Information Theory

Recommended readings:


What is the goal of sensory coding?
Attneave (1954): “A major function of the perceptual machinery is to strip away some of the *redundancy* of stimulation, to describe or encode information in a form more *economical* than that in which it impinges on the receptors.”
Motivation

**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 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?
What is the goal of sensory coding?
Sparse vs. compact codes:

two extremes: PCA (compact) vs. Vector Quantization

Advantages of sparse codes (Field, 1994):

• **signal-to-noise ratio:**
  small set of very active units above “sea of inactivity”

• **correspondence and feature detection:**
  small number of active units facilitates finding the right correspondences, higher order correlations are rarer and thus more meaningful

• **storage and retrieval with associative memory:**
  storage capacity greater in sparse networks (e.g. Hopfield)
What is the goal of sensory coding?
Another answer:

Find independent causes of sensory input.

Example: image of a face on the retina depends on identity, pose, location, facial expression, lighting situation, etc. We can assume that these causes are close to independent.

Would be nice to have a representation that allows easy read out of (some of) the individual causes.

Example: Independence: \[ p(a,b) = p(a)p(b) \]
\[ p(\text{identity, expression}) = p(\text{identity})p(\text{expression}) \]
(factorial code)
What is the goal of sensory coding?
More answers:

- Fish out information relevant for survival (markings of prey and predators).
- Fish out information that lead to rewards and punishments.
- Find a code that allows good generalizations. Situations requiring similar actions should have similar representations.
- Find representations that allow different modalities to talk to one another, facilitate sensory integration.

Information theory gives us a mathematical framework for discussing some but not all of these issues. Only speaks about probabilities of symbols/events but not their meaning.
Efficient coding and compression

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

Example: Retina
When viewing natural scenes, the firing of neighboring rods or cones will be highly correlated.
Retina input: $\sim 10^8$ rods and cones
Retina output: $\sim 10^6$ fibers in optic nerve, reduction by factor $\sim 100$

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

Galileo satellite $\rightarrow$ radio waves $\rightarrow$ earth

parent cell $\rightarrow$ daughter cell

computer memory $\rightarrow$ disk drive $\rightarrow$ computer memory
Recall:

\[
\log_2(x) = \frac{\ln(x)}{\ln(2)}
\]

\[
\log_a x = \frac{1}{\log_b a} \log_b x
\]

\[
\log(xy) = \log(x) + \log(y)
\]

\[
\log(x/y) = \log(x) - \log(y)
\]

\[
\log(x^y) = y \log(x)
\]

\[
\exp(\ln(x)) = x
\]

\[
a^{\log_a x} = x
\]

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

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

**Idea:** Improbable outcomes should somehow provide more information

Let’s try: $I(x_k) = \log(1/P_k) = -\log(P_k)$ “Shannon information content”

**Properties:**
$I(x_k) \geq 0$ since $0 \leq P_k \leq 1$

$I(x_k) = 0$ for $P_k=1$ (certain event gives no information)

$I(x_k) > I(x_i)$ for $P_k < P_i$

logarithm is monotonic, i.e: if $a<b$ then $\log(a)<\log(b)$
**Information gained for sequence of independent events**

Consider observing the outcomes $x_a, x_b, x_c$ in succession; the probability for observing this sequence is $P(x_a, x_b, x_c) = P_a P_b P_c$

Let’s look at the information gained:

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

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

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

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

Information gained is just the sum of individual information gains.
Entropy

**Question:** What is the average information gained 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:**
- entropy always bigger than or equal to zero
- entropy is a measure of the uncertainty in a random variable
- can be seen as generalization of variance
- entropy related to minimum average code length for variable
- related concept in physics and physical chemistry: there entropy is 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?

$$H(X) \equiv -\sum_k P_k \log(P_k)$$

Answer: (just apply definition)

$$H(X) = -P \log(P) - Q \log(Q)$$

Note: entropy zero if one outcome certain, entropy maximized if both outcomes equally likely (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?
\[ H = -(1/2 \log(1/2) + 1/4 \log(1/4) + 1/8 \log(1/8) + 1/16 \log(1/16) + 4 * 1/64 \log(1/64) \]
\[ = 2 \text{ bits} \]

What if each horse had chance of 1/8 of winning?
\[ H = -8 * 1/8 \log(1/8) \]
\[ = 3 \text{ bits (maximum uncertainty)} \]

3. Uniform: for N outcomes entropy maximized if all equally likely:
\[ H(X) = -\sum_{k=1}^{N} P_k \log(P_k) = -N \frac{1}{N} \log\left(\frac{1}{N}\right) = \log(N) \]
Entropy and Data Compression

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).

