Written by Yarden Nativ, mail: nativyardenac@gmail.com

9th Recitation 19.05.22

Information theory

Entropy

Shannon's communication system theory (Shannon and Weaver, 1949) offers a quantitative framework to analyze the transmission of a message from a source to a destination. Shannon's idea was to define a quantitative measure, entropy, representing the potential information content of a given set of messages. Given that entropy refers to a set of messages, the variability and the context of the ensemble is critical for quantifying the information content. Zero variability means that the message will be transformed fully in the first trial. Higher variability can indicate some level of uncertainty of the system (for example the timing of two preceding spikes in a Poisson neuron).

The measure, entropy of a system, represents the uncertainty about the state of that system. It is measured by the number of bits required to fully describe the state of the system. Let X be a discrete random variable with a given distribution p(x):

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

Class Exercise:

a. What is the entropy of a fair dice?

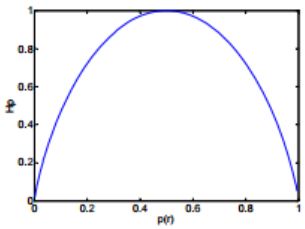
Solution: $H(x) = -6 \cdot \frac{1}{6}\log_2\left(\frac{1}{6}\right) = -\log_2\left(\frac{1}{6}\right)\ bits$

b. Given a coin with probability p for heads, what is its entropy?

Solution: $H(x) = -[p \cdot \log_2(p) + (1-p) \cdot \log_2(1-p)]\ bits$

c. Given the two examples, when is the entropy maximal and when minimal?

Solution: Maximal for uniform distribution and minimal when only one value is possible.



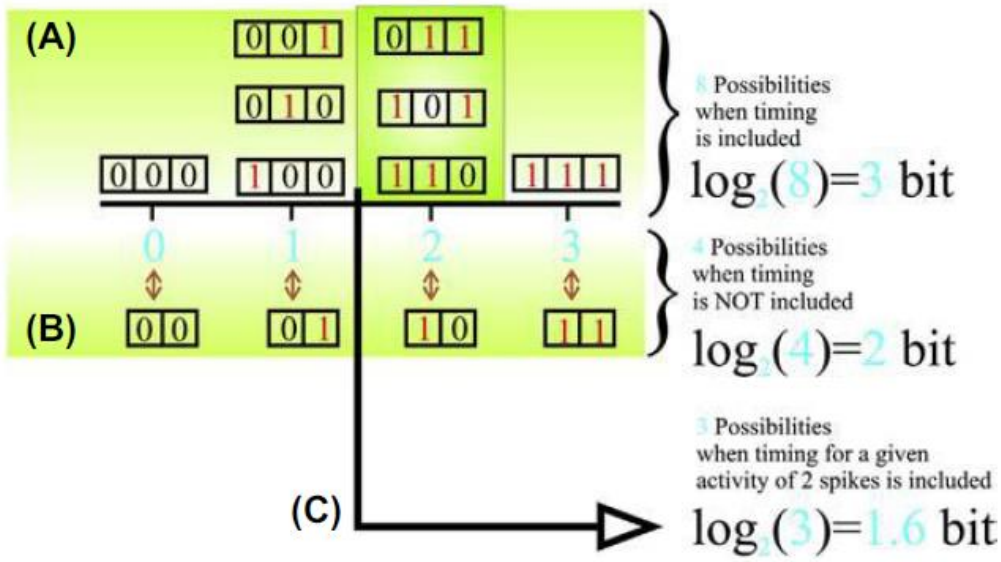
Class discussion: What is the relevant random variable?

1. **The occurrence of a spike in a given time**:
   We have N bins, so we can support $N+1$ possible spike counts (a 0 spike result is also possible). The maximal entropy, assuming that each response is equally likely is $\log_2(N+1)$ bits. This result depends on the number of bins (which we set arbitrarily).
   Fig. B shows this example.

a. **The timing of a spike**:
   We have N_1 spikes which can be sorted across N bins by $\binom{N}{N_1}$. The maximal entropy for a uniform distribution of arrangements is $\log_2\binom{N}{N_1}$.This is still resulting dependency on the bin size, and it assumes the number of spikes is deterministic.
   Fig. C shows this example.

