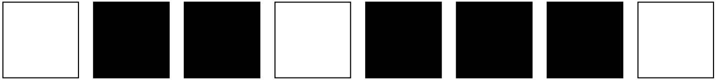
Information Theory, Inference, and Brain Networks



Pedro A.M. Mediano

Computational Neurodynamics Group
Department of Computing

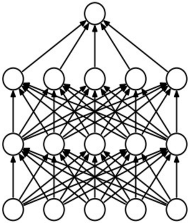
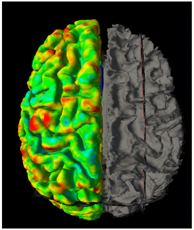
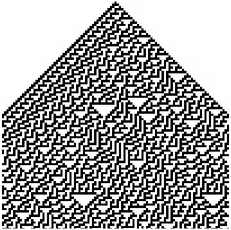
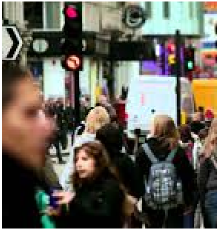
pmediano@ic.ac.uk

Imperial College, London

Motivation

BEFORE WE START...

- Whadda hell are ya doing here?
- Because I like things like these...

- My goal is **the scientific study of the emergence of distributed intelligence.**

PARALLEL DISTRIBUTED PROCESSING

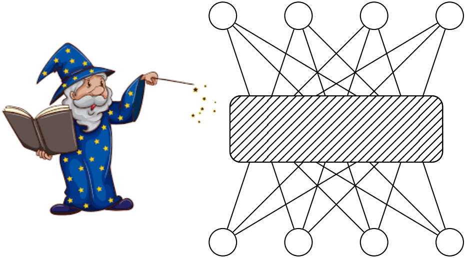
Explorations in the Microstructure of Cognition

Volume 2: Psychological and Biological Models

JAMES L. McCLELLAND, DAVID E. RUMELHART,
AND THE PDP RESEARCH GROUP

The missing piece in PDP

- PDP says that computation can happen, but it doesn't explain how.

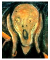


- **Challenge**: how can we describe the interaction between many neurons when performing a computation?
  ⏩ **Information theory.**

THE BASICS

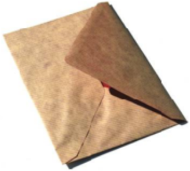
 Very important result/definition.

 Problem of interest.

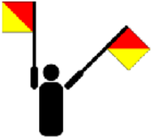
 Exercise. (Some optional, all recommended!)

WHAT IS INFORMATION?

In earlier times, information was identified with the objects that carried it.

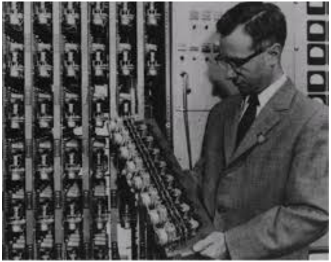


Later, information was carried by waves (sound, light, electromagnetic).

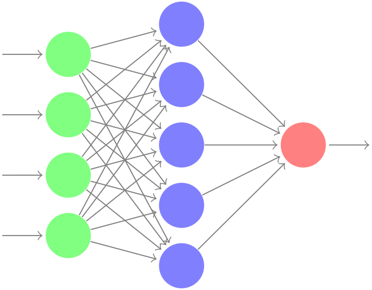


WHAT IS INFORMATION?

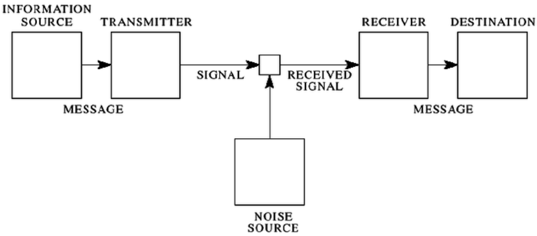
Also true for neural networks!

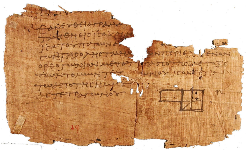


Rosenblatt holding the weights of a perceptron.



WHAT IS INFORMATION?





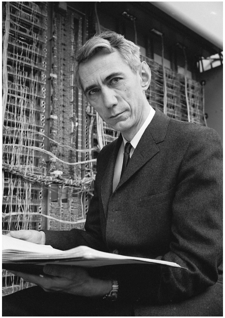
Information needs a **communication protocol**: a priori agreement between sender and receiver.

WHAT IS INFORMATION?

