Teoretične osnove računalništva 2
Teorija informacij

Entropy and information

Enes Pasalic UP FAMNIT študijsko leto 20/21

Lecture topics

- Entropy definition (discrete RV.):
  - Entropy of joint variables, conditional entropy, chain rules, properties of entropy
- Mutual information:
  - connection between entropy and mutual information
- Kullback-Leibler distance (relative entropy)
- Chain rules:
  - entropy
  - mutual information
- Markov chains
  - properties
- AEP:
  - examples,
  - Typical events,
  - coding using AEP,
  - Expected length of codewords using AEP coding

Entropy

Let X be a discrete random variable with **alphabet** $\mathcal{X}$ with probability mass function (pmf) p_X. The amount of **Shannon information** for $x \in \mathcal{X}$ is defined as:

$$I_X(x) = \log_2 \frac{1}{p_X(x)}.$$

It measures "surprise" since the smaller the value of $p_X(x)$ the larger is $I_X(x)$, thus more information is obtained form such an event !

Entropy for discrete random variables

Entropy of a discrete random variable X is the mathematical expectation of Shannon's information:

$$H(X) = E[I_X(x)] = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)}.$$

Entropy: example

Primer izračuna entropije pri znani porazdelitvi verjetnosti

$$X = \begin{cases} a & \text{z verjetnostjo } \frac{1}{2} \\ b & \text{z verjetnostjo } \frac{1}{4} \\ c & \text{z verjetnostjo } \frac{1}{8} \\ d & \text{z verjetnostjo } \frac{1}{8} \end{cases}$$

Entropija X je:

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}\text{bitov}$$

Example

Suppose we are watching cars going past on a highway. For simplicity, suppose 50% of the cars are black, 25% are white, 12.5% are red, and 12.5% are blue.

Consider the flow of cars as an information source with four words: black, white, red, and blue.

A simple way of encoding this source into binary symbols would be to associate each color with two bits, that is:
black = 00, white = 01, red = 10, and blue = 11,
an average of 2.00 bits per color.

A Better Code Using Information Theory

A better encoding can be constructed by allowing for the frequency of certain symbols, or words:
black = 0, white = 10, red = 110, blue = 111.

How is this encoding better?

0.50 black x 1 bit = .500
0.25 white x 2 bits = .500
0.125 red x 3 bits = .375
0.125 blue x 3 bits = .375

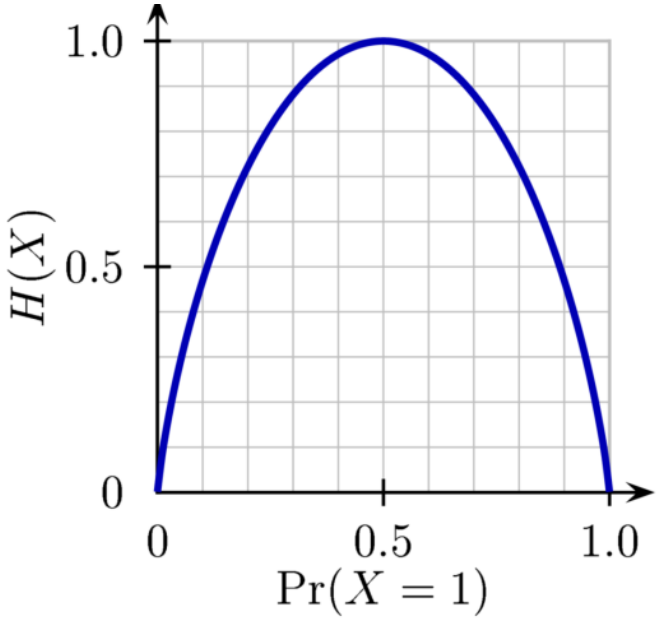
Average-- 1.750 bits per car

Entropy for binary random variables

Entropija binarne porazdelitve

Za binarno porazdelitev spr. X velja $\mathcal{X} = \{0, 1\}$ in $p_X(1) = p, p_X(0) = 1 - p$.

