THEORETICAL NEUROSCIENCE I

Lecture 16: Shannon information basics

Prof. Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

Outline

1. The story so far ...

2. Intuitive motivation: halving the number of remaining possibilities.

3. Shannon information and Shannon entropy of discrete random events.

4. Example: sinking submarines.

5. Mutual information of dependent random variables.

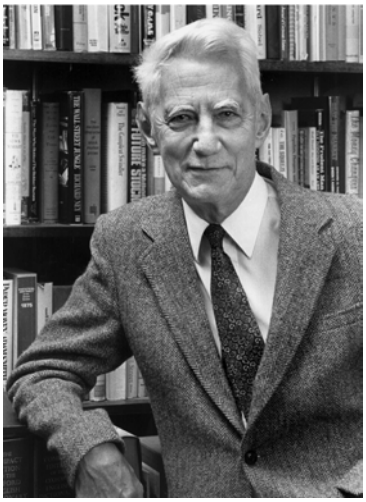


Figure 1: Claude Shannon (1916-2001). [1]

Claude Shannon (1916-2001)

1937 MSc. creates basis for digital circuits and computers.
1940 PhD on theoretical genetics
1942-1945 fire-control systems and cryptography
1948 "A Mathematical Theory of Communication"
Other interests: juggling, unicycling, chess, powered pogo stick, wearable roulette computer, flame throwing trumpet, and more.

1 Intuitive motivation

Shannon information is the minimal number of binary questions that are needed to identify the outcome of a discrete random event.

Being entirely general, Shannon information lets us compare all kinds of random events and processes.

It applies equally to physics, sociology, economics, cryptography, neuroscience, and more ... Even thermodynamical entropy can be subsumed.
Why does this offer a reasonable quantification of information?

Game of 20 questions

In the game of twenty questions, one player thinks of an object, and the other player attempts to guess what it is by asking questions that have yes/no answers. The aim is to identify the object with as few questions as possible.

What is the best strategy for playing this game?

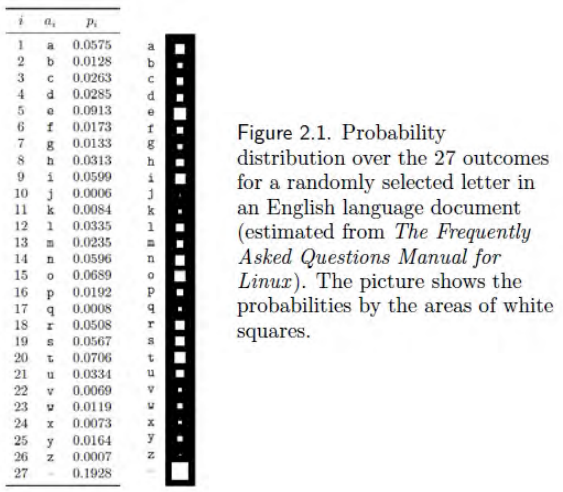
Should you ask 'Is it a Costa Rican parakeet?'.

Should you ask 'Is it living or non-living?'

English characters

How informative are English characters? Knowing one character of the word, how much closer are you to knowing the word?

Are all characters equally informative?



i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

Figure 2: Probability distributions. [2]

Game of 12 balls

You are given 12 identical-looking balls. 11 are equal in weight, but one is lighter or heavier than the others. You are also provided with a scale. What's the most direct way of identifying the odd ball?

Figure 3: Balance and 12 balls. [3]

- Consider information gain in terms of a *multiplicative reduction* of possibilities!
- What *multiplicative reduction* can be achieved by using the balance once?

- What can be achieved by placing *one* ball on each side?
- By placing *six* balls on each side?
- By placing *four* balls on each side?

1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+
1- 2- 3- 4- 5- 6- 7- 8- 9- 10- 11- 12-

Designing informative experiments

- The number of possible states of the world is 24 (which of 12 balls is odd *and* whether it is lighter or heavier).
- Each weighing has three possible outcomes: 'left', 'right', 'balanced'.
- Three successive weighings have $3^3 = 27$ possible outcomes.
- Thus, in principle, three weighings should suffice to determine which ball is odd *and* whether it is lighter or heavier.
- How can you gain information as quickly as possible?
- Choose strategy such that, at each weighing, the three outcomes are as equiprobable as possible!

Game of 63 (optional)

For simplicity, imagine that we are playing a (far duller) game called 'sixty-three':

What is the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

One of several equally good strategies is:

$$1: \qquad is\ x \geq 32?$$

$$
\begin{aligned}
2: &\quad is\ x \bmod 32 \geq 16? \\
3: &\quad is\ x \bmod 16 \geq 8? \\
4: &\quad is\ x \bmod 8 \geq 4? \\
5: &\quad is\ x \bmod 4 \geq 2? \\
6: &\quad is\ x \bmod 2 = 1?
\end{aligned}
$$

In general, we need to halve the number of remaining possibilities six times!