All developed in the seminal 1948 paper,

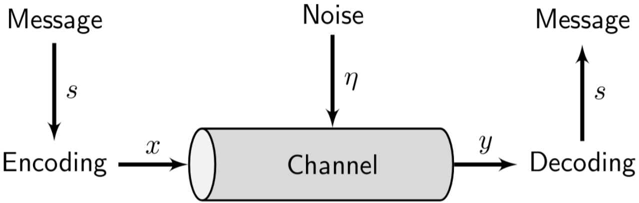
A Mathematical Theory of Communication

By **Claude Shannon.**



GOAL OF INFORMATION THEORY

How can we achieve optimal communication through a noisy channel?



PLAN FOR THESE LECTURES

1. $\boxed{\text{Entropy and coding.}}$
2. KL divergence and mutual information.
3. Links with statistics and maximum likelihood.
4. Research example.

ENCODING EXAMPLE

Let's find the shortest encoding for this message:

- Naive code:

  00 01 10 11

  0010000111000001

- Huffman code:

  1 01 001 000

  10011010001101

Symbol x	**Probability** $p(x)$
	1/2
	1/4
	1/8
	1/8

ENCODING EXAMPLE

- Average code length:

$$\sum_x p(x)L(x)$$

- Key idea: **frequent symbols have shorter sequences.**
- In particular, proportional to $-\log_2 p(x)$.

ENTROPY

Minimal description length for $p(x)$ messages in bits:

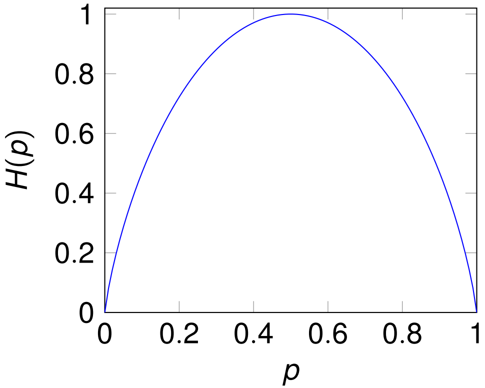
$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

What's the entropy of a random fair coin?

Discuss with your neighbour.

ENTROPY

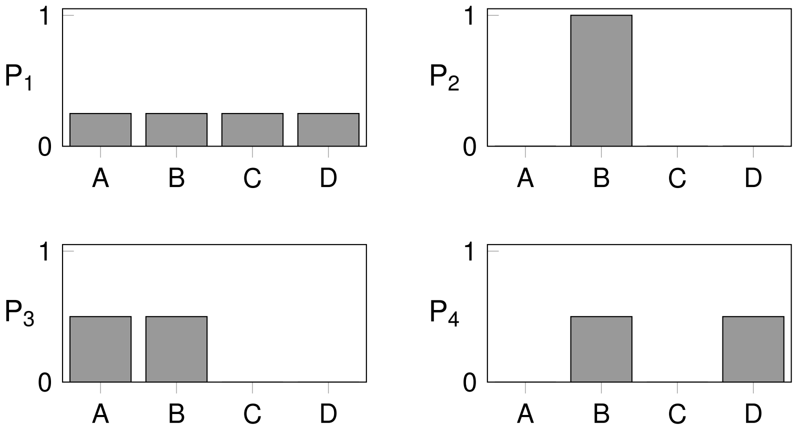
Bernoulli distribution: $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$



Entropy is a measure of *uncertainty* or *randomness*.

ENTROPY

- Rank these distributions from highest to lowest entropy.



JOINT ENTROPY

In addition, we can define these two quantities:

- Joint entropy:

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

- Conditional entropy:

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

KULLBACK-LEIBLER DIVERGENCE

What happens if we use the wrong code?

- Previous code:

  1 01 001 000

  0000010010100000100101

- New optimal code:

  01 1 00

  001101001101

Symbol x	**Assumed** $q(x)$	**Real** $p(x)$
	1/2	0
	1/4	1/4
	1/8	1/2
	1/8	1/4

KULLBACK-LEIBLER DIVERGENCE

 Extra cost incurred if we use the wrong code.

$$
\begin{aligned}
D_{\mathrm{KL}}(p \| q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \underbrace{-\sum_x p(x) \log_2 q(x)}_{\text{Actual message length}} \quad - \quad \underbrace{-\sum_x p(x) \log_2 p(x)}_{\text{Optimal message length}}
\end{aligned}
$$

 Prove that $D_{\mathrm{KL}}(p \| q) = 0$ iff $p = q$.