$$H(X) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}$$

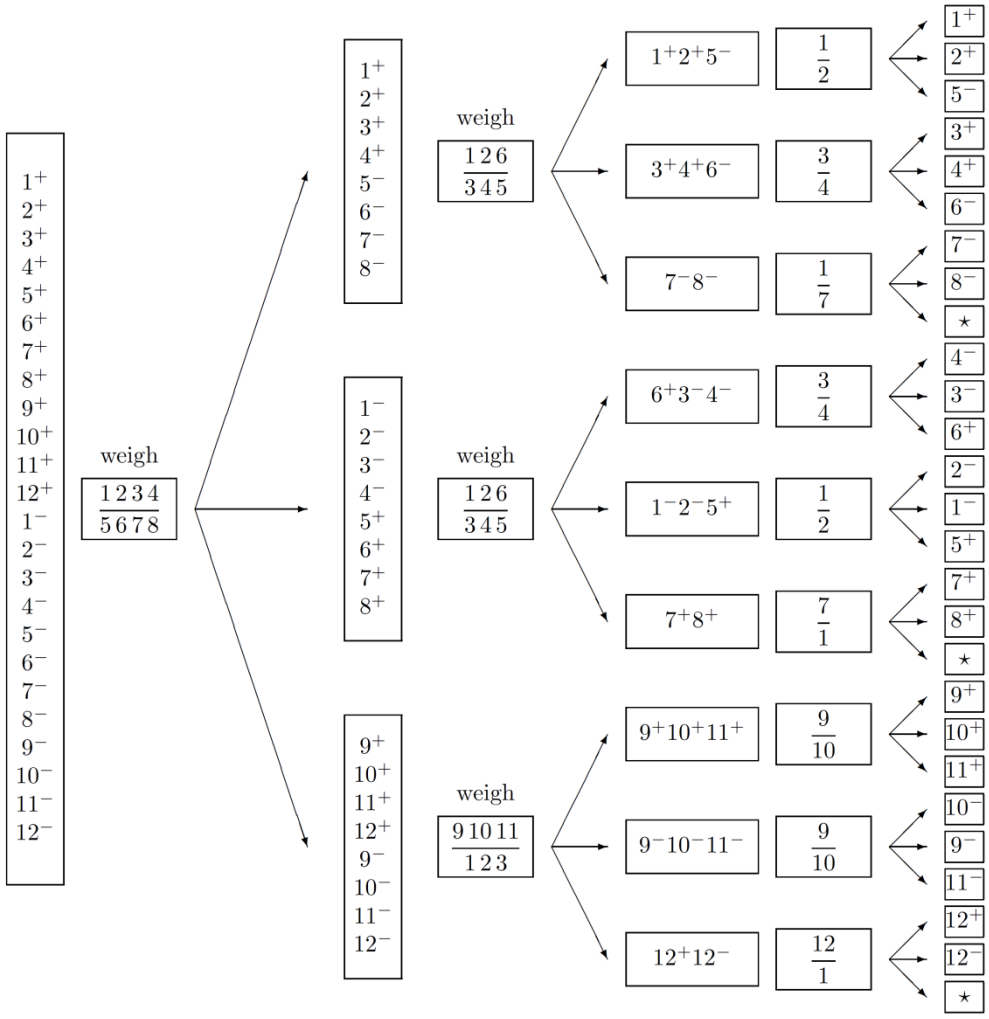


Example: weighing 12 balls to decide defective one

- We have 12 balls and 11 of these have the same weight apart from one, which may be **lighter or heavier**
- We can weigh these balls and the outcome can be :they (putting certain equal number of these balls) weigh the same, one side is lighter, one side is heavier.
- **Task**: How do we weigh (efficiently) these balls to decide:
  - Which of the balls is different
  - Is it lighter or heavier than the other balls
  - This needs to be done with **fewest possible number of weighing** !

Example: weighing 12 balls

Solution:



Taken from: **MacKay**.

Language Wenglish

- Wenglish is ‚similar' to English.
- Wenglish contains $2^{15} = 32768$ words, each of length 5.
- Words are created randomly using probability distribution on the right.
- Number of words in Wenglish is much smaller than the possible number of words with 5 letters ($26^5 \sim 12{,}000{,}000$).
- Notice: In Wenglish, we have more words using highly probable letters.

McKay page 72 !

Dictionary Wenglish

1	aaail
2	aaaiu
3	aaald
⋮	⋮
129	abati
⋮	⋮
2047	azpan
2048	aztdn
⋮	⋮
⋮	⋮
16 384	odrcr
⋮	⋮
⋮	⋮
32 737	zatnt
⋮	⋮
32 768	zxast

Probability distribution of English letters

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

Wenglish

- Assuming the probability of any word is same $p = 1/32768$ we have that Shannon's information equals to $\log_2 32768 = 15$ bits per word. Hence, 3 bits per letter.
- What happens if we analyze information letter by letter ?
- If word starts with 'a' then we get $\log_2 \frac{1}{0.625} \approx 4$ bits. If word starts with 'z' then we get $\log_2 \frac{1}{0.001} \approx 10$ bits of information !
- This means that we get much less information after 'z' than after 'a' !
- Similarity to English: Almost no information after 'xyl...' and much information after 'pro ...' !

Why logarithms in the definition of entropy I ?

- What about the two questions :

  1) " Is the temperature in Thailand over 30 degrees ? "
  This question has only two answers, YES or NO.

  2) " The president of Taiwan has spoken with a certain person from Hsinchu today. With whom ? "

  Here, the question has about 400,000 possible answers (since Hsinchu has about 400,000 inhabitants).

Conclusion

The number of possible answers r should be linked to information.

Why logarithms in the definition of entropy II ?

- Another example - you observe a gambler throwing a fair die:
  - 6 possible outcomes $\{1, 2, 3, 4, 5, 6\}$.
  - You note the outcome and then tell it to a friend giving a piece of information.
- Observe throwing a dice 3x and tell the outcomes to your friend. **The amount of information** that you give to your friend this time is **three times as much as the first time**.

Conclusion

"Information" should be additive in some sense.

Why logarithms in the definition of entropy III ?

- The problem is that the first case have $r = 6$ outcomes and the second $r = 6^3 = 216$ outcomes !
- It would imply that we would get $36 = 216/6$ times more information by comparing r ?
- Obvious solution is to relate amount of information through logarithms as,

$$\log_b 6^3 = 3 \log 6,$$

hence 3 times more information is gained.

Conclusion

Logarithm is a sensitive measure of information !

Decomposing entropy I

- Suppose X a random variable and $\mathcal{X} = \{0, 1, 2\}$.
- Decide by flipping a fair coin if $x = 0$, that is, $P(x = 0) = 1/2$
- Flip a coin again (if $x \neq 0$) to decide whether $x = 1$ or $x = 2$.

- The probability distribution and $H(X)$ are:

$$P(x = 0) = 1/2; \quad P(x = 1) = 1/4; \quad P(x = 2) = 1/4,$$

and

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = 1.5.$$

Decomposing entropy II

- Revealing **if** $x = 0$ **or not** corresponds to a binary variable with entropy $H(1/2, 1/2) = 1$ bit.
- **If** $x \neq 0$ then we need to learn the value of the second flip, again binary random variable $H(1/2, 1/2) = 1$.
- This leads to the total entropy of

$$H(X) = H(1/2, 1/2) + \frac{1}{2}H(1/2, 1/2) = 1.5,$$

  where $\frac{1}{2}$ since we flip second time half of the times.
- For probability distribution $\mathbf{p} = (p_1, p_2, \ldots, p_I)$

$$H(\mathbf{p}) = H(p_1, 1 - p_1) + (1 - p_1)H(\frac{p_2}{1 - p_1} \ldots, \frac{p_I}{1 - p_1}).$$

Reducing entropy

- Consider the following sequence:

  1 2 1 2 4 4 1 2 4 4 4 4 4 4 1 2 4 4 4 4 4 4

- Obtaining the probability from the sequence
  - 1, 2 four times (4/22), (4/22)
  - 4 fourteen times (14/22)

- The entropy H = 0.447 + 0.447 + 0.415 = 1.309 bits

- Since there are 22 symbols, we theoretically would need 22 * 1.309 = 28.798 (29) bits to transmit the information
- However, check the symbols 12, 44
- 12 appears 4/11 and 44 appears 7/11
- H = 0.530 + 0.415 = 0.945 bits
- 11 * 0.945 = 10.395 (11) bits (138 % less!)
- We might possibly be able to find patterns with less entropy

Entropy of joint (vezana), conditional (pogojni)

- **Vezana entropija** dveh (ali več) naključnih spremenljivk:

$$H(X,Y) = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{X,Y}(x,y)\,\log_2 \frac{1}{p_{X,Y}(x,y)}$$

