Compression of radio astronomy data flows

Tim Natusch
IRASR - AUT University

Motivation

- Data transport and storage are costly, particularly at the scale of the SKA.
- Compression reduces data volumes that must be transported/stored, potential cost reductions follow.

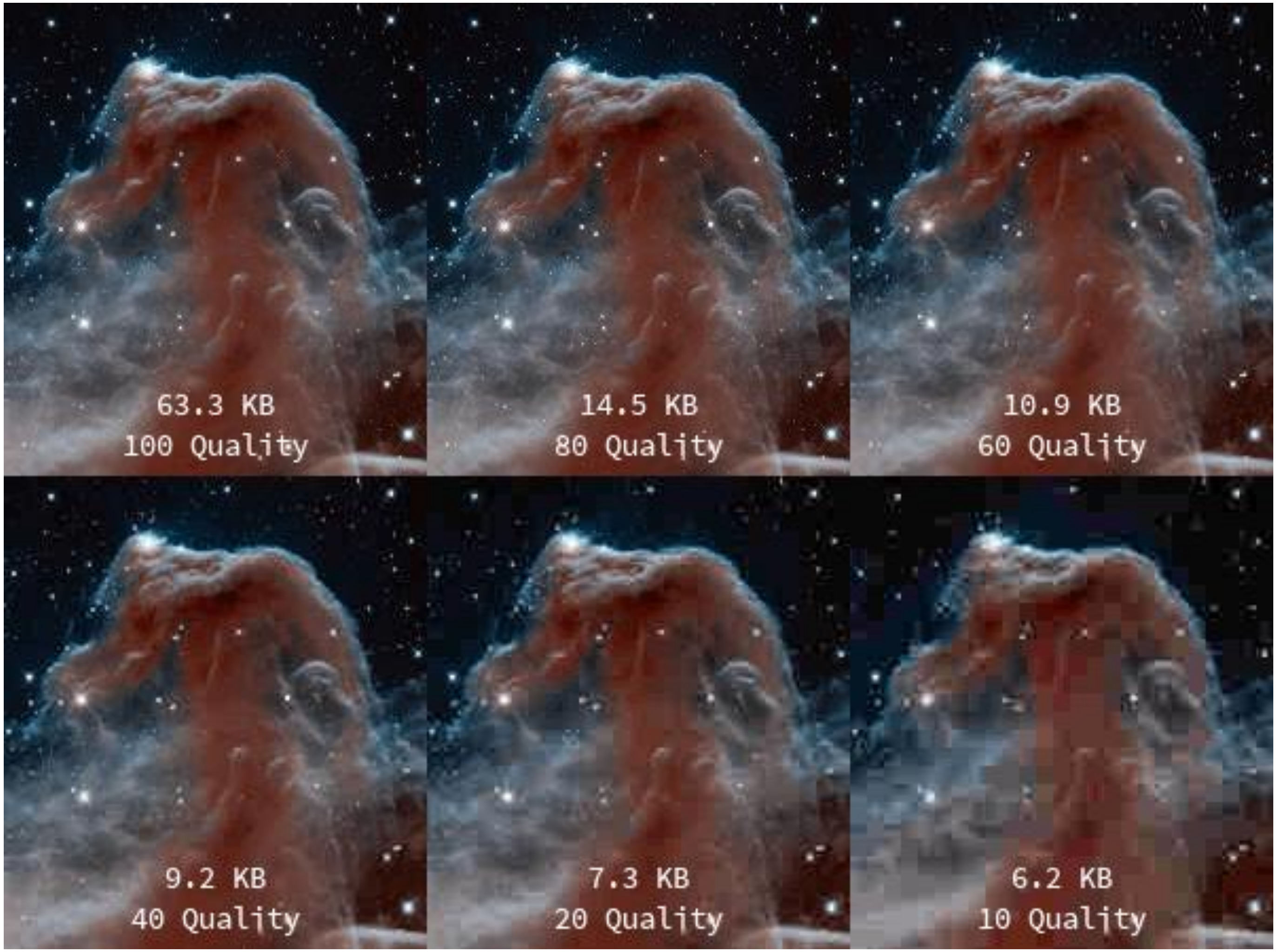
(It comes with its own costs/overheads of course!)

The general scheme

2 types of compression

- Lossless – can completely recover the original information from the compressed data
- Lossy – information lost during the compression process is irretrievable

Lossy JPEG compression



Lossless, simple example

Run Length Logic (RLL)

- Original data string = 55555555555555555555
- RLL string = 555#17
- Send 6 characters instead of 20!

