Information Theory
Credits and recommended readings:

- This is a great book! Thanks for making the figures available!


Motivation

Our world is full of regularities, structure. It is useful for brains to learn about these regularities: brains construct models of the world. Models allow to correctly interpret ambiguous sensory inputs or to predict future events:
• Observation: Brain needs representations that allow efficient access to information and do not waste precious resources. We want efficient representation and communication.

• A natural framework for dealing with efficient coding and transmission of information is information theory, which is based on probability theory.

• Idea: the brain may employ principles from information theory to code sensory information from the environment efficiently.
Observation: (Baddeley, 1997) Neurons in lower (V1) and higher (IT) visual cortical areas, show approximately exponential distribution of firing rates in response to natural video (panel A). Why may this be?
Saving space on your computer hard disk:
- use “zip” program to compress files, store images and movies in “jpg” and “mpg” formats
- lossy versus loss-less compression

Retina: $10^8$ receptors (rods and cones) representing the “input”; only $10^6$ retinal ganglion cells whose axons form “output”: Apparently, visual input is efficiently re-coded to save resources.

Caveat: Information theory just talks about symbols, messages and their probabilities but not their meaning, e.g., how survival relevant they are for the organism.
Communication over noisy channels

- Information theory also provides a framework for studying information transmission across noisy channels where messages may get compromised due to noise.

**Examples:**
- modem $\rightarrow$ phone line $\rightarrow$ modem
- satellite $\rightarrow$ radio waves $\rightarrow$ earth
- parent cell $\rightarrow$ daughter cell
- computer memory $\rightarrow$ disk drive $\rightarrow$ computer memory
Example: the symmetric binary channel
• we transmit a binary number across a channel:
  • transmitted signal $x$
  • received signal $y$
• Because the channel is noisy $y$ will sometimes be different from $x$. Specifically: assume the bit that is received is flipped with probability $f$.

\[
\begin{align*}
P(y = 0 \mid x = 0) &= 1 - f; \\
P(y = 0 \mid x = 1) &= f; \\
P(y = 1 \mid x = 0) &= f; \\
P(y = 1 \mid x = 1) &= 1 - f.
\end{align*}
\]
Can we reduce the number of errors somehow?

- physical solution vs. ‘system’ solution
Repetition codes:

• just send the data multiple times and take majority vote to decode!

• Example: R3 code: send each data item 3 times, take majority vote to decode
Image Example:

- just send the data multiple times, take majority vote!

\[ f = 10\% \]
• Information theory tells how good any code can only be. It turns out that we can transmit information at finite rate with arbitrarily low error rate.
Reminder:
\[
\log_2(x) = \frac{\ln(x)}{\ln(2)}
\]
\[
\log_a x = \frac{1}{\log_b a} \log_b x
\]
\[
\log(xy) = \log(x) + \log(y)
\]
\[
\log(x/y) = \log(x) - \log(y)
\]
\[
\log(x^y) = y \log(x)
\]
\[
\exp(\ln(x)) = x
\]
\[
a^{\log_a x} = x
\]

Matlab:
LOG, LOG2, LOG10

Note: base of logarithm is matter of convention, in information theory base 2 is typically used: information measured in \textit{bits} in contrast to \textit{nats} for base $e$ (natural logarithm).
How to Quantify Information?

discrete random variable $X$ taking values from the set of “outcomes” \( \{x_k, \ k=1,\ldots,N\} \) with probability $P_k$: $0 \leq P_k \leq 1; \sum P_k = 1$

**Question:** what would be a good measure for the information gained by observing the outcome $X = x_k$?

**Intuition:** Improbable outcomes should somehow provide more information (more “surprise”). Let’s try:

\[
I(x_k) = \log(1/P_k) = -\log(P_k)
\]

**Properties:**

\[
\begin{align*}
I(x_k) &\geq 0 & \text{since } 0 \leq P_k \leq 1 \\
I(x_k) &= 0 \text{ for } P_k = 1 & \text{certain event gives no information} \\
I(x_k) &> I(x_i) \text{ for } P_k < P_i & \text{log monotonic, } a < b \implies \log(a) < \log(b)
\end{align*}
\]
Consider observing the outcomes $x_a, x_b, x_c$ in succession; the probability for observing this sequence is (assuming independence): $P(x_a, x_b, x_c) = P_a P_b P_c$

$$I(x_k) = \log(1/P_k) = -\log(P_k)$$

Let's look at the information gained:

$$I(x_a, x_b, x_c) = -\log(P(x_a, x_b, x_c)) = -\log(P_a P_b P_c)$$

$$= -\log(P_a) - \log(P_b) - \log(P_c)$$

$$= I(x_a) + I(x_b) + I(x_c)$$

Information content is just the sum of individual information contents
**Entropy**