- **Entropija pogojne porazdelitve**

$$H(X \mid Y=y) = \sum_{x\in\mathcal{X}} p_{X|Y}(x \mid y)\,\log_2 \frac{1}{p_{X|Y}(x \mid y)}$$

- **Pogojna entropija**

$$\begin{aligned} H(X \mid Y) &= \sum_{y\in\mathcal{Y}} p(y)\, H(X \mid Y=y) \\ &= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{X,Y}(x,y)\,\log_2 \frac{1}{p_{X|Y}(x \mid y)} \end{aligned}$$

Entropy of two (several) random variables

- **Joint entropy** of two random variables measures UNCERTAINTY about the pair (X, Y)
- **Conditional entropy** $H(X|Y = y)$ measures uncertainty about X given the value of Y which equals to y, thus for $Y = y$.
- **Conditional entropy** $H(X|Y)$ measures uncertainty about X given Y

Conditional entropy - example

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

Conditional entropy

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

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

$$= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0$$

$$= \frac{11}{8} \text{ bits.}$$

joint entropy

$$H(X, Y) = \frac{27}{8}$$

When do we get max. entropy ?

Condintional entropy

$$H(Y|X) = \frac{13}{8}$$

$$H(Y|X) \neq H(X|Y)$$

Chain rule for entropy

- For joint probability we had:

$$p_{X,Y}(x,y) = p_Y(y) \times p_{X|Y}(x \mid y)$$

- For entropy:

Verižno pravilo pri entropiji

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

Proof

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X).
\end{aligned}
$$

Equivalently (shorter proof): we can write

$$\log p(X,Y) = \log p(X) + \log p(Y|X)$$

and apply E to both sides. □

Chain rule for entropy

Verižno pravilo pri entropiji

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

- This rule can be easily generalized for more than 2 variables:

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid H_1, \ldots, H_{i-1})$$

In case that X and Y are **independent**, we have

$$H(X \mid Y) = H(X) \Leftrightarrow H(X, Y) = H(X) + H(Y)$$

Mutual (medsebojna) information

Medsebojna informacija

Medsebojna informacija med dvema diskretnima naključnima spr. X in Y je definirana kot

$$I(X;Y) = H(X) - H(X|Y).$$

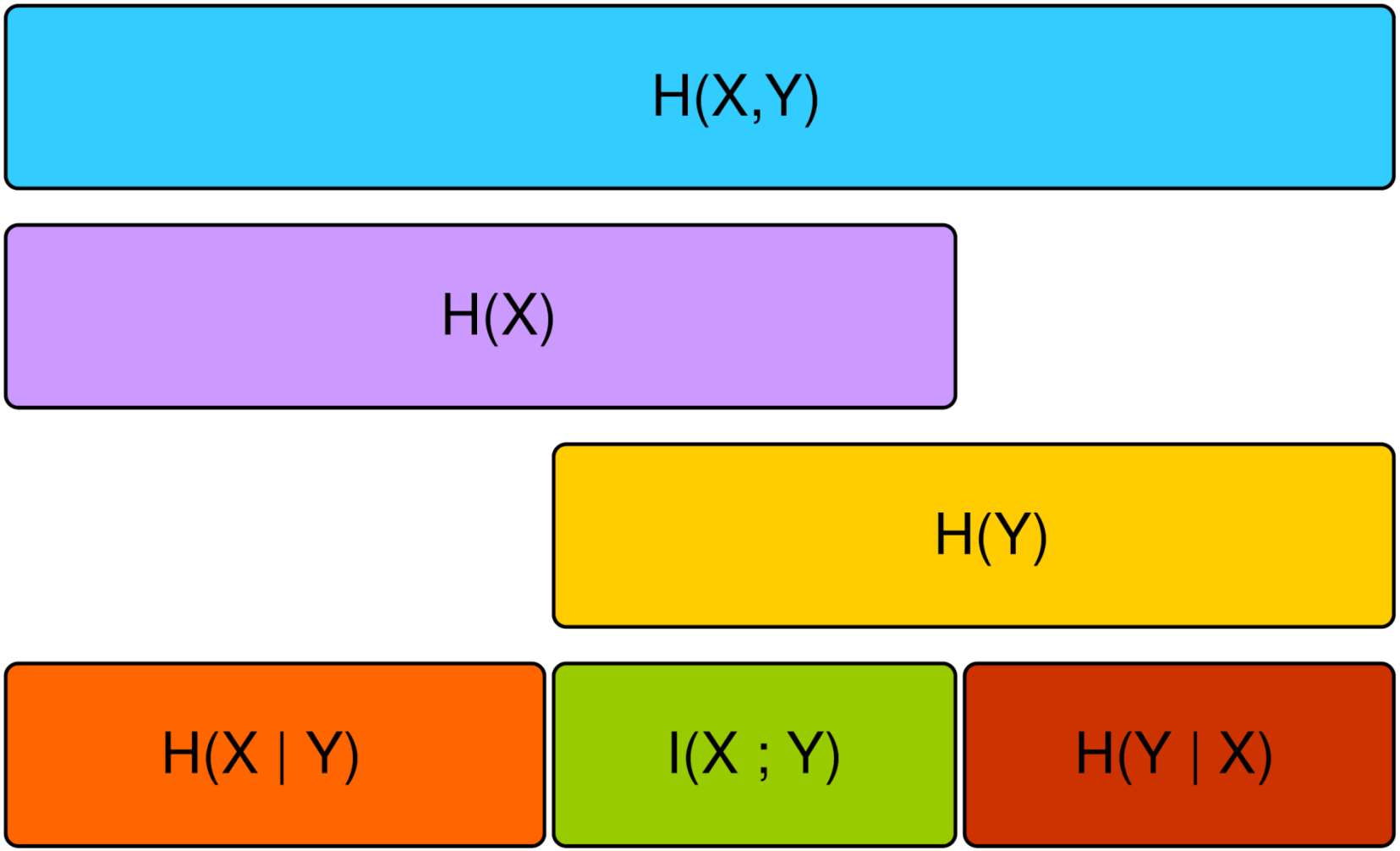
Z njo merimo povprečno nedoločenost spr. X ob znani spr. Y.

- Mutual information is symmetric:

$$\begin{aligned} I(X\,;\,Y) &= H(X) - H(X \mid Y) = (H(X) - H(X,Y)) + H(Y) \\ &= H(Y) - H(Y \mid X) = I(Y\,;\,X)\ . \end{aligned}$$

- Meaning: In average X tells about Y, as much as Y tells about X!