KULLBACK-LEIBLER DIVERGENCE

PROPERTIES

- KL divergence is non-negative:

$$D_{\mathrm{KL}}(p \| q) \geq 0$$

- The equality holds when $p = q$:

$$D_{\mathrm{KL}}(p \| q) = 0 \quad \text{iff} \quad p = q$$

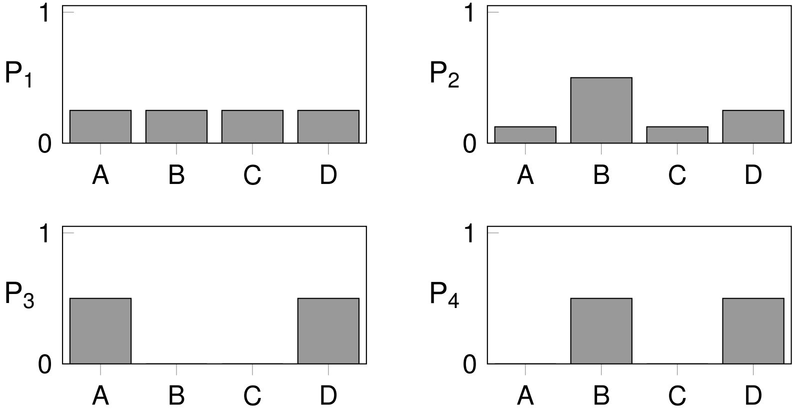
- It is not symmetric:

$$D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(q \| p)$$

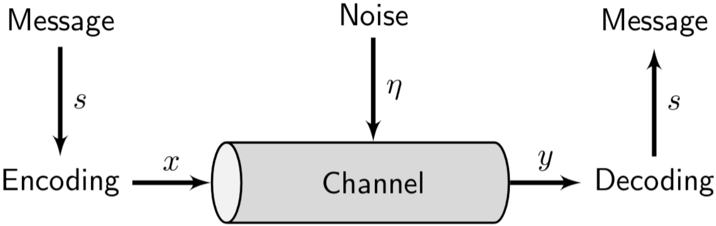
Kullback-Leibler divergence

- Calculate these KL divergences:

$D_{\mathrm{KL}}(P_3 \| P_1)$ $D_{\mathrm{KL}}(P_2 \| P_4)$ $D_{\mathrm{KL}}(P_4 \| P_2)$ $D_{\mathrm{KL}}(P_3 \| P_4)$



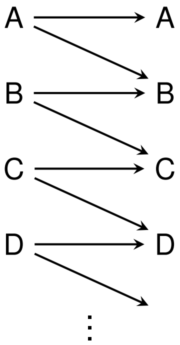
SENDING INFORMATION



THE NOISY TYPEWRITER

- Input X is uniform distribution on N symbols.
- Need $\log_2 N$ bits to encode.
- Symbols will be mixed by channel noise!
- Can only send one of $N/2$ symbols *without loss*.
- Rate of transmission to Y is

$$H(Y) - H(Y|X) = \log_2 N - 1$$
$$= \log_2 \frac{N}{2}$$



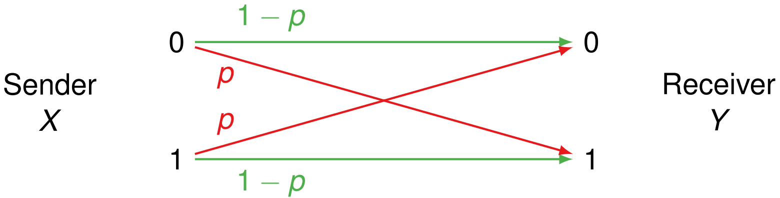
MUTUAL INFORMATION

How much does knowing X tell you about Y.

$$I(X;Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=\underbrace{H(Y)}_{\text{Uncertainty about } Y}-\underbrace{H(Y|X)}_{\text{Uncertainty about } Y \text{ given } X}$$

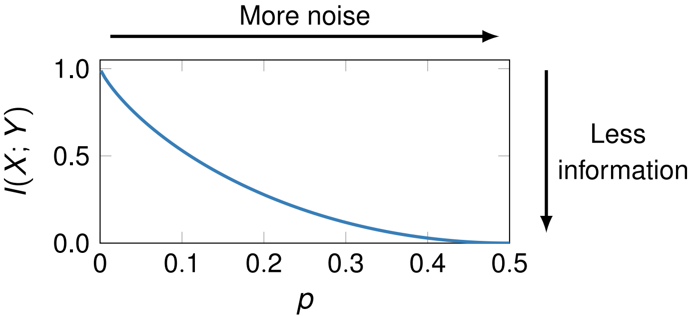
What's the MI of the binary symmetric channel?