**Question:** What is the average information content when observing a random variable over and over again?

**Answer:** Entropy!

\[ H(X) \equiv E[I(X)] = -\sum_k P_k \log(P_k) \]

**Notes:**
- \( H(X) \) always bigger than or equal to zero
- measure of the uncertainty in a random variable
- generalization of variance
- related to minimum average code length for variable
- related concept in thermodynamics: there entropy can be thought of as a measure of the “disorder” of a system
Examples

1. Binary random variable: outcomes are \{0,1\} where outcome '1' occurs with probability \(P\) and outcome '0' occurs with probability \(Q=1-P\).

Question: What is the entropy of this random variable?

\[
H(X) = -\sum_k P_k \log(P_k)
\]

Answer: (just apply definition)
\[
H(X) = -P \log(P) - Q \log(Q)
\]

Note: entropy zero if outcome certain, entropy maximized if both outcomes equally probable, then \(H(X)=1\) bit.
2. **Horse Race**: eight horses are starting and their respective odds of winning are: 1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64

What is the entropy?

\[
H = -(1/2 \log(1/2) + 1/4 \log(1/4) + 1/8 \log(1/8) \\
+ 1/16 \log(1/16) + 4 \times 1/64 \log(1/64))
\]

\[= 2 \text{ bits}\]

What if each horse had chance of 1/8 of winning?

\[
H = -8 \times 1/8 \log(1/8)
\]

\[= 3 \text{ bits}\]

3. **Uniform**: for \(N\) outcomes entropy is maximized if all are equally likely *(proof: see later on blackboard):*

\[
H(X) = - \sum_{k=1}^{N} P_k \log(P_k) = - \frac{1}{N} \log \left( \frac{1}{N} \right) = \log(N)
\]
Note: Entropy is roughly the minimum average code length required for coding a random variable.

Idea: use short codes for likely outcomes and long codes for rare ones. Try to use every symbol of your alphabet equally often on average (e.g. Huffman coding described in Ballard book): basis for data compression methods.

Example: consider horse race again:

- Probabilities of winning: $1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64$
- Naïve code: 3 bit combination to indicate winner: $000, 001, ..., 111$
- Better code: $0, 10, 110, 1110, 111100, 111101, 111110, 111111$, requires on average just 2 bits (the entropy), 33% savings!
**Huffman Coding**

**Idea:** an algorithm for finding a good binary code for a discrete random variable.

Assume the random variable $x$ can take $N$ values

- Step 0: sort the $N$ values according to their probability and consider the $N$ values as leaves of a tree
- Step 1: construct binary tree by successively combining the two nodes with lowest probability
- Step 2: label the leftwards (rightwards) edges with 0 (1) to obtain binary codewords

**Note:** this constructs a prefix-free code, i.e. initial substrings of the codewords are no codewords themselves: no ambiguity (check example)
Example: Huffman code for the horse race problem

e.g., the binary codeword for horse 5 is ‘111100’
Notes:

• we've constructed a binary tree (for every edge with a '0' there’s also one with a '1')
• by construction, the resulting code is a prefix-free code, i.e. any partial codeword beginning at the start of a codeword but terminating early is not a codeword itself
• the most probable events are represented by the shortest messages
• Why are messages unique? Because they always end at the leafs of the tree (prefix-free code)
**Excursion: Convex functions**

**Definition:** $f(x)$ is **convex** in an interval $[x_1, x_2]$, if for any $0 \leq \lambda \leq 1$:

$$
\lambda f(x_1) + (1 - \lambda) f(x_2) \geq f(\lambda x_1 + (1 - \lambda) x_2)
$$

$f(x)$ is **strictly convex** if

$$
\lambda f(x_1) + (1 - \lambda) f(x_2) > f(\lambda x_1 + (1 - \lambda) x_2)
$$

unless $\lambda \in \{0, 1\}$.