Summary intuitive motivation

- Shannon information is the minimal number of binary questions needed to determine the outcome of a discrete random event.
- Shannon information provides a common yardstick with which all kinds of random events and processes can be compared.
- It does this by quantifying the multiplicative reduction in uncertainty (remaining number of possibilities).
- 1 bit = halving the number of possibilities.

2 Shannon information and and Shannon entropy of discrete random events

We wish to quantify the informativeness of **discrete random events**.

- The informativeness of individual events should reflect their rarity or surprise value.

  $\Rightarrow$ measure should decrease monotonically with probability!

- For multiple *independent* events, the joint informativeness should equal the sum of individual informativeness.

  $\Rightarrow$ measure should be logarithmic!

- Intuitive definition in terms of halving possibilities.

  $\Rightarrow$ Choose $\log_2$.

'Shannon information' of individual random events

The 'information' (Shannon information content or SIC) of an individual random event x *decreases* with the *binary logarithm* of its *probability*.

It is defined as

$$h(x) = \log_2 \frac{1}{P(x)} = -\log_2 P(x)$$

where $P(x)$ is the probability of x.

Its unit is called 'bits'.

Example: ordinary coin

When you throw an ordinary coin, the probability of each outcome is $\frac{1}{2}$. Thus, the SIC of each outcome is

$$h = -\log_2 \frac{1}{2} = \log_2 2 = 1$$

With each throw, you gain $1\,bit$ of information (in Shannon's sense of the term).

Example: N-sided dice

When you throw an N-sided die, the probability of each outcome is $\frac{1}{N}$, so that the SIC of each outcome is

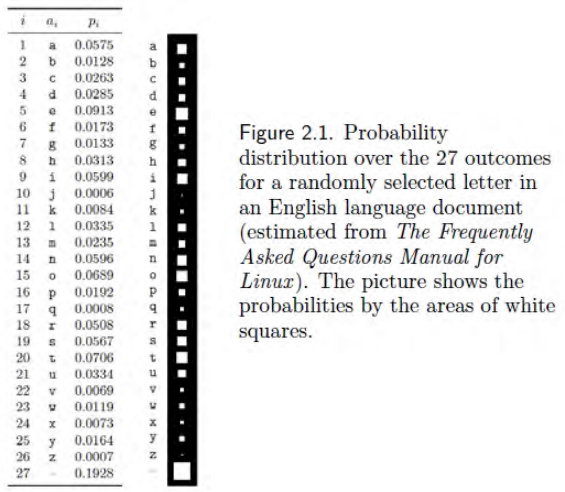
$$h = -\log_2 \frac{1}{N} = \log_2 N$$

With each throw, you gain $\log_2 N\,bits$ of information.

Figure 4: N-sided dice.

Example: English characters

When a random character x is picked from an English document, the SIC of the 27 possible outcomes is listed here. The outcome $x = \mathbf{z}$ yields $10.4\,bits$, while $x = \mathbf{e}$ yields $3.5\,bits$.



i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

Figure 5: Probability distributions. [2]

'Shannon entropy' of random variable or process

The 'entropy' of a random variable X is defined as the average Shannon information content of all possible outcomes x.

$$H(X) = \sum_x P(x)\, h(x) \qquad \textit{(probabilistic average)}$$

where each outcome x contributes entropy $P(x)\, h(x)$. Since

$$h(x) = -\log_2 P(x) = \log_2 \frac{1}{P(x)}$$

we have

$$H(X) = -\sum_x P(x)\, \log_2 P(x) = \sum_x P(x)\, \log_2 \frac{1}{P(x)}$$

The unit of entropy is 'bits'.

Example: ordinary coin

The entropy of the random variable realized by throwing an ordinary coin is

$$H = \sum_x P(x)\, h(x) = \sum_x P(x)\, \log_2 \frac{1}{P(x)}$$

The entropy contribution of each outcome is

$$P(x)\,h(x) = \frac{1}{2}\,\log_2 2$$

Both contributions together add up to

$$H = \sum_x \frac{1}{2}\,\log_2 2 = 2\,\frac{1}{2}\,\log_2 2 = 1$$

The entropy is the average information of individual outcome. Since every outcome is equally informative, the average information equals the individual information. Thus, the entropy is identical to the information of each individual outcome.

Example: N-sided dice

The entropy of the random variable realized by throwing an N-sided die is

$$H = \sum_x P(x)\,h(x) = \sum_x P(x)\,\log_2 \frac{1}{P(x)}$$

The entropy contribution of each outcome is

$$P(x)\,h(x) = \frac{1}{N}\,\log_2 N$$

N contributions together add up to

$$H = \sum_x \frac{1}{N}\,\log_2 N = N\,\frac{1}{N}\,\log_2 N = \log_2 N$$

Certainty and impossibility

Neither certain nor impossible outcomes contribute to the entropy:

$$P(x) = 1 \qquad \Rightarrow \qquad 1 \cdot \log_2 \frac{1}{1} = 0$$

$$P(x) = 0 \qquad \Rightarrow \qquad 0 \cdot \log_2 \frac{1}{0} = 0$$

Uncertainty

In contrast, any uncertain outcome contributes positively to the entropy:

$$0 < P(x) < 1 \qquad \Rightarrow \qquad P \cdot \log_2 \frac{1}{P} > 0$$

Entropy contributions and probability

Whereas Shannon information decreases with probability, the contribution to Shannon entropy peaks at a certain probability value ($P \approx 0.37$)

