Entropy of independent sources

October 16, 2023

1 Entropy of independent sources

1.1 Entropy of binary encoding: the binary entropy function

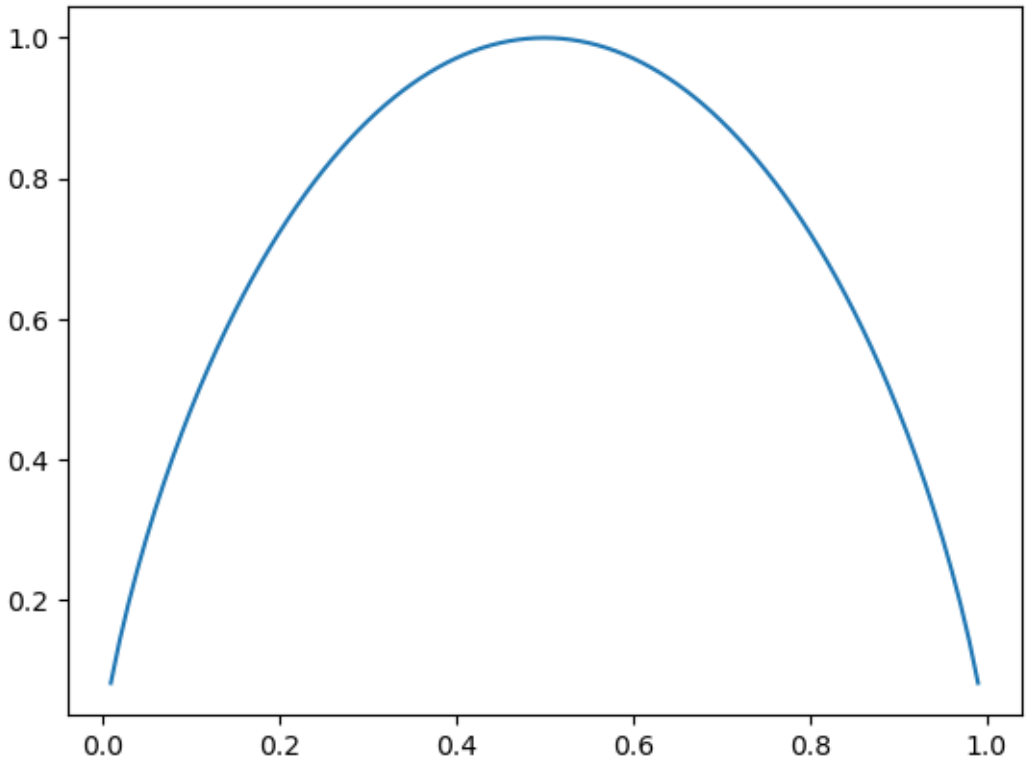
The binary entropy function $H_b(p)$ is defined as the entropy of a Bernoulli process. The Bernoulli process is modeled as a random variable X that can take only $n = 2$ values: 0 and 1 with the probability of success $P[X = 1] = p$ and probability of failure of $P[X = 0] = 1 - p$.

Proof analytically that the entropy equals

$$H(X) = H_b(p) = \frac{1}{\ln 2}\left(p \ln \frac{1-p}{p} - \ln(1-p)\right)$$

Plot the function and interprete the points for $p = 0$, $p = 0.25$, $p = 0.5$, and $p = 1$. Note that the function evaluation for $p = 0$ and $p = 1$ requires special attention!

[14]: `[<matplotlib.lines.Line2D at 0x1b8a5567fd0>]`



1.2 Entropy of acadenic example

Consider a discrete source X with an alphabet of length $n = 4$ for which the probabilities for the occurrences is given by $p(x_i) = [0.4, 0.3, 0.2]$ for $i = 1, \ldots, 3$. Compute the entropy of the source (expressed in bit/symbol)

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i))$$

and the maximal entropy

$$H_{max} = \log_2(n)$$

which is obtained for uniform distributed sources with $p(x_i) = 1/n$.

Since p_i represents a probability, we have to ensure that the sum of all p_i s is equal to one

$$\sum_{i=1}^{n} p(x_i) = 1.$$

```
p_i = [0.4 0.3 0.2 0.1]
H_X = 1.84644 bit/symbol
H_max = 2.00000 bit/symbol
```

1.2.1 Information rate and channel capacity

The entropy measures the amount of information generated by a source. The information rate measures the amount of information generated **per time unit**.