By combining the two possibilities we get the distribution of spike count for a small bin, representing a distribution like in Poisson neuron (fig A).

Nevertheless, there is still an ongoing debate in neuroscience should we consider spike rate as carrying information or other measurements like synchrony (like in synfire chain), membrane potentials, chemical signals involving glia cells and more.



Written by Yarden Nativ, mail: nativyardenac@gmail.com

Mutual and Conditional Entropy:

Given two random variables, we want to quantify using the entropy measure how much information the first gives about the second. For this we use mutual entropy and conditional entropy, when only the latter represents a possible causal relationship between the variables:

Mutual entropy:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2[p(x,y)]$$

Conditional entropy:

$$H(Y|X) = \sum_{x_i\in X} p(x_i)H(Y|x=x_i) =$$

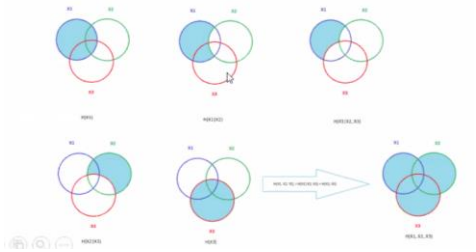
$$\sum_{x_i\in X} p(x_i)\cdot\sum_{y_j\in Y} -p\left(y_j|x_i\right)\log_2\left(p\left(y_j|x_i\right)\right)$$

The chain rule connecting between them:

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

Example of the chain rule for 3 Random Variables:

$$H(X_1,X_2,X_3) = H(X_1|X_2,X_3) + H(X_2,X_3) = H(X_1|X_2,X_3) + H(X_2|X_3) + H(X_3)$$



Basic properties of mutual and conditional entropy:

- $0 \le H(X) \le log_2|X|$ (extremum when X is uniformly distributed, number of elements)
- In fully independent distributions: $H(X,Y) = H(X) + H(Y)$, $H(Y|X) = H(Y)$
- $H(Y|X) \le H(Y)$
- $H(X|Y) \ne H(Y|X)$
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- $H(X_1,X_2,\dots,X_n) \le \sum_n H(X_j)$ based on the chain rule

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Class Exercise:

Let (X, Y) have the following joint distribution:

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

a. Find H(X), H(Y)

b. Find H(X|Y), H(Y|X)

c. Find H(X,Y)

Solution:

a. To find the distributions of X and Y, we need to sum the rows and columns of the matrix and get the marginals:

$$Marginal_X = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$$

$$Marginal_Y = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

For this reason:

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - 2\cdot\frac{1}{8}\log_2\frac{1}{8} = \frac{1}{2} + \frac{2}{4} + 2\cdot\frac{3}{8} = \frac{7}{4}\ bits$$

$$H(Y) = -4\cdot\frac{1}{4}\log_2\frac{1}{4} = 2\ bits$$

b. For a given X from this distribution, we look at the distribution of the Y and normalize it by dividing to the full sum:

$$H(X|Y) = \sum_{i=1}^{4} p(Y=i)H(X|Y=i) =$$

$$= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4}H(1,0,0,0)$$

$$= \frac{1}{4}\cdot\frac{7}{4} + \frac{1}{4}\cdot\frac{7}{4} + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 0 = \frac{11}{8}bits$$

Similarly $H(Y|X) = \frac{13}{8}$, $H(X,Y) = H(Y|X) + H(X) = \frac{27}{8}$ (can be calculated also by definition).

Note that the entropy $H(X|Y) < H(X)$, therefore the Random Variables aren't independent. The level of the dependency between them is quantified by subtraction between the two entropies and defines the level of mutual information.