This is 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!
Differential Entropy

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

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

**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

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

\[ H(X) = -\int_{0}^{a} \frac{1}{a} \log\left(\frac{1}{a}\right) \, dx = \log(a) \]
**Maximum Entropy Distributions**

**Idea:** maximum entropy distributions are “most random”
For discrete RV, uniform distribution has maximum entropy

For continuous RV, need to consider additional constraints on the distributions:

Neurons at very different ends of the visual system show the same *exponential distribution* of firing rates in response to “video watching” (Baddeley et al., 1997)
Why exponential distribution? 

\[ p(r) = \frac{1}{\mu} \exp\left(-\frac{r}{\mu}\right) \]

Exponential distribution maximizes entropy under the constraint of a fixed mean firing rate \( \mu \).

**Maximum entropy principle**: find density \( p(x) \) with maximum entropy that satisfies certain constraints, formulated as expectations of functions of \( x \):

\[
\int p(x) F_i(x) dx = c_i, \; i = 1, \ldots, n
\]

Answer:

\[
p(x) = A \exp\left(\sum_i a_i F_i(x)\right)
\]

Interpretation of mean firing rate constraint: each spike incurs certain metabolic costs: goal is to maximize transmitted information given a fixed average energy expenditure.
Examples

Two important results:
1) for a fixed variance, the Gaussian distribution has the highest entropy (another reason why the Gaussian is so special)

2) for a fixed mean and $p(x)=0$ if $x \leq 0$, the exponential distribution has the highest entropy. Neurons in brain may 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
Intrinsic Plasticity

Not only the synapses are plastic! Now surge of evidence for adaptation of dendritic and somatic excitability. Typically associated with *homeostasis* ideas, e.g.:

- neuron tries to keep mean firing rate at desired level
- neuron keeps variance of firing rate at desired level
- neuron tries to attain particular distribution of firing rates
**Question:** how do intrinsic conductance properties need to change to reach specific firing rate distribution?

Stemmler&Koch (1999): derived learning rule for two compartment Hodgkin-Huxley model:

\[
\frac{dV}{dt} = \sum_i g_i m_i^{p_i} h_i^{q_i} (E_i - V),
\]

where \( g_i \) is the (peak) value of the \( i \)-th conductance adaptation of \( g_i \) leads to learning...

... of uniform firing rate distribution

**Question 1:** formulations of intrinsic plasticity for firing rate models?

**Question 2:** how do intrinsic and synaptic learning processes interact?
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:

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

**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 \( D(Q \parallel P) \) \( (i.e. D(.\parallel .) \) is not a *metric*).

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

**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.
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:** KL-divergence is quantitative measure of “alikeness” of distributions of two random variables. Can we find a quantitative measure of independence of two random variables?

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

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

We define as the mutual information the KL-divergence between the joint distribution and the product of the marginal distributions:

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

**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”

**Generalization to multiple RVs:**

\[ I(X_1, X_2, \ldots, X_n) = D(p(X_1, X_2, \ldots, X_n) \parallel p(X_1)p(X_2)\ldots p(X_n)) \text{ deviation from independence} \]

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

\[ H(X_1, X_2, \ldots, X_n) \equiv -E[\log p(X_1, X_2, \ldots, X_n)] \]
Mutual Information as Objective Function

Haykin

Linsker’s Infomax

ICA

Imax by Becker and Hinton

Note: trying to maximize or minimize MI in a neural network architecture sometimes leads to biologically implausible non-local learning rules
Cocktail Party Problem

**Motivation:** cocktail party with many speakers as sound sources and array of microphones, where each microphone is picking up a different mixture of the speakers.