Connection between entropy and mutual information



Relative entropy

Kullback-Lebler distance - relative entropy

Relative entropy is so-called Kullback-Lebler distance between two (discrete) distributions (of RV X) defined as:

$$D(p_X||q_X) = \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{p_X(x)}{q_X(x)}.$$

It measures suitability of using q_X for p_X ! In general, $D(p_X||q_X) \neq D(q_X||p_X)$.

Information inequality

For arbitrary (discrete) distributions p_X and q_X (w.r.t. RV X) we have:

$$D(p_X||q_X) \geq 0.$$

The equality happens IFF $p_X(x) = q_X(x)$ for every $x \in \mathcal{X}$.

Proof: Jensens' inequality.

Proof of information inequality

We use Jenesen's inequality which claims that for a convex function f (in our case $-\log_2$) we have $E[f(X)] \geq f(E[X])$!

$$\begin{aligned} D(p_X||q_X) &= -E[\log_2 \frac{q_X(x)}{p_X(x)}] \geq -\log_2 E[\frac{q_X(x)}{p_X(x)}] \\ &= -\log_2 \left(\sum_x p_X(x)\frac{q_X(x)}{p_X(x)}\right) = -\log_2 1 = 0. \end{aligned}$$

Example – relative entropy

$\mathcal{X} = \{0, 1\}$, $p(0) = 1 - r$, $p(1) = r$, $q(0) = 1 - s$, $q(1) = s$.

$$D(p \parallel q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

$$D(q \parallel p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}$$

If $r = s$, then $D(p \parallel q) = D(q \parallel p)$, but for $r = \frac{1}{2}$, $s = \frac{1}{4}$

$$D(p \parallel q) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} = 0.207\,52 \text{ bit}$$

$$D(q \parallel p) = \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} = 0.188\,72 \text{ bit}$$

An application of Kullback – Leibler distance

- From information inequality we have:

$$I(X\,;\,Y) \geq 0$$

- *Proof:*

$$\begin{aligned} I(X\,;\,Y) &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{\substack{x\in\mathcal{X} \\ y\in\mathcal{Y}}} p_{X,Y}(x,y)\,\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)} \\ &= D(p_{X,Y} \parallel p_X p_Y) \geq 0\ . \end{aligned}$$

Moreover: $D(p_{X,Y}||p_X p_Y) = 0$ IFF $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. In other words, X and Y are independent. This also implies that $I(X;Y) = 0$.

Relative entropy – standard usage

- The relative entropy $D(p \parallel q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$.
- If, instead, we used the code for a distribution q, we would need $H(p) + D(p \parallel q)$ bits on the average to describe the random variable.

Properties of entropy

- $H(X) \geq 0$
- $H(X) \leq \log_2 |\mathcal{X}|$
  - Uniform distribution w.r.t X gives maximal entropy among all possible distributions of X. Equality arises IFF X is uniformly distributed.
- $H(X \mid Y) \leq H(X)$ equality only in the case when X and Y are independent
  - More information does not hurt: If we nkow (observe) Y, we reduce uncertainty about X.
  - It may happen that for certain $Y = y$, $H(X|Y=y) > H(X)$, but in average the above inequality is valid.

Dokaz II

We claim that if $|\mathcal{X}| = \mathcal{L}$ then $H(X) \leq \log_2 L$ with equality IFF $p_X(x_i) = 1/L$, for all $x_i \in \mathcal{X}$.

$$
\begin{aligned}
H(X) - \log_2 L &= -\sum_{i=1}^{L} p_X(x_i) \log_2 p_X(x_i) - \log_2 L \\
&= \sum_{i=1}^{L} p_X(x_i) \left(\log_2 \frac{1}{p_X(x_i)} - \log_2 L \right) \\
&= \sum_{i=1}^{L} p_X(x_i) \log_2 \frac{1}{L p_X(x_i)} \leq \\
&\leq \sum_{i=1}^{L} p_X(x_i) \left(\frac{1}{L p_X(x_i)} - 1 \right) \log_2 e = \\
&= \left(\sum_{i=1}^{L} \frac{1}{L} - \sum_{i=1}^{L} p_X(x_i) \right) \log_2 e = \\
&= (1 - 1) \log_2 e = 0.
\end{aligned}
$$

Chain rule (verižno pravilo) for entropy

Verižno pravilo za entropijo

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

Zgornja meja vezane entropije

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Enakost velja natanko tedaj, ko so X_i med seboj neodvisne naključne spremenljivke.

Equality above IFF X_i are independent.

Conditional mutual information and chain rule

- **Conditional mutual information is defined as**

$$I(X\,;\,Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

Verižno pravilo za medsebojno informacijo

$$I(Y; X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} I(Y; X_i | X_1, \ldots, X_{i-1})$$

- If X_i are mutually independent, we have

$$I(Y\,;\,X_1, \ldots, X_n) = \sum_{i=1}^{n} I(Y\,;\,X_i)$$

Markov chains

Markov chain

(Discrete) random variables X, Y and Z create a **Markov chain** $X \to Y \to Z$ if Z conditionally depends ONLY on Y and not on X. For Markov chain we have

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

Example of Markov chain

For instance, let Y measures X with added noise and $Z = f(Y)$ is output of some (deterministic or random) process, then $X \to Y \to Z$.

Properties of the Markov chain

Če $X \to Y \to Z$, potem

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

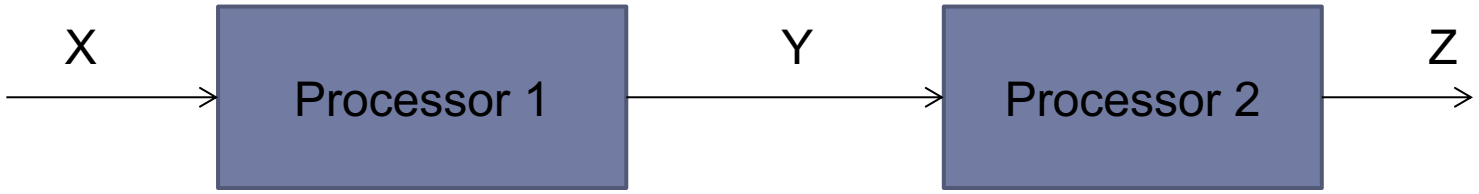
Zato sledi:

$$I(X; Z|Y) = H(Z|Y) - H(Z|Y, X) = 0.$$

To pomeni, da ko poznamo Y, z Z-jem ne dobimo dodatne informacije o X-u in obratno.