Mutual information

Given that two variables are dependent, we can quantify $I(X;Y)$ as the relative entropy between the joint distribution and the product distribution $p(X) \cdot p(Y)$. This measure represents the level in which each variable reduces the entropy of the other variable. If X is fully dependent on Y, then the mutual information is maximal, and the conditional entropy of X given Y is 0. The quantification of the distance between the two distribution is given by D_{KL} from which we can derive:

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$

Important properties:

- $I(X;X) = H(X) - H(X|X) = H(X)$
- $0 \leq I(X;Y) \leq \min\big(H(X), H(Y)\big)$- when one variable is contained in the other than the mutual information is the entropy of the smaller variable.
- Data processor inequality: if $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$

Class Exercise:

Neuron A may be recorded at two rates: 40spikes/s (p=0.7) or 80spikes/s (p=0.3). Neuron B fires at two rates as well: 20spikes/s or 40spikes/s. When A fires at 40spikes/s the probability of B to fire 40spikes/s is 0.8. When A fires at 80spikes/s the probability of B to fire 20spikes/s is 0.8.

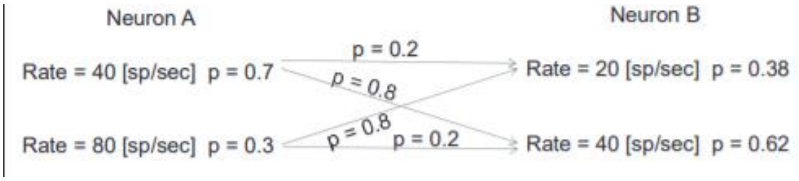
a. Calculate the entropy of A, B

b. Calculate the conditional entropy of B given A .

c. Calculate the mutual information

Solution:

Here is a full illustration of the relationship between the two neurons:



Now we can calculate:

$H(A) = H(0.7,0.3) = 0.88\ bits$

$H(B) = H(0.38,0.62) = 0.96\ bits$

The entropy of A is lower, and therefore there is less uncertainty regarding its spiking rate. Yet, the entropy of B is closer to uniform distributions.

The conditional entropy:

$$H(B|A) = \sum_{A} P(A) \cdot H(B|A)$$

$= 0.7_{A=40} \cdot H(0.2,0.8) + 0.3_{A=80} \cdot H(0.2,0.8) = H(0.2,0.8) = 0.72\ bits$

The mutual information:

$I(B;A) = H(B) - H(B|A) = 0.24\ bits$

Or:

$H(B,A) = H(0.14,0.56,0.24,0.06) = 1.60\ bits$

$I(B;A) = H(A) + H(B) - H(B,A) = 0.24\ bits$

The joint probability is calculated by:

$rate\ A$	$rate\ B$	P
40	20	$0.7 \cdot 0.2 = 0.14$
40	40	0.56
80	20	0.24
80	40	0.06

The mutual information indicates dependency of the two neurons one on the other.

Class Exercise:

The following table describes the probability of a patient coming into the clinic to be suffering from either Tourette syndrome (TS) or obsessive compulsive disorder (OCD):

TS	OCD	P
-	-	0.9
-	+	0.06
+	-	0.02
+	+	0.02

a. Calculate the entropy of TS and OCD and their joint entropy, explain the result.
b. Calculate the mutual information between the two disorders, explain the resul

Solution:

$P(TS = +) = 0.02 + 0.02 = 0.04\ bits$

$H(TS) = H(0.04,0.96) = 0.24\ bits$

$P(OCD = +) = 0.08\ bits$

$H(OCD) = H(0.08,0.92) = 0.4\ bits$

We can notice that H(OCD) is higher because it is more uniform than the TS. To calculate the joint entropy we can use the table:

$$H(TS, OCD) = -\sum P(TS_i, OCD_j) \cdot log_2\left(P(TS_i, OCD_j)\right)$$
$$= -0.9 \cdot log_2 0.9 - 0.06 \cdot log_2 0.06 - 2 \cdot 0.02 \cdot log_2 0.02 = 0.6\ bits$$