Data source characteristics

Need to know;

- Source Alphabet = set of all data symbols/states the source can generate

  A={a, b, c, d, … ,z}

  A = {00,01,10,11}

- Symbol probability distribution

  P={p(a), p(b), … , p(z)}

  P={p(00), p(01), p(10), p(11)}

Some information theory

- $$I(a_i) = \log\left(\frac{1}{P(a_i)}\right) = -\log(P(a_i))$$
- If base of log = 2 then unit = the bit

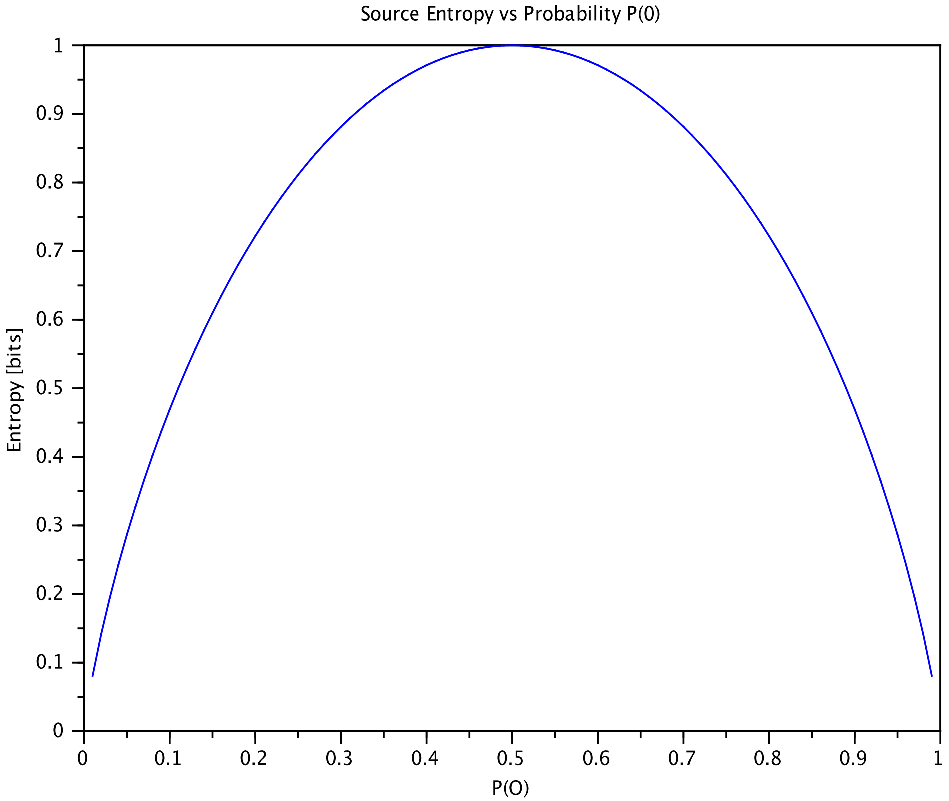
$$I(a_i) = -\log_2(P(a_i)) \text{ bits}$$

- Source Entropy = average no of bits per symbol

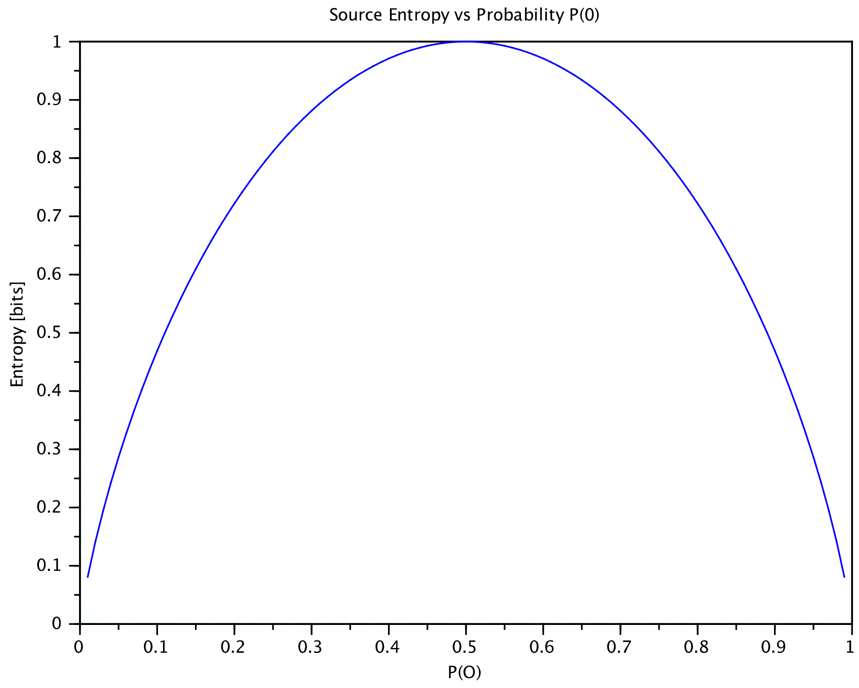
$$H(A,P) = \sum_{i=1}^{N} P(a_i) I(a_i) = -\sum_{i=1}^{N} P(a_i) \log_2(P(a_i))$$