Last sentence: If we know (observe) Y, then Z does not help in getting additional information about X, or vice versa.

Data processing lemma with Markov chain



Can we get more information by processing the information through processors (assuming that Z depends on X only through Y) ?

Mutual information for data processing

Data processing inequality

If X, Y and Z build a **Markov chain** $X \to Y \to Z$ then

$$I(X;Z) \leq I(X;Y).$$

Additional processing of data ($Z = f(Y)$) DOES NOT increase information about X !!

Proof:

$$\begin{aligned} I(X\ ;\ Y,Z) &= I(X\ ;\ Z) + I(X\ ;\ Y \mid Z) \\ &= I(X\ ;\ Y) + I(X\ ;\ Z \mid Y) \end{aligned}$$

Using: $I(X\ ;\ Z \mid Y) = 0$ And: $I(X\ ;\ Y \mid Z) \geq 0$

Data processing inequality with nonMarkov chain

If X, Y, Z do not form a Markov chain it is possible that $I(X;Y|Z) > I(X;Y)$. For example, if X and Y are independent fair binary RVs and $Z = X + Y$, then $I(X;Y) = 0$,

Firstly, $I(X, Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z)$.

We know that $H(X|Z) \leq H(X) = 1$ bit. But $H(X|Z) > 0$ as

$$H(X|Z) = P(Z=1)H(X|Z=1) + P(Z=0)H(X|Z=0) = \ldots = 1/2.$$

Fanno's lemma – introduction (OPTIONAL)

- Suppose we wish to estimate $X \sim p(x)$
- We observe Y related to X by the conditional distribution $p(y|x)$. From Y we calculate $g(Y) = \widehat{X}$; $\widehat{X}$ is an estimate of X over the alphabet $\widehat{\mathcal{X}}$.
- $X \to Y \to \widehat{X}$ forms a Markov chain
- Define the probability of error

$$P_e = P\left\{\widehat{X} \neq X\right\}.$$

Fanno‘s lemma – introduction (OPTIONAL)

- Let X and Y be random variables with values in $\mathcal{X} = \{x_1, \ldots, x_L\}$, so $|\mathcal{X}| = \mathcal{L}$.
- **PROBLEM:** We observe Y and want to know "more" about X, i.e., what is the value of $H(X|Y)$?

- **IDEA:** To determine X we may ask when $X = Y$. If $X = Y$ (for particular realization) then we are done !
- If $X \neq Y$ then there are $L - 1$ possible values for X.
  - Leads to (Fanno's lemma)

$$H(X|Y) \leq H(P_e) + P_e \log_2(L-1),$$

where $P_e = P(X \neq Y)$.

Fanno's lemma – proof (OPTIONAL)

- Introduce error indicator (binary random variable)

$$Z = \begin{cases} 0, & X = Y, \\ 1, & X \neq Y. \end{cases}$$

- First, $H(Z) = H(P_e)$!

- Then, $H(X, Z|Y) = H(X|Y) + H(Z|X, Y) = H(X|Y)$ since X and Y determines Z uniquely !
  (For identity $p_{x,z|y} = \frac{p_{x,y,z}}{p_y} = \frac{p_{z|x,y}p_{x,y}}{p_y}$...).

- Therefore,

$$\begin{aligned} H(X|Y) &= H(X, Z|Y) = \\ &= H(Z|Y) + H(X|YZ) \leq \\ &\leq H(Z) + H(X|YZ). \end{aligned}$$

Fano's lemma proof cont. (OPTIONAL)

- Two things should be noticed

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

since we know X after observing Y and

$$H(X|Y, Z = 1) \leq \log_2(L - 1),$$

as there are only $L - 1$ values of X when $X \neq Y$. Thus,

$$H(X|YZ) \leq P(Z = 1)\log_2(L - 1) = P_e \log_2(L - 1)$$

so finally

$$H(X|Y) \leq H(P_e) + P_e \log_2(L - 1).$$

Weak law of large numbers and entropy

- If $X_1, X_2, \ldots$ is a sequence of I. I. D. random variables, having identical distribution over alphabet $\mathcal{X}$, then:

$$\log_2 \frac{1}{p_X(X_1)}, \log_2 \frac{1}{p_X(X_1)}, \ldots$$

  also a sequence of I. I. D. random variables.

- Mathematical expectation (average) of any of these random variables $\log_2 \frac{1}{p_X(X_i)}$ equals to the entropy:

$$E[\log_2 \frac{1}{p_X(X_i)}] = \sum_x p_X(x) \log_2 \frac{1}{p_X(x)} = H(x) \quad \text{for all } i \in \mathbb{N}.$$

Weak law of large numbers and entropy

- From our hypothesis that X_i are I.I.D. random variables we get:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i) \qquad \frac{1}{p(x_1, \ldots, x_n)} = \prod_{i=1}^{n} \frac{1}{p_X(x_i)}$$

$$\log_2 \frac{1}{p(x_1, \ldots, x_n)} = \log_2 \prod_{i=1}^{n} \frac{1}{p_X(x_i)}$$

$$\log_2 \frac{1}{p(x_1, \ldots, x_n)} = \sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)}$$

$$\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_X(x_i)}$$

Weak law of large numbers and entropy - statement

- Thus, for I.I.D. random variables X_i we have:

$$\frac{1}{n}\log_2\frac{1}{p(x_1,\ldots,x_n)} = \frac{1}{n}\sum_{i=1}^{n}\log_2\frac{1}{p_X(x_i)}$$

Po šibkem zakonu velikih števil zaporedje na desni strani konvergira v verjetnosti k povprečju, to je k entropiji:

$$\lim_{n\to\infty} Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}\log_2\frac{1}{p_X(X_i)} - H(X)\right| < \epsilon\right] = 1 \text{ za vsak } \epsilon > 0.$$

We are applying above weak law of large numbers mentioned earlier

AEP statement

- Asymptotic Equipartition Property

Lastnost AEP

Za zaporedje enako porazdeljenih in neodvisnih naključnih spremenljivk velja

$$\lim_{n\to\infty} Pr\left[\left|\frac{1}{n}\log_2\frac{1}{p_X(x_1,\ldots,x_n)} - H(X)\right| < \epsilon\right] = 1$$

za vsak $\epsilon > 0$.