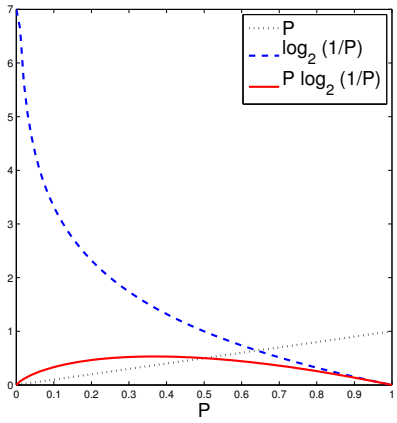


Figure 6: Shannon information and entropy.

Upper limit of Shannon entropy

Given an ensemble X with individual outcomes x and probabilities $P(x)$. How should we choose the $P(x)$ such as to maximize the Shannon entropy?

Can we choose $P(x) = 0.37$ for all outcomes x?

Can we choose $P(x) = 0.37$ for two outcomes $x_{1,2}$ and smaller values for all others?
Perhaps there is a more principled way?

Equiprobable ensembles

Ensemble entropy is maximal when all N outcomes are equally probable. Thus, entropy has the upper bound

$$H(X) \leq \log_2 N$$

$$H(X) = \log_2 N \qquad \text{iff} \qquad P(x) = \frac{1}{N} \text{ for all } x$$

since, in this case,

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)} = -\sum_x \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N$$

Example: 2 outcomes

Two outcomes, $(+)$ and $(-)$, with probabilities:

$$P_+, \; P_- \qquad\qquad P_- = 1 - P_+$$

$$H = \sum_x P(x) \log_2 \frac{1}{P(x)} = P_+ \log_2 \frac{1}{P_+} + (1 - P_+) \log_2 \frac{1}{1 - P_+}$$

$H = f(P_+)$ is maximal when $P_+ = 1/2$, in other words, when outcomes are equiprobable!
$H = f(P_+)$

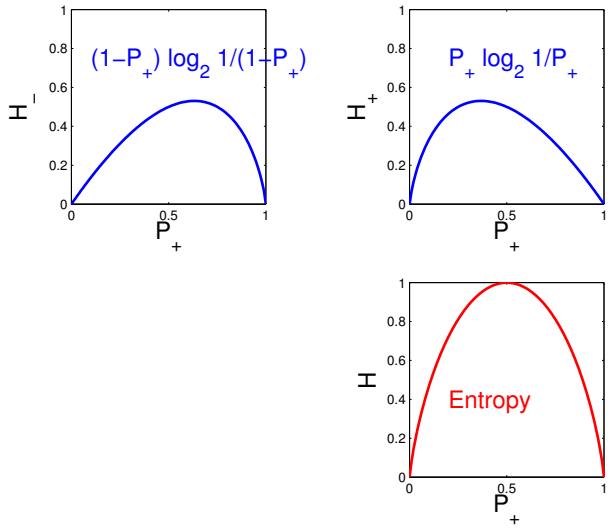


Figure 7: Plots of outputs.

$H = f(P_+)$ is maximal for equiprobable outcomes $P_+ = P_- = 1/2$.

N outcomes (optional)

Of N outcomes

$$P_1, P_2, \ldots, P_N$$

consider two particular outcomes 1 and 2. Compare the case $P_1 = P_2 = P$

$$H_1 + H_2 = P \log_2 \frac{1}{P} + P \log_2 \frac{1}{P}$$

to the case $P_1 = P + \epsilon$ and $P_2 = P - \epsilon$

$$H_1' + H_2' = (P + \epsilon) \log_2 \frac{1}{P + \epsilon} + (P - \epsilon) \log_2 \frac{1}{P - \epsilon}$$

by forming the difference of entropies

$$\Delta H = H_1' - H_1 + H_2' - H_2$$

$$\begin{aligned}
\Delta H &= P\,\log_2\left(\frac{P+\epsilon}{P}\right) + P\,\log_2\left(\frac{P-\epsilon}{P}\right) + \epsilon\,\log_2\left(\frac{P-\epsilon}{P+\epsilon}\right) = \\
&= P\,\log_2\left(1+\frac{\epsilon}{P}\right) + P\,\log_2\left(1-\frac{\epsilon}{P}\right) + \epsilon\,\log_2\left(1-\frac{2\epsilon}{P+\epsilon}\right) \approx \\
&\approx -\frac{2\epsilon^2}{P+\epsilon}
\end{aligned}$$

where we have used $\log(1+x) \approx x$ for small x.

Thus, any difference ϵ in the probabilities decreases the ensemble entropy! It follows that ensemble entropy is maximal when all outcomes are equiprobable.