Note: **concave** is same thing with '$\leq$'.
Examples: (of convex functions)

Note: reversing the sign makes a (strictly) convex function (strictly) concave and vice versa. E.g. $-x^2$ is concave.
Jensen's inequality

**Theorem:** let \( x \) be a random variable and let \( f(x) \) be a convex function. Then:

\[
E(f(x)) = \sum_i p_i f(x_i) \geq f\left(\sum_i p_i x_i\right) = f(E(x))
\]

If \( f(x) \) is strictly convex and \( E(f(x)) = f(E(x)) \), then the random variable \( x \) is a constant.
Illustration of Jensen’s inequality:
(for proof idea see MacKay book):

If a collection of masses $p_i$ are placed on a convex ~ curve $f(x)$ at locations $(x_i, f(x_i))$, then the centre of gravity of those masses, which is at $(\mathcal{E}[x], \mathcal{E}[f(x)])$, lies above the curve.

This means that:
\[
f(E(x)) \leq E(f(x))
\]
The uniform distribution has maximal entropy

Proof: (blackboard)
**Differential Entropy**

**Idea:** generalize to continuous random variables described by pdf:

\[
H(X) \equiv - \int_{-\infty}^{\infty} p(x) \log(p(x)) \, dx
\]

**Notes:**
- differential entropy can be negative, in contrast to entropy of discrete random variable
- but still: the smaller differential entropy, the “less random” is \( X \)

**Example:** uniform distribution

\[
p(x) = \begin{cases} 
\frac{1}{a}, & \text{for } 0 \leq x \leq a \\
0, & \text{otherwise}
\end{cases}
\]

\[
H(X) = - \int_{0}^{a} \frac{1}{a} \log(\frac{1}{a}) \, dx = \log(a)
\]
• What is the (differential) entropy of the exponential distribution?

• Homework: What is the (differential) entropy of the Gaussian distribution?
Maximum Entropy Distributions

Idea: maximum entropy distributions are “most random”
• For discrete RV, uniform distribution has maximum entropy (see earlier)
• For continuous RV, need to consider additional constraints on the distributions:

Important results:
• for a fixed range of the RV, i.e. \( \min < x < \max \), the uniform distribution has maximum entropy
• for a fixed variance, the Gaussian distribution has the highest entropy (another reason why the Gaussian is so special)
• for a fixed mean and \( p(x)=0 \) if \( x \leq 0 \), the exponential distribution has the highest entropy. Do cortical neurons have exponential firing rate distributions because this allows them to be most “informative” given a fixed average firing rate, which corresponds to a certain level of average energy consumption?
**Kullback Leibler Divergence**

**Idea:** Consider you want to compare two probability distributions $P$ and $Q$ that are defined over the same set of outcomes.

A “natural” way of defining a “distance” between two distributions is the so-called *Kullback-Leibler divergence (KL-distance)*, or relative entropy:

$$D(P \parallel Q) = E_P \left[ \log \frac{P(X)}{Q(X)} \right] = \sum_k P(x_k) \log \frac{P(x_k)}{Q(x_k)}$$
\[ D(P \parallel Q) \equiv E_P \left[ \log \frac{P(X)}{Q(X)} \right] = \sum_k P(x_k) \log \frac{P(x_k)}{Q(x_k)} \]

**Properties of KL-divergence:**

\( D(P \parallel Q) \geq 0 \) and \( D(P \parallel Q) = 0 \) if and only if \( P = Q \), i.e., if two distributions are the same, their KL-divergence is zero otherwise it's bigger. \( D(P \parallel Q) \) in general is not equal to \( D(Q \parallel P) \) (i.e. \( D(\cdot \parallel \cdot) \) is not a metric)

The KL-divergence is a quantitative measure of how “alike” two probability distributions are.