So that the mutual information is:

$$I(TS; OCD) = H(TS) + H(OCD) - H(TS, OCD) = 0.04 \text{ bits}$$

Or:

$$I(TS; OCD) = H(TS) - H(TS|OCD)$$

$$H(TS|OCD) = P(OCD = +) \cdot H\left(\frac{0.02}{0.08}, \frac{0.06}{0.08}\right) + P(OCD = -) \cdot H\left(\frac{0.9}{0.92}, \frac{0.02}{0.92}\right)$$
$$= 0.06 + 0.14 = 0.2 \text{ bits}$$

$$I(TS; OCD) = H(TS) - H(TS|OCD) = 0.24 - 0.2 = 0.04 \text{ bits}$$

Almost no mutual information between the two distributions, therefore they are independent.

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Class Exercise:

The following table describes the probability a patient Testis Nervosis suffers from two symptoms (S1 & S2).

$S1$	$S2$	P
-	-	0.4
-	+	0.3
+	-	0.1
+	+	0.2

Two random variables are constructed:

X – has 4 possible values based on the S1 & S2 symptom occurance.

Y – has 3 possible values based on the number of symptoms occurring (0,1,2)

a. Calculate the entropy of X and the entropy of Y, explain the results.
b. Calculate the mutual information between the variables, explain the result.

Solution:

We'll update the table:

$S1$	$S2$	P	X	Y
-	-	0.4	1	0
-	+	0.3	2	1
+	-	0.1	3	1
+	+	0.2	4	2

Therefore the entropies are:

$H(X) = H(0.4,0.3,0.2,0.1) = 1.84$ bits

$H(Y) = H(0.4,0.3 + 0.1,0.2) = 1.5$ bits

The maximal entropy for Y is for a uniform distribution $-\log_2\left(\frac{1}{3}\right) = 1.58$, so the calculated H(Y) is very close to the maximal entropy.

We can now notice that the distribution of X contains Y because if we know for every subject where he is in X than we know for certain what is his Y. Therefore:

$$H(X;Y) = H(X)$$

And the mutual information is:

$$I(X;Y) = H(X) + H(Y) - H(X;Y) = H(Y) = 1.5 \text{ bits}$$

Or:

$$H(Y|X) = 0$$

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) = 1.5 \text{ bits}$$

And also the mutual information predicts full dependency of Y on X.

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Class Exercise:

Roughly 10% of Parkinson's disease patients are early onset (i.e. under the age of 50). A late onset parkinsonian patient displays primarily either akinesia (25%) or tremor (75%) symptoms. However, early onset patients display a high percentage (90%) of akinesia associated symptoms.

a. Calculate the entropy of the symptoms.
b. Calculate the conditional entropy of the symptoms given the age of onset.
c. What is the mutual information between the age and symptom?

Solution:

a. We will first calculate the probability of akinesia:

$p(akinesia) = 0.9 \cdot 0.25 + 0.1 \cdot 0.9 = 0.315$

Now we need the calculate the entropy of all possible states in Parkinson, with akinesia and without:

$H(symptoms) = H(0.315,0.685) = 0.8989\ bits$

b. $H(symptoms|age) = 0.1 \cdot H(0.9,0.1) + 0.9 \cdot H(0.25,0.75) = 0.777\ bits$

c. $I(symptoms; age) = H(symptoms) - H(symptoms|age) = 0.8989 - 0.777 = 0.1219\ bits$

Therefore akinesia symptoms are almost independent of age.

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Class Exercises for self-learning:

Exercise 1: Gruesome Gnomes have the ability to walk either upright (p0(s=U) = 0.75) or upside down (p0(s=D)=0.25). Two neurons (N1 and N2) in the Gnome's vestibular system of the fire differently when the Gnome is standing upright (U) or upside down (D).

%	RN1=100 spikes/s	RN1=0 spikes/s
U	80	20
D	30	70

%	RN2=100 spikes/s	RN2=0 spikes/s
U	90	10
D	50	50

a. Calculate the entropy of the state of the Gnome.
b. Which of the neurons provide more information about the state of the Gnome? Calculate the information that it provides on the state.