Summary 'information' and 'entropy' of discrete random variables

- The 'Shannon information' $h(x)$ of a random event x depends on its probability $P(x)$:

$$h(x) = -\log_2\, P(x)$$

- The 'Shannon entropy' $H(X)$ of a random variable or process is the average information of its events:

$$H(X) = -\sum_x\, P(x)\, h(x) = -\sum_x\, P(x)\, \log_2\, P(x)$$

- The unit of information and entropy is 'bits'.
- The maximal entropy of a random variable with N discrete outcomes is

$$H(X) \leq \log_2 N$$

3 Sinking submarines

In the game of battleships, each player hides a fleet of ships in a sea represented by a square grid. On each turn, one player attempts to hit the other's ships by firing at one square in the opponent's sea. The response to a targeted square such as 'G3' is either 'miss', 'hit', or 'destroyed'.

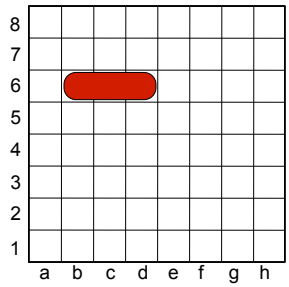


Figure 8: 8 x 8 grid, where the ships are placed.

In a less exciting version called *submarine*, each player hides a submarine in just one square of an **eight-by-eight** grid.

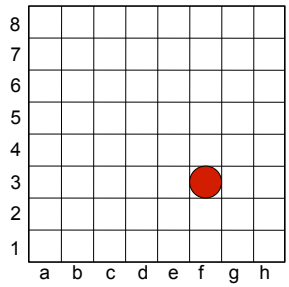


Figure 9: 8 x 8 grid, where the submarines are placed.

At each turn, the two possible outcomes are $\{y, n\}$, corresponding to hit and miss. The probabily $P(x)$ of a hit depends on the state of the board (how many squares are still untried).

Example runs

Given an eight-by-eight grid, the chances for our first shot are $P(y) = 1/64$ and $P(n) = 63/64$.

If we are lucky enough to hit the submarine with the first shot, we have gained

$$h(x) = h_{(1)}(y) = \log_2 64 = 6 \,\text{bits}$$

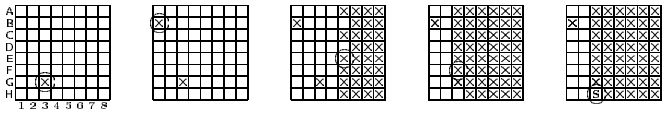
of information and finish the run.

Note that we have reduced the number of possibilities 64-fold, which corresponds to 6 consecutive 2-fold reductions. Hence we have gained 6 bits.
If the first shot misses, the chances for the second shot are $P(y) = 1/63$ and $P(n) = 62/63$. If the second misses, too, the chances for the third shot are $P(y) = 1/62$ and $P(n) = 61/62$. Even though our first shot missed, we have still gained

$$h(x) = h_{(1)}(n) = \log_2 \frac{64}{63} = 0.0227 \,\text{bits}$$

The second miss gained us

$$h_{(2)}(n) = \log_2 \frac{63}{62} = 0.0230 \,\text{bits}$$



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

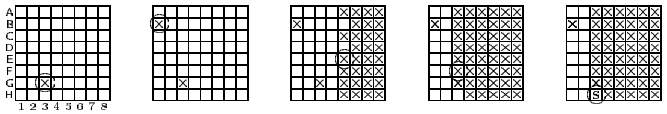
Figure 10: Example information of Sinking submarines.

After 32 consecutive misses (firing at a new square each time), the total information gained is

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{33}{32} = \log_2 \frac{64}{32} =$$

$$= 0.0227 + 0.0230 + \ldots + 0.0430 = 1 \,\mathrm{bits}$$

corresponding to a 2-fold reduction of the remaining possibilities ($2^1 = 2$).



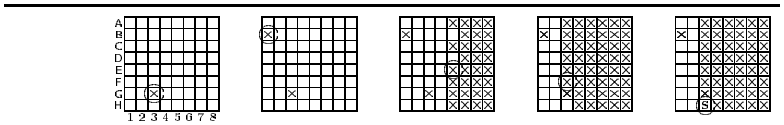
move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{n}$	$x = \mathrm{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Figure 11: Example information of Sinking submarines.

After 48 unsuccessful shots, the information gained is 2 bits:

$$\log_2 \frac{64}{63} + \ldots + \log_2 \frac{17}{16} = \log_2 \frac{64}{16} = 2 \,\mathrm{bits}$$

corresponding to a 4-fold reduction of the remaining possibilities ($2^2 = 4$): The unknown location has been narrowed down to one quarter of the original number of possibilities.



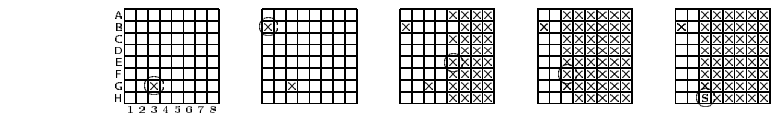
move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Figure 12: Example information of Sinking submarines.