**Proof:** (see blackboard)

Hint: start from definition; use Jensen’s inequality:

\[ E(f(x)) \geq f(E(x)) \]
Example:

Consider two distributions for binary random variable $X$: $P(X=1)=p$, $P(X=0)=1-p$; $Q(X=1)=q$, $Q(X=0)=1-q$

$$D(P\parallel Q) = p \log (p/q) + (1-p) \log ((1-p)/(1-q))$$

Case 1: $p=q$: $D(P\parallel Q) = 0$
Case 2: $p=1/2$, $q=1/4$:
$$D(P\parallel Q) = \frac{1}{2} \log(2) + \frac{1}{2} \log(2/3) = 1 - \frac{1}{2} \log(3) = 1.2075$$

Generalization to continuous distributions:

$$D(p(x) \parallel q(x)) \equiv E_p \left[ \log \frac{p(x)}{q(x)} \right] = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx$$

The same properties as above hold.
Maximum Entropy Distributions

• exponential distribution: has maximum entropy of any distribution of positive RV with fixed mean

• uniform distribution: has maximum entropy of any distribution defined over a fixed interval

• Gaussian distribution: has maximum entropy of any distribution with a fixed variance

Now we can prove all this quite elegantly with the properties of the KL-divergence!

\[ \text{Corr}(X, Y) = E(XY) \]  

\[ \text{Cov}(X, Y) = E((X - \mu_x)(Y - \mu_y)) \]  

\[ \rho_{xy} = \frac{E((X - \mu_x)(Y - \mu_y))}{\sigma_x \sigma_y} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \]  

**Note:** independence implies uncorrelatedness, i.e. if two random variables \( X \) and \( Y \) are independent, it follows that their covariance is zero, but not vice versa. Proof: see blackboard
**Joint Entropy**

**Definition:** Consider two RVs X, Y with joint distribution P(X,Y). The joint entropy is defined as:

\[ H(X, Y) \equiv -\sum_{x,y} P(x, y) \log P(x, y) \]

**Notes:**
- if X, Y are independent, then H(X,Y) = H(X)+H(Y) (proof see blackboard)
- straightforward generalization to many RVs:
  \[ H(X_1, \ldots, X_n) \equiv -\sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n) \log P(x_1,\ldots,x_n) \]
- if \( X_i \) independent: \( H(X_1,\ldots,X_n) = H(X_1)+\ldots+H(X_n) \)
Mutual Information

Consider two random variables $X$ and $Y$ with a joint probability mass function $P(x, y)$, and marginal probability mass functions $P(x)$ and $P(y)$.

**Goal:** Can we find a quantitative measure of the degree of dependence of two random variables?

**Idea:** recall definition of independence of two RVs $X, Y$:

$$P(x, y) \equiv P(X = x, Y = y) = P(X = x)P(Y = y) \equiv P(x)P(y)$$

Mutual information defined as KL-divergence between the joint distribution and the product of marginals:

$$I(X; Y) \equiv D(P(X, Y) \parallel P(X)P(Y)) = \sum_x \sum_y P(x, y) \log \left( \frac{P(x, y)}{P(x)P(y)} \right)$$
\[ I(X;Y) = D(p(X,Y) \parallel p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \]

Properties of Mutual Information:

1. \( I(X;Y) \geq 0 \), equality if and only if \( X \) and \( Y \) are independent

2. \( I(X;Y) = I(Y;X) \) (symmetry)

3. \( I(X;X) = H(X) \), entropy is “self-information”

Notes:
- property 1 follows directly from the properties of the KL divergence discussed earlier
- property 2 follows from the symmetry of the definition
- property 3 follows from the definition
\[ I(X;Y) = D(p(X, Y) \parallel p(X)p(Y)) = \sum_{x} \sum_{y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \]

**Example:** symmetric binary channel, \( p(x=0) = p(x=1) = 1/2 \)

What is \( I(x;y) \)?

\[
I(x;y) = \begin{cases} 
(1-f)/2 \log((1-f)/2 / (1/4)) & | x=0, y=0 \\
+ f/2 \log( f/2 / (1/4)) & | x=0, y=1 \\
+ (1-f)/2 \log((1-f)/2 / (1/4)) & | x=1, y=1 \\
+ f/2 \log( f/2 / (1/4)) & | x=1, y=0 \\
\end{cases}
\]

\[ = (1-f) \log(1-f) + (1-f) \log(2) \]
\[ + f \log(f) + f \log(2) \]
\[ = 1 + f \log(f) + (1-f)\log(1-f) \]
cont'd: $I(x;y) = 1 + f \log(f) + (1-f)\log(1-f)$

$f = 0 : I(x;y) = 1 \text{ bit} \mid \text{perfect transmission}$
$f = \frac{1}{2} : I(x;y) = 0 \text{ bit} \mid \text{no transmission}$
Mutual Information, Independence, and Correlation

• independence is equivalent to zero mutual information (follows from properties of MI as KL-divergence)

• independence implies zero covariance, i.e. independent random variables are uncorrelated (see earlier)

• zero covariance does not imply independence!