Discuss with your neighbour.



MUTUAL INFORMATION

- MI is maximal when X and Y are identical and minimal when they are independent.



MI is a generalised measure of correlation.

MUTUAL INFORMATION

PROPERTIES

- MI is symmetric:

$$I(X;Y) = I(Y;X)$$

- MI is non-negative:

$$I(X;Y) \geq 0 \quad , \quad I(X;Y) = 0 \text{ iff } X \perp\!\!\!\perp Y$$

- MI is a KL divergence:

$$I(X;Y) = D_{\mathrm{KL}}\left(p(x,y) \| p(x)p(y)\right)$$

RECAP

- ✓ Entropy measures *uncertainty* or *randomness*.
- ✓ KL divergence measures *differences* between distributions.
- ✓ MI measures *correlation* between variables.

MORE EXAMPLES

Calculate the following:

1. $H(X)$
2. $H(X,Y)$
3. $I(X;Y)$
4. $D_{\mathrm{KL}}(p(x)\|p(y))$

$p(x,y)$	$x=0$	$x=1$
$y=0$	0.1	0
$y=1$	0.1	0.3
$y=2$	0.3	0.2

Continuous variables

CONTINUOUS VARIABLES

So far, we've used discrete variables only...

But in ML we use $\mathbb{R}^D$!

Can we extend these definitions to continuous variables?

ENTROPY IN $\mathbb{R}^1$

- We have a variable $X \in \mathbb{R}$ with pdf $f(x)$.
- We use bins of width Δ to get a discrete variable X^Δ with

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

Now we take $H(X^\Delta)$ as $\Delta \to 0$:

$$\begin{aligned} H(X^\Delta) &= -\sum p_i \log p_i \\ &= -\sum f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\underbrace{\sum \Delta f(x_i) \log f(x_i)}_{\text{Riemann integral}} - \underbrace{\log \Delta}_{\text{Divergent term}} \end{aligned}$$

DIFFERENTIAL ENTROPY

Ignoring the $\log \Delta$, we get the formula for differential entropy:

$$H(X) = -\int f(x) \log f(x) dx$$

Differential entropy is not a "real" entropy!

Mutual information in $\mathbb{R}^1$

Magically, for MI the divergent terms cancel out, and...

- Continuous MI is actually a real MI!

Summary:

✓ MI in continuous variables is interpretable.

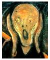
✗ Entropy in continuous variables is not.

MI invariance

MI is invariant to invertible mappings.

$$I(U;V) = I(X;Y) \quad \text{where} \quad U = f(X), V = g(Y)$$

if f and g are smooth and invertible.



Prove this result.

Tip: Use the fact that densities transform as $p(u, v) = |\boldsymbol{J}| p(x, y)$, with $\boldsymbol{J}$ the appropriate Jacobian, and the Jacobian is block-diagonal.

ENTROPY IN GAUSSIAN DISTRIBUTIONS

Let's calculate the entropy of a Gaussian $p(x) = \mathcal{N}(x|\mu, \mathbf{\Sigma})$:

$$H(X) = -\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \mathbf{\Sigma}) \; \log \mathcal{N}(x|\mu, \mathbf{\Sigma}) dx$$

$$= \frac{1}{2}\mathbb{E}[\log|2\pi\mathbf{\Sigma}|] + \frac{1}{2}\mathbb{E}[(x-\mu)^\top \mathbf{\Sigma}^{-1}(x-\mu)]$$

$$= \frac{1}{2}\log|2\pi\mathbf{\Sigma}| + \frac{1}{2}\mathbb{E}[(x-\mu)^\top \mathbf{\Sigma}^{-1}(x-\mu)]$$

Entropy in Gaussian distributions

For the second term:

$$\begin{aligned}\mathbb{E}\left[\operatorname{tr}\left((x-\mu)^\top \mathbf{\Sigma}^{-1}(x-\mu)\right)\right] &= \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbb{E}\left[(x-\mu)(x-\mu)^\top\right]\right)\\ &= \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\right)\\ &= D\end{aligned}$$

Overall:

$$H(X) = \frac{1}{2}\log|2\pi e\mathbf{\Sigma}|$$

Information measures have analytical solutions for Gaussian distributions.