What if we hit on the 49th shot, when there were 16 squares let? The Shannon information gained is

$$h_{(49)} = \log_2 16 = 4 \,\mathrm{bits}$$

corresponding to a 16-fold reduction of the remaining possibilities ($2^4 = 16$).



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Figure 13: Example information of Sinking submarines.

Total information gained

What is the total information gained by 48 consecutive misses and a hit with the 49^{th} shot?

$$\sum_{i=1}^{48} h_{(i)} = 2\,\text{bits}, \qquad\qquad h_{(49)} = 4\,\text{bits}$$

so that

$$\sum_{i=1}^{49} h_{(i)} = 6\,\text{bits}$$

This result holds regardless of when we happen to hit the submarine! We always gain 6 bits, because we always end up with a 64-fold reduction of possibilities ($2^6 = 64$).

Summary sinking submarines

- The game of sinking submarines illustrates the consistency of Shannon's measure of information.
- Some shots (hits) gain us far more information than others (misses).
- In the end, however, we must gain 6 bits of information to sink the submarine, no matter how many shots it takes.

4 Mutual information of dependent random variablesl

Finally, we consider *joint ensembles* formed by two random variables. In general, such variables may exhibit some degree of mutual dependence, indicating shared information.

We wish to quantify such dependencies (shared information) in terms of Shannon information/entropy.

Once again we take a strictly statistical view, avoiding any consideration of mechanism or physical causation.

A noisy channel with input x and output y constitutes a *joint ensemble*. How much information does y reveal about x?

A neural response r to a sensory event s also constitutes a *joint ensemble*. How much does r reveal about s?

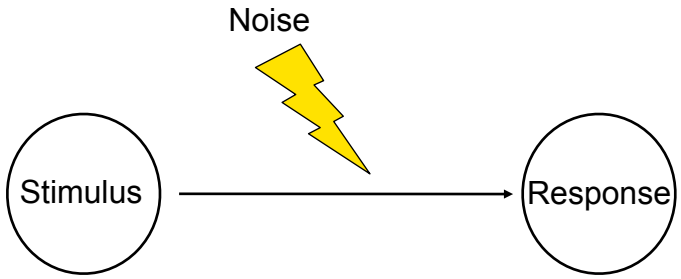


Figure 14: Response of the stimulus affected by noise.

Joint entropy and individual entropy

The *joint* entropy of two random variables X and Y is defined as

$$H(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{1}{P(x,y)}$$

It compares to the *individual* or *marginal* entropies of X, and of Y,

$$H(X) = \sum_x P(x)\, \log_2 \frac{1}{P(x)}, \qquad\qquad H(Y) = \sum_y P(y)\, \log_2 \frac{1}{P(y)}$$

Independent variables

In the case of *independent* variables, the joint entropy equals the *sum* of the individual entropies:

$$P(x,y) = P(x)\, P(y) \quad \Rightarrow \quad H(X,Y) = H(X) + H(Y)$$

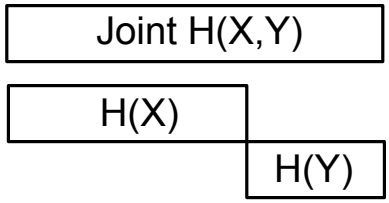


Figure 15: Joint entropy equals the *sum* of the individual entropies.

Proof:

$$\begin{aligned} H(X,Y) &= \sum_{xy} P(x)\, P(y)\, \log_2 \frac{1}{P(x)\, P(y)} = \\ &= \sum_y P(y) \sum_x P(x)\, \log_2 \frac{1}{P(x)} + \\ &+ \sum_x P(x) \sum_y P(y)\, \log_2 \frac{1}{P(y)} = \\ &= H(X) + H(Y) \end{aligned}$$

Example independent variables

A joint ensemble has the joint distribution:

$P(x,y)$ x			
		$+$	$-$
y	$+$	$1/4$	$1/4$
	$-$	$1/4$	$1/4$

$$H(X) = H(Y) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1$$

$$H(X,Y) = \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = 2$$

Overlap:

$$I_m = H(X) + H(Y) - H(X,Y) = 1 + 1 - 2 = 0$$

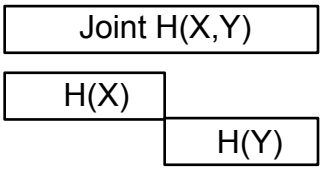


Figure 16: Overlap.

Fully dependent variables

In the case of *deterministically* dependent variables, the joint entropy *equals* each individual entropy:

$$P(X,Y) = P(X) = P(Y) \quad \Rightarrow \quad H(X,Y) = H(X) = H(Y)$$

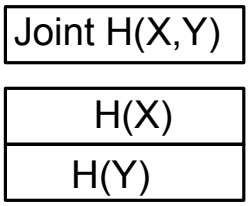


Figure 17: The joint entropy *equals* each individual entropy.