For a sequence of I.I.D. random variables X_i the above value converges to $H(X)$ for sufficiently large n

AEP interpretation

- AEP states, that for every $\epsilon > 0$ there is sufficiently large n, so that

$$\Pr\left[\left|\frac{1}{n}\log_2\frac{1}{p(x_1,\ldots,x_n)} - H(X)\right| < \epsilon\right] \approx 1$$

$$H(X) - \epsilon < \frac{1}{n}\log_2\frac{1}{p(x_1,\ldots,x_n)} < H(X) + \epsilon$$

$$2^{n(H(X)-\epsilon)} < \frac{1}{p(x_1,\ldots,x_n)} < 2^{n(H(X)+\epsilon)}$$

$$2^{-n(H(X)+\epsilon)} < p(x_1,\ldots,x_n) < 2^{-n(H(X)-\epsilon)}$$

$$\Leftrightarrow \quad \Pr\left[p(x_1,\ldots,x_n) = 2^{-n(H(X)\pm\epsilon)}\right] \approx 1$$

Typical sequences having probability almost 1

Subset of typical sequences

Subset of typical sequences, denoted $A_\epsilon^{(n)}$ contains all sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ satisfying:

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

The AEP property then gives:

$$\lim_{n\to\infty} Pr[X^n \in A_\epsilon^{(n)}] = 1.$$

In other words, for all (small) $\epsilon > 0$ there is (large) n so that

$$Pr[X^n \in A_\epsilon^{(n)}] > 1 - \epsilon.$$

Typical sequences - properties

- How many sequences are in this set $A_\epsilon^{(n)}$?
- We use the property that the probability of each such sequence is at least $2^{-n(H(X)+\epsilon)}$

- Since the total probability of all sequences in $A_\epsilon^{(n)}$ is almost 1, we cannot have many of these. How many ?

$$
\begin{aligned}
1 &\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) \\
&\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)} \left|A_\epsilon^{(n)}\right| \\
\Leftrightarrow \quad & \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)} \quad .
\end{aligned}
$$

Typical sequences properties II

- Can we have the case that cardinality of $A_\epsilon^{(n)}$ is very small ?
- Now we use the property that any sequence in $A_\epsilon^{(n)}$ has probality less than $2^{-n(H(X)-\epsilon)}$.

- AEP states that the total probability (for large n) is greater than $1-\epsilon$.

$$
\begin{aligned}
1-\epsilon &< \Pr\left[X^n \in A_\epsilon^{(n)}\right] \\
&\leq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} \left|A_\epsilon^{(n)}\right| \\
&\Leftrightarrow \left|A_\epsilon^{(n)}\right| > (1-\epsilon)2^{n(H(X)-\epsilon)} \quad .
\end{aligned}
$$

Typical sequences - summary

- Thus, AEP for arbitrary small ε and sufficiently large n claims:
  - Typical sequences have large (total) probability (almost 1).
  - The cardinality of $A_\epsilon^{(n)}$ is approximately $2^{nH(X)}$

- What does that means ?

- Cardinality of all sequences $(x_1, \ldots, x_n) \in \mathcal{X}$ of length n is $|\mathcal{X}|^n$. Maximal entropy is $\log_2 |\mathcal{X}|$. Now, if the entropy of our sequences is $H(X) = \log_2 |\mathcal{X}|$ then

$$|A_\epsilon^{(n)}| \approx 2^{nH(X)} = 2^{n \log_2 |\mathcal{X}|} = |\mathcal{X}|^n,$$

  which means that the cardinality of typical sequences can be as large as $|\mathcal{X}|^n$.

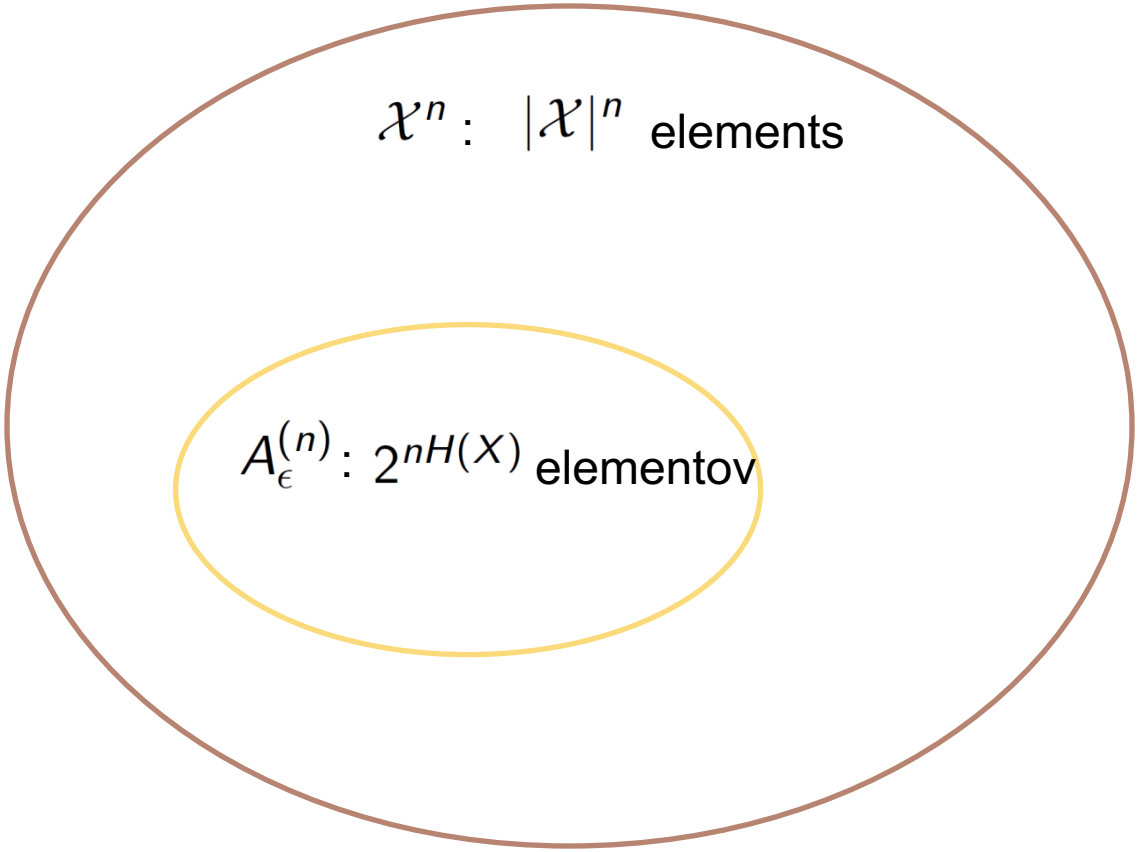
Typical sequences when $H(X) < \log_2 |\mathcal{X}|$

- On the other hand:

  - Cardinality of all sequences $(x_1, \ldots, x_n) \in \mathcal{X}$ of length n is $|\mathcal{X}|^n$. Maximal entropy is $\log_2 |\mathcal{X}|$. In general, $H(X) < \log_2 |\mathcal{X}|$ which actually implies that **the cardinality of typical sequences is exponentially smaller than the cardinality of all sequences**:

$$\frac{2^{nH(X)}}{2^{n \log_2 |\mathcal{X}|}} = 2^{-n\delta} \xrightarrow[n \to \infty]{} 0, \quad \text{if } \delta = \log_2 |\mathcal{X}| - H(X) > 0.$$

Typical sequences



A subset of typical sequences is almost negligible part of all sequences but their total probability is almost one !!