%	RN1=100 RN2=100	RN1=0 RN2=100	RN1=100 RN2=0	RN1=0 RN2=0
U	80	10	0	10
D	30	20	0	50

c. If the neurons display the joint probability shown in the table above. What is their joint entropy? Compare to the joint entropy in the independent case and explain.

Exercise 2: A neuron responds to different stimuli (stimulus A or B) with a burst of 0-3 spikes at times 10ms, 20ms and 30ms after the stimulus, according to the table below:

10ms	0	0	0		0	1		1	1	1
20ms	0	0	1		1	0		0	1	1
30ms	0	1	0		1	0		1	0	1
Stimulus	A	A	A	B	A	A	B	B	B	B
P	0.05	0.2	0.05	0.05	0.05	050.	050.	20.	0.1	0.2

a. Calculate the entropies of (i) the stimulus (1 point) and (ii) the spike count.
b. Calculate the conditional entropy of the spike count given the spike pattern.
c. Calculate which spike time (10 or 30 ms) gives the most information about the stimulus?

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Short solutions to exercises for self learning:

Exercise 1

a.

H(state) = H(0.75, 0.25) = 0.8113 bits

b.

p(Rn1 = 100) = 0.8*0.75+0.3*0.25 = 0.675

H(Rn1) = H(0.675, 0.325) = 0.9097

H(Rn1|state) = 0.75*H(0.8, 0.2) + 0.25*H(0.3, 0.7) = 0.7618

I(Rn1, state) = H(Rn1) – H(Rn1|state) = 0.9097 - 0.7618 = 0.1479 bits

(Rn2 = 100) = 0.9*0.75+0.5*0.25 = 0.8

H(Rn2) = H(0.8, 0.2) = 0.7219

H(Rn2|state) = 0.75*H(0.9, 0.1) + 0.25*H(0.5, 0.5) = 0.6017

I(Rn2, state) = H(Rn2) – H(Rn2|state) = 0. 7219- 0. 6017= 0.1202 bits

Neuron 1 provides more information

c.

The joint probability of Rn1 and Rn2:

0.75*0.8+0.25*0.3 = 0.675

0.75*0.1+0.25*0.2=0.125

0.75*0 +0.25*0=0

0.75*0.1+0.25*0.5=0.2

H(Rn1, Rn2) = H(0.675, 0.125, 0, 0.2) = 1.2221 bits

In the independent case:

H(Rn1, Rn2) = H(Rn1) + H(Rn2) = 0. 9097+ 0. 7219= 1.6316 bits

Written by Yarden Nativ, mail: nativyardenac@gmail.com

Exercise 2

a.

stimulus entropy = H(0.4, 0.6) = 0.971 bits

Spike count entropy = H(0.05, 0.4, 0.35, 0.2) = 1.7394 bits

b. H(spike count| spike pattern) = 0 bits

c.

H(S|spikeTime = 10ms)

= p(no spike at 10ms)*H(S|no spike at 10ms) + p(spike at 10ms)*H(S|spike at 10ms)

= 0.4*H(0.875, 0.125) + 0.6*H(0.9167, 0.0833)

= 0.4*0. 5436 + 0.6*0. 4137 = 0.4657

I(S, spike time at 10ms) = H(S) - H(S|spikeTime = 10ms) = 0.9710-0.4657=0.5053 bits

H(S|spikeTime = 30ms)

= p(no spike at 30ms)*H(S|no spike at 30ms) + p(spike at 30ms)*H(S|spike at 30ms)

= 0.35*H(0.4286, 0.5714) + 0.65*H(0.3846, 0. 6154)

= 0.35*0.9852+ 0.65*0.9612= 0.9696

I(S, spike time at 10ms) = H(S) - H(S|spikeTime = 10ms) = 0.9710-0.9696=0.0014 bits