Proof:

$$H(X,Y) = \sum_{xy} P(x,y) \log_2 \frac{1}{P(x,y)} =$$

$$= \sum_{x} P(x) \log_2 \frac{1}{P(x)} =$$

$$= \sum_{y} P(y) \log_2 \frac{1}{P(y)}$$

Example of fully dependent variables

Another joint ensemble has the joint distribution:

$P(x,y)$ x			
		$+$	$-$
y	$+$	0	$1/2$
	$-$	$1/2$	0

$$H(X) = H(Y) = \frac{1}{2}\,\log_2\,2 + \frac{1}{2}\,\log_2\,2 = 1$$

$$H(X,Y) = \frac{1}{2}\,\log_2\,2 + \frac{1}{2}\,\log_2\,2 = 1$$

$$I_m = H(X) + H(Y) - H(X,Y) = 1 + 1 - 1 = 1$$

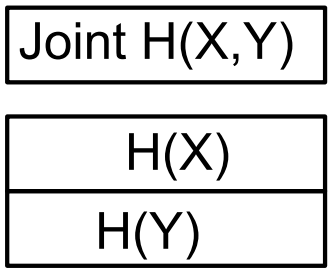


Figure 18: Fully independent variables.

Partly dependent variables

In the case of *stochastically* dependent variables, the joint entropy is *less* than the sum of the individual entropies:

$$P(X,Y) \neq P(X) \cdot P(Y) \quad \Rightarrow \quad H(X,Y) < H(X) + H(Y)$$

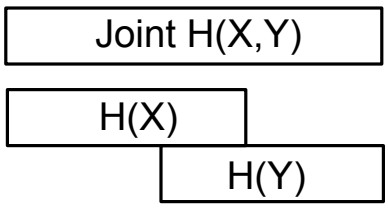


Figure 19: Partly dependent variables.

Example of partially dependent variables

Another joint ensemble has the joint distribution:

$P(x,y)$ x			
		$+$	$-$
y	$+$	$1/8$	$3/8$
	$-$	$3/8$	$1/8$

$$H(X) = H(Y) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1$$

$$H(X,Y) = \frac{2}{8}\log_2 8 + \frac{6}{8}\log_2 \frac{8}{3} = \frac{3}{4} + \frac{9}{4} - \frac{3}{4}\log_2 3 = 1.81$$

$$I_m = H(X) + H(Y) - H(X,Y) = 1 + 1 - 1.81 = 0.19$$

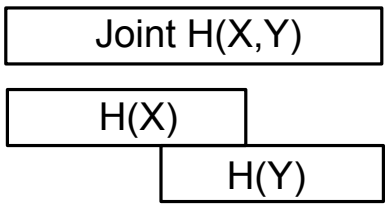


Figure 20: Partly dependent variables.

Mutual information

The 'overlap', or shared information, is called the 'mutual information'.

It quantifies how much information about the other variable is revealed by knowing the value of one variable.

It is necessarily symmetric!

Summary mutual information

- Given two random variables, it is useful to compare joint entropy and individual entropies.
- If combined individual entropies exceed joint entropy, the excess is 'mutual information' .
- MI is information shared between variables, revealed by one about the other, uncertainty reduced about the other (as shown in next lecture).

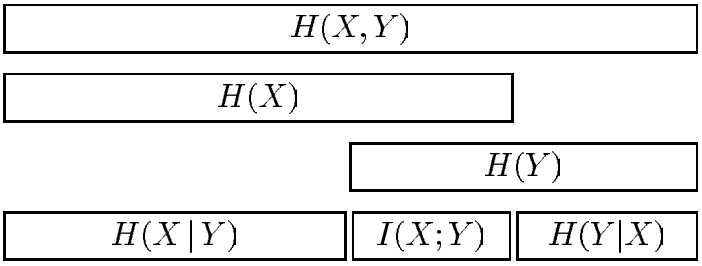


Figure 21: Joint entropy and individual entropies.

Overall summary

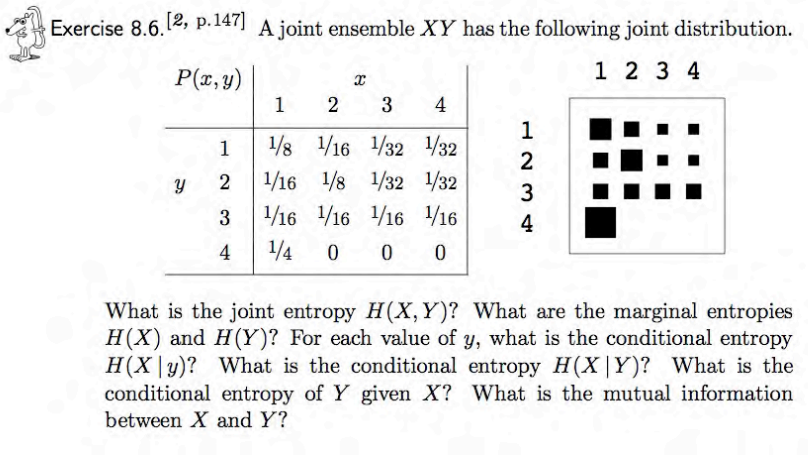
1. Intuitive motivation: halving the number of remaining possibilities.