AEP example

Dragging 5 times one ball from a hat - there are 2 white balls and one black

Non-typical →

	probabilty	ϵ − typical($\epsilon = 0.138$)
● ● ● ● ●	1/3 1/3 1/3 1/3 1/3 → 0.0041	
● ● ● ● ○	1/3 1/3 1/3 1/3 2/3 → 0.0082	
● ● ● ○ ●	1/3 1/3 1/3 2/3 1/3 → 0.0082	
● ● ● ○ ○	1/3 1/3 1/3 2/3 2/3 → 0.0165	
● ● ○ ● ●	1/3 1/3 2/3 1/3 1/3 → 0.0082	
● ● ○ ● ○	1/3 1/3 2/3 1/3 2/3 → 0.0165	
● ● ○ ○ ●	1/3 1/3 2/3 2/3 1/3 → 0.0165	
● ● ○ ○ ○	1/3 1/3 2/3 2/3 2/3 → 0.0329	⋆
● ○ ● ● ●	1/3 2/3 1/3 1/3 1/3 → 0.0082	
● ○ ● ● ○	1/3 2/3 1/3 1/3 2/3 → 0.0165	
● ○ ● ○ ●	1/3 2/3 1/3 2/3 1/3 → 0.0165	
● ○ ● ○ ○	1/3 2/3 1/3 2/3 2/3 → 0.0329	⋆
● ○ ○ ● ●	1/3 2/3 2/3 1/3 1/3 → 0.0165	
● ○ ○ ● ○	1/3 2/3 2/3 1/3 2/3 → 0.0329	⋆
● ○ ○ ○ ●	1/3 2/3 2/3 2/3 1/3 → 0.0329	⋆
● ○ ○ ○ ○	1/3 2/3 2/3 2/3 2/3 → 0.0658	⋆
○ ● ● ● ●	2/3 1/3 1/3 1/3 1/3 → 0.0082	
○ ● ● ● ○	2/3 1/3 1/3 1/3 2/3 → 0.0165	
○ ● ● ○ ●	2/3 1/3 1/3 2/3 1/3 → 0.0165	
○ ● ● ○ ○	2/3 1/3 1/3 2/3 2/3 → 0.0329	⋆
○ ● ○ ● ●	2/3 1/3 2/3 1/3 1/3 → 0.0165	
○ ● ○ ● ○	2/3 1/3 2/3 1/3 2/3 → 0.0329	⋆
○ ● ○ ○ ●	2/3 1/3 2/3 2/3 1/3 → 0.0329	⋆
○ ● ○ ○ ○	2/3 1/3 2/3 2/3 2/3 → 0.0658	⋆
○ ○ ● ● ●	2/3 2/3 1/3 1/3 1/3 → 0.0165	
○ ○ ● ● ○	2/3 2/3 1/3 1/3 2/3 → 0.0329	⋆
○ ○ ● ○ ●	2/3 2/3 1/3 2/3 1/3 → 0.0329	⋆
○ ○ ● ○ ○	2/3 2/3 1/3 2/3 2/3 → 0.0658	⋆
○ ○ ○ ● ●	2/3 2/3 2/3 1/3 1/3 → 0.0329	⋆
○ ○ ○ ● ○	2/3 2/3 2/3 1/3 2/3 → 0.0658	⋆
○ ○ ○ ○ ●	2/3 2/3 2/3 2/3 1/3 → 0.0658	⋆
○ ○ ○ ○ ○	2/3 2/3 2/3 2/3 2/3 → 0.1317	
	0.9998	$\sum p^\star = 0.6580$

Usual misinterpretation

- ϵ-typical sequences **are not always most likely**
- It can happen that non-typical sequences have larger probability than typical (see all white balls)

- Consider casting a biased coin with $P(head) = p$ and $P(tail) = q = 1 - p$.
- Casting n times you expect to get head np times. If $q > p$ then getting “all tails" is more likely but non-typical !!

AEP example cont.

- Easy to compute

$$H(X) = H(1/3) = 1/3 \log_2 3 + 2/3 \log_2 3/2 = 0.918.$$

  Notice that $\epsilon = 0.138$, i.e. 15% of $H(1/3)$.
- Using $2^{-n(H(X)+\epsilon)} \leq p_X(x) \leq 2^{-n(H(X)-\epsilon)}$ one gets

$$0.027 \leq p_X(x) \leq 0.068$$

- We have 15 $\epsilon-$typical sequences and our estimate says

$$12.4 \leq |A_\epsilon(X)| \leq 37.6$$

AEP example cont.

- We lower $\epsilon = 0.046$, i.e., 5% of $H(1/3)$, and consider large n.

n	$(1-\epsilon)2^{n(H(X)-\epsilon)}$	$\lvert A_\epsilon(X)\rvert$	$2^{n(H(X)+\epsilon)}$	$P(A_\epsilon(X))$
100	$2^{86.6}$	$2^{92.6}$	$2^{96.4}$	0.660
500	$2^{436.1}$	$2^{474.9}$	$2^{482.1}$	0.971
1000	$2^{872.4}$	$2^{953.4}$	$2^{964.2}$	0.998

Fraction of $\epsilon-$typical sequences for $n = 1000$ is

$$\frac{2^{953.4}}{2^{1000}} = 2^{-46}!!!$$

and they contribute with 0.998 probability !

- Idea is to only use $\epsilon-$typical sequences (ignore the rest) - Shanon source coding.