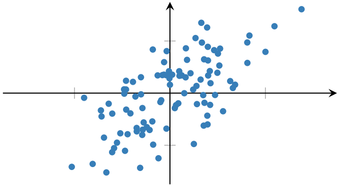
INFORMATION IN GAUSSIAN DISTRIBUTIONS

Given that the entropy of a Gaussian $\mathcal{N}(x|\mu, \mathbf{\Sigma})$ is:

$$H(X) = \frac{1}{2} \log |2\pi e \mathbf{\Sigma}|$$

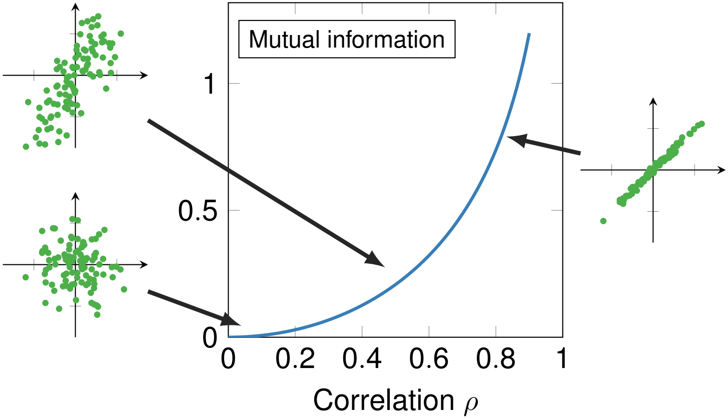
What's the mutual information between two Gaussians?

Discuss with your neighbour.



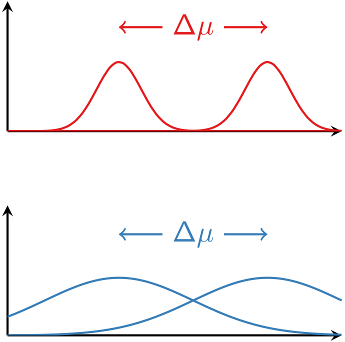
MI in Gaussian distributions

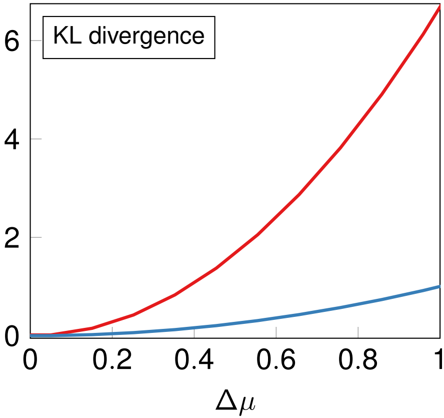
$$I(X;Y) = -\frac{1}{2}\log(1-\rho^2)$$



KL in Gaussian distributions

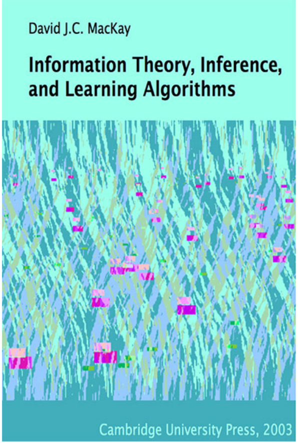
$$D_{\mathrm{KL}}(p_1(x) \| p_2(x)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{\sigma_2^2} - \frac{1}{2}$$





Lots of other stuff!

- Data processing inequalities.
- Rate-distortion theories.
- Error-correcting codes.



Statistics

All that encoding was ok, but...

What's the point?

Statistical interpretation of information theory

ENTROPY AND LIKELIHOOD

- Assume we have data $\boldsymbol{x}_i \in \mathbb{R}^D$ generated from $p^*(\boldsymbol{x})$.
- Take family of models $p \in \mathcal{P} = \{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^M\}$.
- Assume there exists a $\boldsymbol{\theta}^*$ such that $p(\boldsymbol{x}|\boldsymbol{\theta}^*) = p^*(\boldsymbol{x})$.
- Consider maximum-likelihood estimator $\boldsymbol{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}[p(\boldsymbol{x}|\boldsymbol{\theta})]$. Then:

$$\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta}_{\mathrm{ML}})] = \mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta}^*)] = -H[p^*(\boldsymbol{x})]$$

Entropy is the negative log-likelihood of the best model!