2. Information and entropy of discrete random variables.

3. Example: sinking submarines.

4. Mutual information of dependent random variables.

Appendix A: ITILA example



Exercise 8.6.[2, p.147] A joint ensemble XY has the following joint distribution.

$P(x,y)$		x = 1	2	3	4
y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

What is the joint entropy $H(X,Y)$? What are the marginal entropies $H(X)$ and $H(Y)$? For each value of y, what is the conditional entropy $H(X\,|\,y)$? What is the conditional entropy $H(X\,|\,Y)$? What is the conditional entropy of Y given X? What is the mutual information between X and Y?

Figure 22: Example exercise. [4]

An ensemble of random variables (x, y) has the following marginal and joint probabilities P:

P(x,y)		P(x)			
		1/2	1/4	1/8	1/8
	1/4	1/8	1/16	1/32	1/32
P(y)	1/4	1/16	1/8	1/32	1/32
	1/4	1/16	1/16	1/16	1/16
	1/4	1/4	0	0	0

Information gained h

h(x,y)		h(x)			
		1	2	3	3
	2	3	4	5	5
h(y)	2	4	3	5	5
	2	4	4	4	4
	2	2	0	0	0

Information h

$$h(x) = \log_2 \frac{1}{P(x)}, \qquad h(y) = \log_2 \frac{1}{P(y)}, \qquad h(x,y) = \log_2 \frac{1}{P(x,y)}$$

Entropy contributions $P \cdot h$

P(x,y)	h(x,y)		P(x)	h(x)		
			1/2	2/4	3/8	3/8
		2/4	3/8	4/16	5/32	5/32
P(y)	h(y)	2/4	4/16	3/8	5/32	5/32
		2/4	4/16	4/16	4/16	4/16
		2/4	2/4	0	0	0

Individual and joint entropies

$$H(X) = \sum_x P(x)\, h(x) \qquad\qquad H(Y) = \sum_y P(y)\, h(y)$$

$$H(X,Y) = \sum_{xy} P(x,y)\, h(x,y)$$

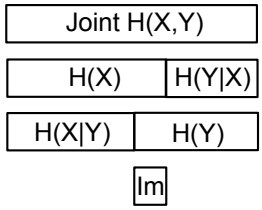
$$H(X) = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1\,\frac{3}{4} \qquad\qquad H(Y) = \frac{2}{4} + \frac{2}{4} + \frac{2}{4} + \frac{2}{4} = 2$$

$$\begin{aligned} H(X,Y) \;&= \tfrac{3}{8} + \tfrac{4}{16} + \tfrac{5}{32} + \tfrac{5}{32} + \\ &+ \tfrac{4}{16} + \tfrac{3}{8} + \tfrac{5}{32} + \tfrac{5}{32} + \\ &+ \tfrac{4}{16} + \tfrac{4}{16} + \tfrac{4}{16} + \tfrac{4}{16} + \\ &+ \tfrac{2}{4} \qquad\qquad = 3\,\tfrac{3}{8} \end{aligned}$$

$$H(X,Y) = 3\,\frac{3}{8} \qquad H(X) = 1\,\frac{3}{4} \qquad H(Y) = 2$$

$$I_m = H(X) + H(Y) - H(X,Y) = \frac{3}{8}$$

$$H(X|Y) = H(X) - I_m = 1\,\frac{3}{8} \qquad H(Y|X) = H(Y) - I_m = 1\,\frac{5}{8}$$



$$H(X|Y) \;= \Sigma_y\; P(y)\; \Sigma_x\; P(x|y)\, h(x|y) =$$

$$\begin{aligned}
&= \tfrac{1}{4}\,\left[\tfrac{1}{2} + \tfrac{2}{4} + \tfrac{3}{8} + \tfrac{3}{8}\right] + \\
&+ \tfrac{1}{4}\,\left[\tfrac{2}{4} + \tfrac{1}{2} + \tfrac{3}{8} + \tfrac{3}{8}\right] + \\
&+ \tfrac{1}{4}\,\left[\tfrac{2}{4} + \tfrac{2}{4} + \tfrac{2}{4} + \tfrac{2}{4}\right] + \\
&+ \tfrac{1}{4}\,\left[\tfrac{0}{1} \qquad\qquad\qquad\quad\right] = 1\,\tfrac{3}{8}
\end{aligned}$$

$$H(Y|X) \;= \Sigma_x\; P(x)\; \Sigma_y\; P(y|x)\, h(y|x) =$$

$$\begin{aligned}
&= \tfrac{1}{2}\,\left[\tfrac{2}{4} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{2}\right] + \\
&+ \tfrac{1}{4}\,\left[\tfrac{2}{4} + \tfrac{1}{2} + \tfrac{2}{4} \qquad\right] + \\
&+ \tfrac{1}{8}\,\left[\tfrac{2}{4} + \tfrac{2}{4} + \tfrac{1}{2} \qquad\right] + \\
&+ \tfrac{1}{8}\,\left[\tfrac{2}{4} + \tfrac{2}{4} + \tfrac{1}{2} \qquad\right] = 1\,\tfrac{5}{8}
\end{aligned}$$