Other examples: tossing a coin

- Tossing a coin introduces RV X with alphabet $\mathcal{X} = \{0, 1\}$, e.g. "head =0" and "tail =1". In general, assign $p_X(1) = p$ and $p_X(0) = 1 - p$. If we toss a coin n times we have I.I.D. RVs and consequently:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

- In this example, typical sequences $A_\epsilon^{(n)}$ consists of those sequences whose $\sum_i x_i$ is approximately np. (Think about tossing a coin 100 times what do you expect ?)
- For these sequences we have:

$$\begin{aligned} \log_2 \frac{1}{p_X(x_1, \ldots, x_n)} &\approx \log_2 \frac{1}{p^{np}(1-p)^{n(1-p)}} \\ &= n(p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}) = nH(X). \end{aligned}$$

Another example: casting a dice

- Casting a dice introduces RV X with alphabet $\mathcal{X} = \{1, 2, \ldots, 6\}$. The probability $p_X(j) = p_j$ for all $j \in \mathcal{X}$. If we cast a dice n times we have I.I.D. RVs and consequently:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{j=1}^{6} p^{k_j},$$

  where k_j is the number of cases we get j as the outcome.
- In this example, typical sequences $A_\epsilon^{(n)}$ consists of those sequences such that $k_j \approx np_j$, for all $j \in \mathcal{X}$.
- For these sequences we have:

$$\begin{aligned} \log_2 \frac{1}{p_X(x_1, \ldots, x_n)} &\approx \log_2 \frac{1}{\prod_{j=1}^{6} p_j^{np_j}} \\ &= n\left(\sum_{j=1}^{6} p_j \log_2 \frac{1}{p_j}\right) = nH(X). \end{aligned}$$

Coding using AEP

- We specify coding using sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ as binary sequences $\{0,1\}^*$ of arbitrary length.
- Let $x^n \in \mathcal{X}^n$ denote a sequence $(x_1, \ldots, x_n)$ and $\ell(x^n)$ be the length of a codeword stemming from x^n.
- **Our goal** is to build a code whose average length of codewords is close to the entropy of the source:

$$E[\frac{1}{n}\ell(x^n)] = H(X) + \epsilon,$$

  for sufficiently large n.
- This is **best** we can achieve with **lossless source coding**.

Coding with AEP

We divide sequences $x^n \in \mathcal{X}^n$ into two groups:

- Typical sequences $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$.
- NON-typical sequences $x^n \in \mathcal{X}^n \setminus A_\epsilon^{(n)}$.

There are at most $|\mathcal{X}^n|$ non-typical sequences. It is sufficient with $\log_2 |\mathcal{X}^n| + 1$ bits to encode these.

Extra bit is used to distinguish where the codeword comes from (typical or not) !

Average length when coding with AEP

$$
\begin{aligned}
E[\ell(X^n)] &= E\left[\ell(X^n) \,\middle|\, X^n \in A_\epsilon^{(n)}\right] \Pr\left[X^n \in A_\epsilon^{(n)}\right] \\
&\quad + E\left[\ell(X^n) \,\middle|\, X^n \notin A_\epsilon^{(n)}\right] \Pr\left[X^n \notin A_\epsilon^{(n)}\right] \\
&= (n(H(X)+\epsilon)+2) \Pr\left[X^n \in A_\epsilon^{(n)}\right] \\
&\quad + (n \log_2 |\mathcal{X}| + 2) \Pr\left[X^n \notin A_\epsilon^{(n)}\right] \\
&\leq n(H(X)+\epsilon) + n \log |\mathcal{X}| \epsilon + 2 \quad \text{(AEP)} \\
&= n(H(X)+\epsilon') \ ,
\end{aligned}
$$

Remark:

$\epsilon' = \epsilon + \epsilon \log_2 |\mathcal{X}| + \frac{2}{n}$ Is sufficiently small if

$\epsilon > 0$ Is arbitrary small and n suffic. large

Example – source coding of a Swedish text

- Consider length $n = 15$ and $|\mathcal{X}| = 29$ - nmb. of letters in Swedish language.
- Requires

$$15 \log_2 29 = 15 \cdot 4.86 = 73\text{bits},$$

to represent a sequence of 15 letters.

- AEP is also valid for other models (not only IID) - e.g. for languages. Entropy of Swedish is c.a. 1/3 compared to the entropy of using letter independently !!
- Thus, instead we use

$$15 \cdot 1/3 \cdot 4.86 = 25\text{bits}.$$

Optimality of source coding using AEP

Average length of codewords using AEP

Average length of codewords using AEP satisfies:

$$E[\frac{1}{n}\ell(x^n)] \leq H(X) + \epsilon$$

for arbitrary (small) ϵ and sufficiently large n.

Optimality:

We also have that there are $2^{nH(X)}$ sequences with approx. same probability $2^{-nH(X)}$. These sequences can be encodes using $n(H(X) - \delta)$ bits which is only a fraction $2^{-n\delta}$ of total number of sequences. Thus, the average length of codewords is at least $H(x)$.

Optimality using AEP

Average length of codewords using AEP

Average length of codewords using AEP satisfies:

$$H(X) \leq E[\frac{1}{n}\ell(x^n)] \leq H(X) + \epsilon$$

for arbitrary (small) ϵ and sufficiently large n.

In other words:

We can encode source X, so that the average length of codewords is approximately H(X) for large *n*. With AEP coding we can decode Without errors (lossless coding).

Shannon‘s result on source coding

Shannon's result on source coding

For a random variable X with alphabet $\mathcal{X}$ there exists prefix-free coding $C : \mathcal{X} \to \{0,1\}^*$ with average length of codewords:

$$H(X) \leq E[\frac{1}{n}\ell(x^n)] \leq H(X) + 1$$

Shannon‘s result on coding

Shannon's result on coding

If the source has entropy H (bits/symbol) and channel has capacity C (bits/second) then we can encode source and transmit at most $C/H - \epsilon$ (symbol/second) for arbitrary small ϵ. **Cannot transmit faster without errors !**.

Example: Assume we have error-free channel transmitting $0 \to 0$ and $1 \to 1$. Its capacity $C = 1$ and consequently the rate is:

$$R = \lim_{n\to\infty} \frac{n}{\ell(x^n)} < \frac{1}{H(X)}$$

.

Shannon‘s result on source coding with AEP

- AEP states the following:

Shannon's result on source coding

Information about N outcomes of I.I.D. RVs with entropy $H(X)$ can be encoded with at least $NH(X)$ bits for large N and we can decode later (lossless compression). If we try to compress with less than $NH(X)$ bits we inevitably get errors at decoding.