ENTROPY AND LIKELIHOOD

Sketch of a proof:

$$\begin{array}{rcl} D_{\mathrm{KL}}(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{\theta})) & \geq & 0 \\ & \Downarrow & \\ -\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})] & \geq & -\mathbb{E}[\log p^*(\boldsymbol{x})] \\ & \Downarrow & \\ -\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})] & \geq & H[p^*(\boldsymbol{x})] \end{array}$$

1. If $p^* \in \mathcal{P}$, the maximum is achieved iff $p(\cdot|\boldsymbol{\theta}) = p^*(\cdot)$ and therefore $\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta}_{\mathrm{ML}})] = -H[p^*(\boldsymbol{x})]$.

2. If $p^* \notin \mathcal{P}$, the margin between the MLE and the true model is $D_{\mathrm{KL}}(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{\theta}_{\mathrm{ML}})) > 0$.

ENTROPY AND LIKELIHOOD

- Another derivation:

$$D_{\mathrm{KL}}(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{\theta})) = -\underbrace{H[p^*(\boldsymbol{x})]}_{\text{Doesn't depend on } \theta} - \underbrace{\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})]}_{\text{Likelihood}}$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{\theta})) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})]$$

- To show that $\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})]$ is the normal likelihood, consider dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim p^*(\boldsymbol{x})$:

$$\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta})] = \int p^*(\boldsymbol{x}) \log p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \underset{\uparrow}{=} \frac{1}{N}\sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

Sampling

ENTROPY AND LIKELIHOOD

Maximising likelihood is equivalent to minimising KL!

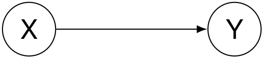
$$\boldsymbol{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} \mathbb{E}[p(\boldsymbol{x}|\boldsymbol{\theta})]$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \operatorname*{argmin}_{\theta} D_{\mathrm{KL}}(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{\theta}))$$

MODEL SELECTION

Are variables X and Y statistically independent?

M$_1$ (full):



$p(x, y)$

M$_2$ (restricted):

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

$$I(X;Y) = \int dx\, dy\, p(x,y) \log \frac{p(y|x)}{p(y)}$$

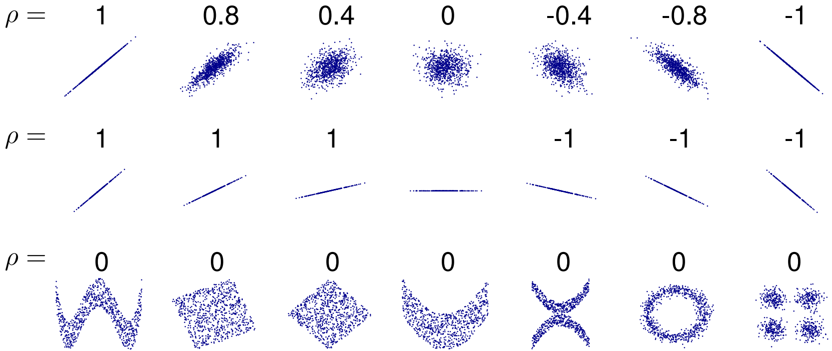
$p(y|x)$ ← Full model

$p(y)$ ← Restricted model

$I(X;Y) > 0$ iff M$_1$ explains the data better than M$_2$.

MUTUAL INFORMATION AND CORRELATION

In non-Gaussian distributions, MI acts as a generalised correlation.



RECAP

- ✓ Entropy functionals (MI, KL) arise from optimal communication principles.
- ✓ Alternative interpretation in terms of likelihood.
- ✓ All that's left is specifying a model $p(x|\theta)$.

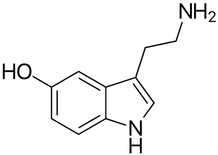
  $\rightarrow$ **Sampling and density estimation.**

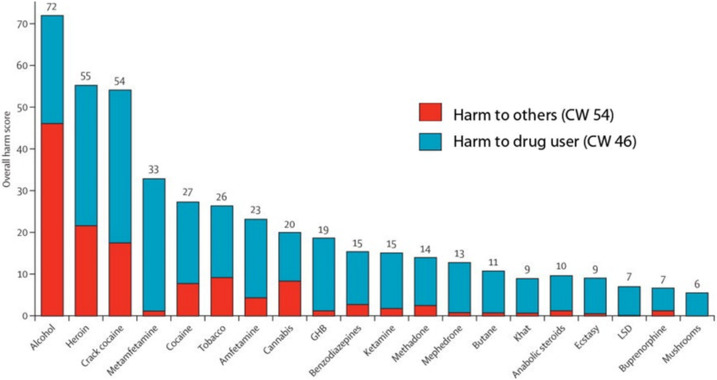
Neurodynamics

Psychedelics and hallucinogenics

Psychedelics and hallucinogenics

- Psychedelics affect the serotonergic system.
- Safest among common recreational drugs.