The information rate can be written for a binary coder as

$$R(X) = \frac{1}{< N >} H(X)$$

with $< N >$ the average length

$$< N >= \sum_{i=1}^{n} p(x_i) N(x_i).$$

1.3 Entropy using letter frequency in English

Consider that the different letters/symbols within English are independent and that each letter from 'a' to 'z' + 'space' represent a symbol with $p_i('a', \ldots, 'z') = [0.065, 0.012, 0.022, 0.035, 0.104, 0.020, 0.016, 0.049, 0.056, 0.001, 0.005, 0.033, 0.020, 0.056, 0.060, 0.014, 0.001, 0.050, 0.052, 0.073, 0.022, 0.008, 0.017, 0.001, 0.015, 0.001]$ and $p_i('space') = 0.192$.

Determine the self-information

$$S_i = -\log_2(p_i)$$

of each character and the entropy of an English text (assuming that all charaters are independent).

Symbol	p_i	S_i
a	0.06500	3.94342
b	0.01200	6.38082
c	0.02200	5.50635
d	0.03500	4.83650
e	0.10400	3.26534
f	0.02000	5.64386
g	0.01600	5.96578
h	0.04900	4.35107
i	0.05600	4.15843
j	0.00100	9.96578
k	0.00500	7.64386
l	0.03300	4.92139
m	0.02000	5.64386
n	0.05600	4.15843
o	0.06000	4.05889
p	0.01400	6.15843
q	0.00100	9.96578
r	0.05000	4.32193
s	0.05200	4.26534
t	0.07300	3.77596
u	0.02200	5.50635
v	0.00800	6.96578
w	0.01700	5.87832
x	0.00100	9.96578
y	0.01500	6.05889
z	0.00100	9.96578
space	0.19200	2.38082

Determine the entropy of the English source

$$H_X = \sum_{i=1}^{n} p_i S_i.$$

`Entropy: H_X = 4.07164 bit/symbol`

Determine the maximal entropy (for uniformly distributed symbols).

`Max. Entropy: H_max = 4.75489 bit/symbol`

Compute the the efficiency.

`Efficiency: e_X = 85.6%`

2 Entropy of morse code

Consider the above probabilities for the English characters and assume that the relative lenght of a morse code for characters from 'a' to 'z' is given as [8, 12, 14, 10, 4, 12, 12, 10, 6, 16, 12, 12, 10, 8, 14, 14, 16, 10, 8, 6, 10, 12, 12, 14, 16, 14] and equals 4 for a space. This is expressed in time

units (a single dot equals 2 time units). Remark that the entropy of the source remains identical to the previous question, namely the entropy of the letter frequency in English.

```
Entropy: H_X = 4.07164 bit/symbol
```

Determine the average duration of a symbol and the the information rate.

```
tau_avg = 8.07800 bit/unit
Symbol  p_i     S_i     tau_i
a       0.06500 3.94342 8
b       0.01200 6.38082 12
c       0.02200 5.50635 14
d       0.03500 4.83650 10
e       0.10400 3.26534 4
f       0.02000 5.64386 12
g       0.01600 5.96578 12
h       0.04900 4.35107 10
i       0.05600 4.15843 6
j       0.00100 9.96578 16
k       0.00500 7.64386 12
l       0.03300 4.92139 12
m       0.02000 5.64386 10
n       0.05600 4.15843 8
o       0.06000 4.05889 14
p       0.01400 6.15843 14
q       0.00100 9.96578 16
r       0.05000 4.32193 10
s       0.05200 4.26534 8
t       0.07300 3.77596 6
u       0.02200 5.50635 10
v       0.00800 6.96578 12
w       0.01700 5.87832 12
x       0.00100 9.96578 14
y       0.01500 6.05889 16
z       0.00100 9.96578 14
space   0.19200 2.38082 4

Information rate: R_X = 0.50404 bit/unit
```

Determine the maximal entropy H_{max} assuming uniformly distributed symbols.

```
Max. entropy: H_max = 4.75489 bit/symbol
```

3 Entropy of Calligraphy in English

What is the amount of information per symbol when using the "calligraphical" information of an English text (i.e. compare "Amount of information is normal font compared to capital letters" to "AMOUNT OF INFORMATION IS NORMAL FONT COMPARED TO CAPITAL LETTERS") Therefore consider 5 groups of symbols 1. "aceimnorsuvwxz": neither sticks nor tails, 2. "bdhklt": only sticks, 3. "gjpqy": only tails, 4. "f": both sticks and tails, and 5. " ": a space.

```
p_cal = [0.534 0.207 0.047 0.02  0.192]
```

Compute the entropy of Calligraphy in English

```
Entropy: H_X_cal = 1.73100 bit/symbol
```