Average conditional information x

$P(x\|y)$		x			
		1	2	3	4
$P(y)$	1/4	1/2	1/4	1/8	1/8
	1/4	1/4	1/2	1/8	1/8
	1/4	1/4	1/4	1/4	1/4
	1/4	1	0	0	0

$P(x\|y)\ \log_2 {}^1/_{P(x\|y)}$		x			
		1	2	3	4
$P(y)$	1/4	1/2	2/4	3/8	3/8
	1/4	2/4	1/2	3/8	3/8
	1/4	2/4	2/4	2/4	2/4
	1/4	0/1	0	0	0

Average conditional information y

$P(y\|x)$		$P(x)$			
		1/2	1/4	1/8	1/8
y	1	1/4	1/4	1/4	1/4
	2	1/8	1/2	1/4	1/4
	3	1/8	1/4	1/2	1/2
	4	1/2	0	0	0

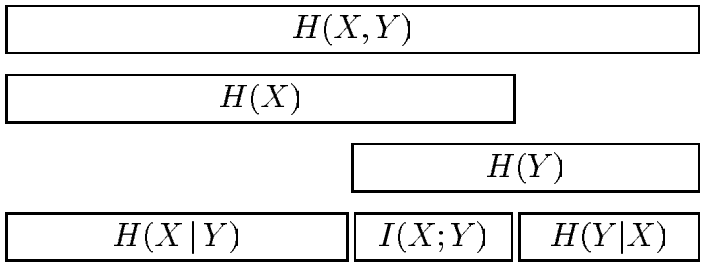
$P(y\|x)\ \log_2 {}^1/_{P(y\|x)}$		$P(x)$			
		1/2	1/4	1/8	1/8
y	1	2/4	2/4	2/4	2/4
	2	3/8	1/2	2/4	2/4
	3	3/8	2/4	1/2	1/2
	4	1/2	0	0	0

Summary ITILA example

$$H(X,Y) = 3\frac{3}{8} \qquad H(X) = 1\frac{3}{4} \qquad H(Y) = 2$$

$$H(Y|X) = 1\frac{5}{8} \qquad H(X|Y) = 1\frac{3}{8}$$

$$I_m = H(X) - H(X|Y) = H(Y) - H(Y|X) = \frac{3}{8}$$



Appendix B: Solution for 12 balls

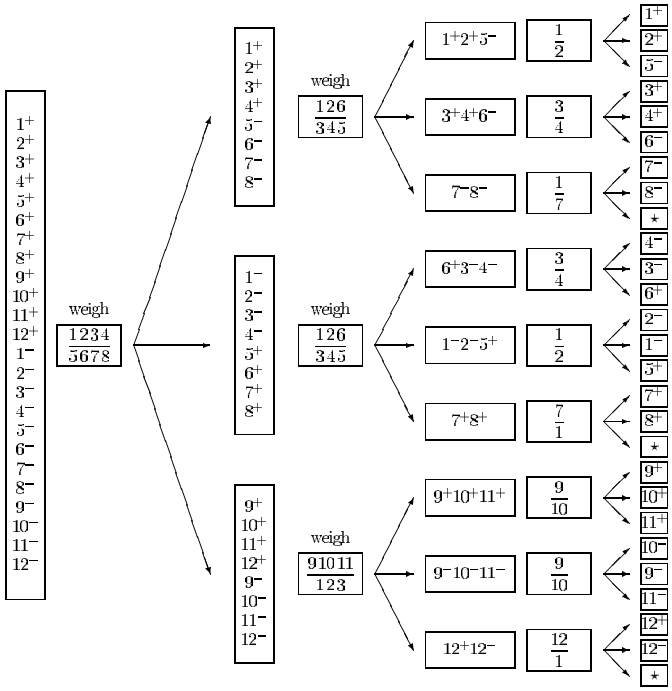


Figure 23: Solution for 12 balls. [5]

5 Bibliography

1. Physics Today, Claude Shannon, Anonymous, 2018, Image 1 Ref: https://physicstoday.scitation.org/do/10.1063/pt.6.6.20180430a/full/

2. David J.C. MacKay "Information Theory, inference, and learning algorithms." (ITILA), Figure 2.1 Ref: http://www.inference.phy.cam.ac.uk/mackay/itila/

3. TBooks, Why are checks and Balances Important in your Accounting System?, Teresa Sanders, 2019, Image 1 Ref: https://tbooks.com/checks-balances-important-accounting-system/

4. David J.C. MacKay "Information Theory, inference, and learning algorithms." (ITILA), Exercise 8.6 Ref: http://www.inference.phy.cam.ac.uk/mackay/itila/

5. David J.C. MacKay "Information Theory, inference, and learning algorithms." (ITILA), Figure 4.2 Ref: http://www.inference.phy.cam.ac.uk/mackay/itila/