Counter example:
Problems:

- what is the correlation between $x$ and $y$?
- are $x$ and $y$ independent?
- what are the entropies of $x$ and $y$?
- what is the joint entropy of $x$ and $y$, is it smaller than the sum of the individual entropies?
- what is the mutual information between $x$ and $y$?

P=1/4 each
\[
I(X;Y) = D(p(X,Y) \| p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)
\]

Different equivalent views of Mutual Information:

- deviation from independence:
  \[
  I(X,Y) = D(p(X,Y) \| p(X)p(Y))
  \]

- savings in encoding:
  \[
  I(X,Y) = H(X) + H(Y) - H(X,Y)
  \]

- why are these equivalent? (see black board)
Conditional Entropy

Define conditional entropy $H(Y \mid X)$ as expected value of entropies of $Y$ averaged over fixed values of $X$:

$$H(Y \mid X) = \sum_x p(x) H(Y \mid X = x)$$

$$= - \sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x)$$

$$= - \sum_x \sum_y p(x, y) \log p(y \mid x)$$

$$= - E_{p(x,y)} \log P(Y \mid X)$$

Chain rule:

$$H(X, Y) = H(X) + H(Y \mid X)$$
Chain rule for conditional entropy:

\[ H(X, Y) = H(X) + H(Y | X) \]

Proof: start with \( H(X,Y) \) and decompose it.

This relation allows another view of mutual information as the entropy of random variable \( X \) minus the conditional entropy of that random variable given another random variable \( Y \):

\[ I(X; Y) = H(X) - H(X | Y) \]

Proof: follows directly from definitions and laws of probability.
\[ I(X;Y) = D(p(X,Y) \parallel p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \]

**Generalization of Mutual Information to multiple RVs:**

\[ I(X_1, X_2, \ldots, X_n) = D(p(X_1, X_2, \ldots, X_n) \parallel p(X_1)p(X_2)\ldots p(X_n)) \]

\[ I(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n) \text{, where} \]

\[ H(X_1, X_2, \ldots, X_n) \equiv -E[\log p(X_1, X_2, \ldots, X_n)] \]
Consider recording from neuron (response $r$) while you present different stimuli $s$ to the animal.

The conditional entropy of the neuron’s response $r$ given the stimulus $s$ is sometimes called the noise entropy:

$$H_{\text{noise}} = H(r \mid s) = \sum_{s,r} P(s)P(r \mid s) \log P(r \mid s)$$

$$= \sum_{s,r} P(r,s) \log P(r \mid s)$$

The mutual information between stimulus and response can then be expressed as:

$$I(r; s) = H(r) - H_{\text{noise}}$$
Consider single neuron with continuous response $r$. When does distribution of neuron's responses have the maximum (differential) entropy?

Need to introduce constraints, e.g.:
- fixed firing rate range $\rightarrow$ uniform
- fixed mean of firing rate $\rightarrow$ exponential
- fixed variance of firing rate $\rightarrow$ gaussian

Different constraints lead to different maximum entropy distributions (see last lecture).
Making a neuron's response uniform

Question: how much should neuron respond to stimulus of strength $s$ with distribution $p(s)$ such that the pdf of its response $r$ is uniform in interval $[0, R]$?

We seek a monotonic function $f$ such that the distribution of $r=f(s)$ is uniform.

Solution:

$$f(s) = R \int_{s_{\min}}^{s} p(s') ds' = RF_s(s)$$

cumulative distribution function
Proof:

Probability of stimulus $s$ falling in $[s, s+\Delta s]$ is $p(s)\Delta s$. This produces responses falling in $[f(s), f(s+\Delta s)]$.

If distribution of output responses is flat ($p(r) = 1/R$), then the probability of the response falling in this interval should be $|f(s+\Delta s)-f(s)|/R$.