**Question:** can you “tease apart” the individual speakers just given $x$, i.e. without knowing how the speakers’ signals got mixed? (*Blind source separation (BSS)*)
Blind Source Separation Example

original time series of source $s$

$s_1(t)$

$s_2(t)$

linear mixture $x$

$x_1(t)$

$x_2(t)$

unmixed signals:

$x_1(t) = a_{11}s_1 + a_{12}s_2$

$x_2(t) = a_{21}s_1 + a_{22}s_2$

Demos of blind audio source separation: [http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html](http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html)
Definition of ICA

Note: there are several, this is the simplest and most restrictive:
• sources are independent and non-gaussian: $s_1, \ldots, s_n$
• sources have zero mean
• $n$ observations are linear mixtures: $x(t) = As(t)$

• the inverse of $A$ exists: $W = A^{-1}$

Goal: find this inverse. Since we do not know the original, cannot compute it directly but we will have to estimate it: $y(t) = \hat{W}x(t)$

• once we have our estimate, we can compute the sources: $\hat{s}(t) = \hat{W}x(t)$

Relation of BSS and ICA:
• ICA is one method of addressing BSS, but not the only one
• BSS is not the only problem where ICA can be usefully applied
Restrictions of ICA

Need to require: (it’s surprisingly little we have to require!)
- sources are independent
- sources are non-gaussian (at most one can be gaussian)

Ambiguities of solution:
- sources can be estimated only up to a constant scale factor
- may get permutation of the sources, i.e. sources may not be recovered in their right order

\[ x_1(t) = a_{11}s_1 + a_{12}s_2 \]
\[ x_2(t) = a_{21}s_1 + a_{22}s_2 \]

a) multiplying source with constant and dividing that source’s matrix entries by same constant leaves \( x \) unchanged
b) switching these columns leaves \( x \) unchanged
Principles for Estimating $W$: $y(t) = \hat{W}x(t)$

Several possible objectives (contrast functions):
- requiring outputs $y_i$ to be uncorrelated is not sufficient
- outputs $y_i$ should be maximally independent
- outputs $y_i$ should be maximally non-gaussian, (central limit theorem), (*projection pursuit*)
- maximum likelihood estimation

**Algorithms:**
- various different algorithms based on these different objectives with interesting relationships between them
- to evaluate criteria like independence (MI), need to make approximations since distributions unknown
- typically, *whitening* is used as a pre-processing stage (reduces number of parameters that need to be estimated)
Artifact Removal in MEG Data

original MEG data

ICs found with ICA algorithm
ICA on Natural Image Patches

The basic idea:
• consider 16 by 16 gray scale patches of natural scenes
• what are the “independent sources" of these patches?

Result:
• they look similar to V1 simple cells: localized, oriented, bandpass

Hypothesis:
Finding ICs may be principle of sensory coding in cortex!

Extensions:
stereo/color images, non-linear methods, topographic ICA
Discussion: Information Theory

Several uses:
• Information theory is important for understanding sensory coding and information transmission in nervous systems

• Information theory is a starting point for developing new machine learning and signal processing techniques such as ICA

• Such techniques can in turn be useful for analyzing neuroscience data as seen in the EEG example

Caveat:
• Typically difficult to derive biologically plausible learning rules from information theoretic principles. Beware of non-local learning rules.
Some other limitations

The “standard approach” for understanding visual representations:
The visual system tries to find efficient encoding of random natural image patches or video fragments that optimizes statistical criterion: sparseness, independence, temporal continuity or slowness

Is this enough for understanding higher level visual representations?

Argument 1: the visual cortex doesn’t get to see random image patches but actively shapes the statistics of its inputs

Argument 2: approaches based on trace rules break down for discontinuous shifts caused by saccades

Argument 3: vision has to serve action: the brain may be more interested in representations that allow efficient acting, predict rewards etc.

Yarbus (1950s): vision is active and goal-directed
Building Embodied Computational Models of Learning in the Visual System

Anthropomorphic robot head:
• 9 degrees of freedom, dual 640x480 color images at 30 Hz

Autonomous Learning:
• unsupervised learning, reinforcement learning (active), innate biases
Simple illustration: clustering of image patches an “infant vision system” looked at

- **a:** system learning from random image patches
- **b:** system learning from “interesting” image patches (motion, color)

The biases, preferences, and goals of the developing system may be as important as the learning mechanism itself:

the standard approach needs to be augmented!