(Nutt et al. 2010)

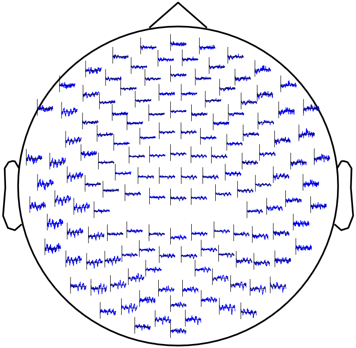
PSYCHEDELIC PHENOMENOLOGY

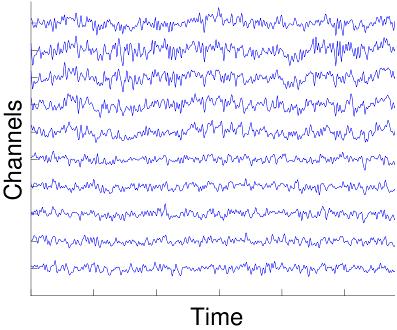
- Onset of audiovisual hallucinations.
  ▸▸ “*With eyes closed, I saw geometric patterns*.”
- Distortion of self models.
  ▸▸ “*I experienced a disintegration of my ’self’ or ’ego’*.”
- Increased cognitive flexibility.
  ▸▸ “*My thoughts wandered freely*.”

How does LSD alter information processing in the brain?

THE DATA

- High-frequency magnetoencephalographic (MEG) data.

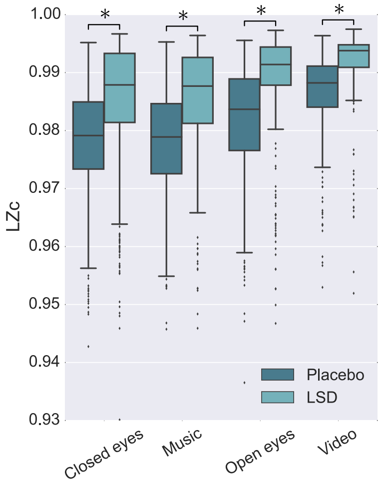




BRAIN ENTROPY

- Entropy estimator for sequential data known as *Lempel-Ziv*.
- Calculate average entropy of cortical surface.

Under LSD, brain has much higher entropy than usual.



(Marchesi & Mediano 2016)

THE ENTROPIC BRAIN

frontiers in
HUMAN NEUROSCIENCE

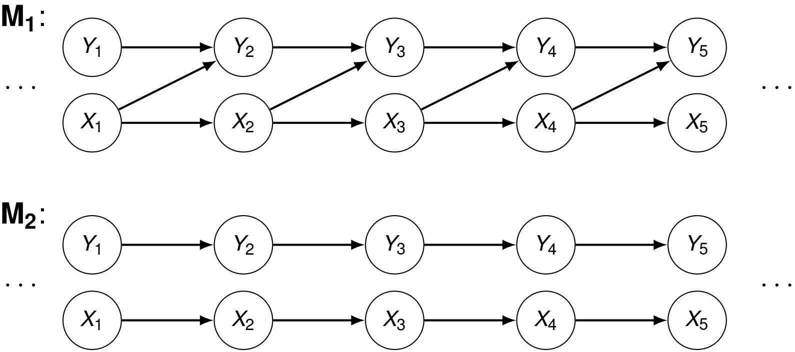
HYPOTHESIS AND THEORY ARTICLE
published: 03 February 2014
doi: 10.3389/fnhum.2014.00020

The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs

Robin L. Carhart-Harris[1]*, Robert Leech[2], Peter J. Hellyer[2], Murray Shanahan[3], Amanda Feilding[4], Enzo Tagliazucchi[5], Dante R. Chialvo[6] and David Nutt[1]

[1] *Division of Brain Sciences, Department of Medicine, Centre for Neuropsychopharmacology, Imperial College London, London, UK*
[2] *C3NL, Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK*
[3] *Department of Computing, Imperial College London, London, UK*
[4] *The Beckley Foundation, Beckley Park, Oxford, UK*
[5] *Neurology Department and Brain Imaging Center, Goethe University, Frankfurt am Main, Germany*
[6] *Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Buenos Aires, Argentina*

CONNECTIVITY INFERENCE



- Evidence for connected model M_1 over M_2 is:

$$I(X_t; Y_{t+1} | Y_t) = \int p(x_t, y_t, y_{t+1}) \log \frac{p(y_{t+1} | y_t, x_t)}{p(y_{t+1} | y_t)}$$

Connectivity inference

Two problems with this:

1. Compute integral $\int p(w_t)f(w_t)dw_t$, where $w_t = \{x_t, y_t, y_{t+1}\}$.
2. Evaluate likelihoods $p(y_{t+1}|y_t, x_t), p(y_{t+1}|y_t)$.