Thus we want: $|f(s+\Delta s)-f(s)|/R = p(s)\Delta s$

Now assuming that $f$ is monotonically increasing and considering the limit $\Delta s \to 0$, we have:

$$\frac{df(s)}{ds} = Rp(s)$$

The solution to this is the result from the previous slide.
Uniform distribution of responses in large monopolar cell of fly

contrast response of neuron compared to integral of natural contrast probability distribution
Histogram equalization

Passing a random variable through its own cumulative distribution function creates a random variable that is uniformly distributed over $[0,1]$.

\[ X \xrightarrow{F_x} Y \]

This is frequently used in signal processing, e.g. image processing.
Cortical Neurons

- neurons in visual cortex of cat and monkey have close to exponential firing rate distributions; may maximize entropy for fixed energy consumption [Baddeley et al., 1997]

Could intrinsic plasticity contribute to maximizing transmitted information? [Stemmler&Koch, 1999]

How does it interact with synaptic plasticity? [Triesch, 2004]
Plasticity Mechanisms for Learning

**Synaptic plasticity:**
- well-established mechanisms but still many open questions
- thought of as weight changes in neural networks

**Intrinsic plasticity:** for review: [Zhang & Linden, 2003]
- change how vigorously neuron responds to inputs; not as well studied, some evidence consistent with homeostasis
- thought of as changes to neuron’s activation function

brain uses local learning rules!
Intrinsic Plasticity Example

- after activity deprivation, cultured neurons are more excitable [Desai et al., 1999]
A Simple Model of Intrinsic Plasticity

Idea: use parametric sigmoid nonlinearity, two adjustable parameters $a \in \mathbb{R}^{>0}$, $b \in \mathbb{R}$

$$y = S_{ab}(h) = \frac{1}{1 + \exp(-(ah + b))}$$

Effect of varying $a$ and $b$: change shape of sigmoid
Gradient Rule

Idea: consider Kullback Leibler divergence between firing rate distribution and desired exponential of mean $\mu$:

$$D = d(f_y \parallel f_{\text{exp}}) = \int f_y(y) \log \left( \frac{f_y(y)}{\frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right)} \right) dy$$

$$= -H(y) + \frac{1}{\mu} E(y) + \log \mu$$

$H(y)$ and $E(y)$ depend on sigmoid parameters $a,b$. Derive stochastic gradient descent rule for $a,b$ to minimize $D$ (see blackboard):

$$\Delta a = \eta \left( a^{-1} + h - (2 + \mu^{-1})hy + \mu^{-1}h^2 \right)$$
$$\Delta b = \eta \left( 1 - (2 + \mu^{-1})y + \mu^{-1}h^2 \right)$$

this rule is strictly local!
Example 1: Gaussian input $h$ to the transfer function

distribution of $h$

- input distribution
- optimal transfer fct.
- learned transfer fct.

activation fct.: learned, optimal

output distribution

frequency

activity $y$
Example 2: uniform input $h$ to the transfer function

![Graph showing distribution of $h$ and output distribution](image)
**Example 3:** exponential input $h$ to the transfer function

**distribution of $h$**

**output distribution**

**activation fct.: learned, optimal**
Example 4: Recovery from sensory deprivation
Entropy maximization for population of neurons

Just making each neuron informative individually does not imply that the population codes effectively: neurons could be redundant.

Ideal situation is when neurons are each maximizing entropy, i.e. they should have the same marginal distribution (*probability equalization*) and they should be independent (*factorial code*). For a factorial code the joint entropy becomes the sum of the individual entropies.

Independence is usually difficult to achieve. Decorrelation is usually easier to obtain.
Recall: independence implies decorrelation but not vice versa.

If decorrelation is the goal we want:

\[ \text{Cov}(r_i, r_j) = \sigma^2 \delta_{ij} \]

This also makes all variances identical. This problem is generally more tractable (e.g. see Dayan&Abbott book for details).

Nice examples: prediction of receptive field properties in the Retina and LGN based on decorrelation idea.
**Mutual Information as Objective Function in Neural Nets**

Note: trying to maximize or minimize MI in a neural network architecture sometimes leads to biologically implausible non-local learning rules.
Transfer Entropy

• causality measure like Granger causality based on information theory (Schreiber, 2000)

\[ T_{J \rightarrow I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1} | i_n^{(k)}, j_n^{(l)})}{p(i_{n+1} | i_n^{(k)})} \]

\[ i_n^{(k)} = (i_n, \ldots, i_{n-k+1}) \]

\[ j_n^{(l)} = (j_n, \ldots, i_{n-l+1}) \]

• how different is our prediction of \( i \) if we do or don’t take the recent history of \( j \) into account?