Solution 1: sampling!

$$
\begin{aligned}
I(X_t; Y_{t+1}|Y_t) &= \int p(w_t)\log\frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)}dw_t \\
&\approx \frac{1}{T}\sum_{i=1}^{T}\log\frac{p(y^i_{t+1}|y^i_t, x^i_t)}{p(y^i_{t+1}|y^i_t)}
\end{aligned}
$$

Solution 2: probabilistic regression!

1. Predict y_{t+1} from y_t.
2. Predict y_{t+1} from both y_t and x_t.

} Check if (2) is better than (1)

BRAIN NETWORKS

Algorithm: Iterative network inference.

Data: Set of brain regions $\mathcal{R}$

for $Y \in \mathcal{R}$ **do**

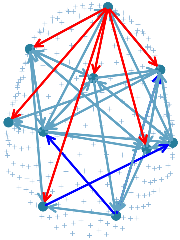
  Initialise $\mathrm{pa}(Y) = \emptyset$

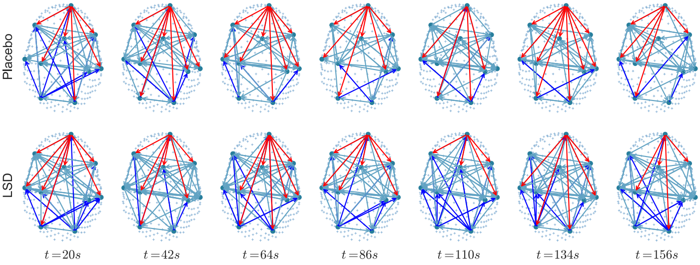
  while $\max_X I(X_t; Y_{t+1} | Y_t, \mathrm{pa}(Y)_t) > 0$ **do**

    $\mathrm{pa}(Y) \leftarrow \mathrm{pa}(Y) \cup \operatorname{argmax}_X I(X_t; Y_{t+1} | Y_t, \mathrm{pa}(Y)_t)$

  end

end

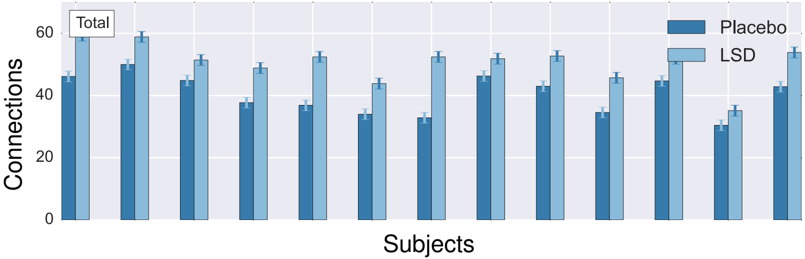




GLOBAL CONNECTIVITY

- Count total number of significant connections.

Under LSD, the brain is more interconnected than usual.

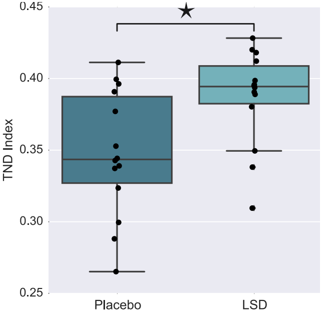


(Marchesi & Mediano 2016)

TRANSIENT NETWORK DISSIMILARITY

- Build transient networks $N_{t_0}, N_{t_1}, \ldots$
- How quickly do they change?
- Transient Network Dissimilarity (TND), average number of "rewirings:"

$$\mathbb{E}\left[\left|N_{t_i} - N_{t_{i+1}}\right|\right]$$



Under LSD, brain connectivity changes faster than usual.

Metastability.

(Marchesi & Mediano 2016)

CONCLUSION

- ✓ Information theory uses probability to study optimal communication.
- ✓ There is an alternative statistical interpretation of IT.
- ✓ We can combine ML and IT to study complex systems.
- ✓ Under LSD, the brain is more interconnected, more metastable, and more entropic.

Thank you for listening!