Entropy in 1 bit case

- $A = \{0, 1\}$, $P = \{p(0), p(1)\}$, $p(0)+p(1)=1$



- Entropy is maximised when symbol emission is equiprobable!
- Entropy reduces as the distribution moves from the equiprobable state!

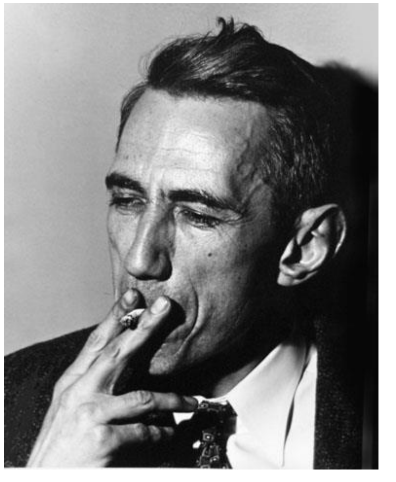


Shannon lossless source coding theorem

$$H(A,P) \leq L \leq H(A,P) + \frac{1}{N}$$

- L = minimum possible length of (losslessly) compressed data string
- For $N \to \infty$ then

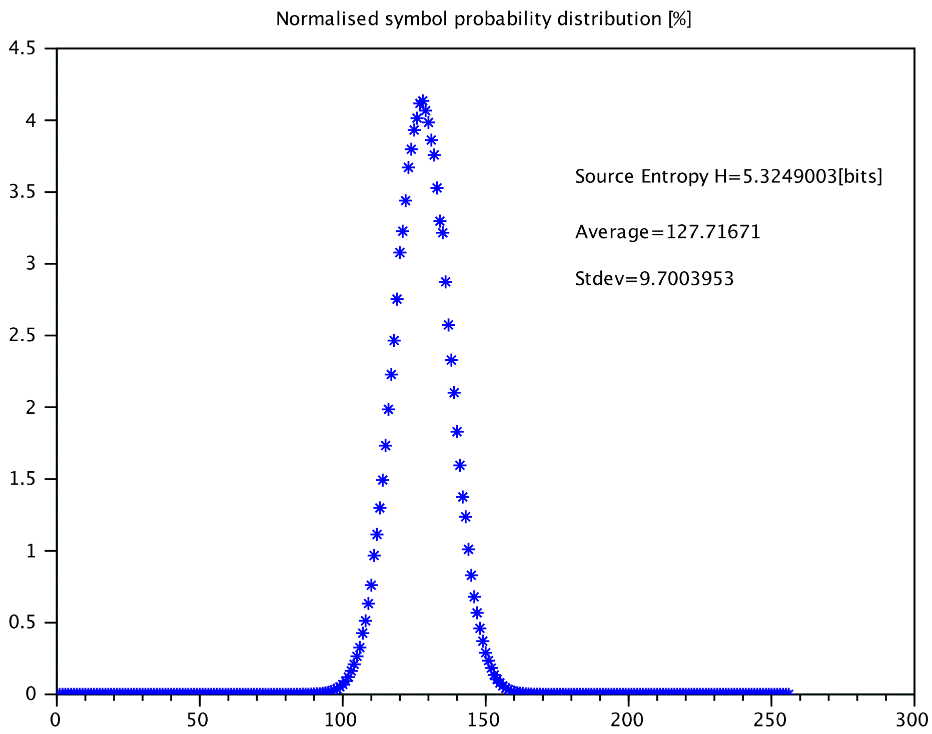
$$L = H(A,P)$$



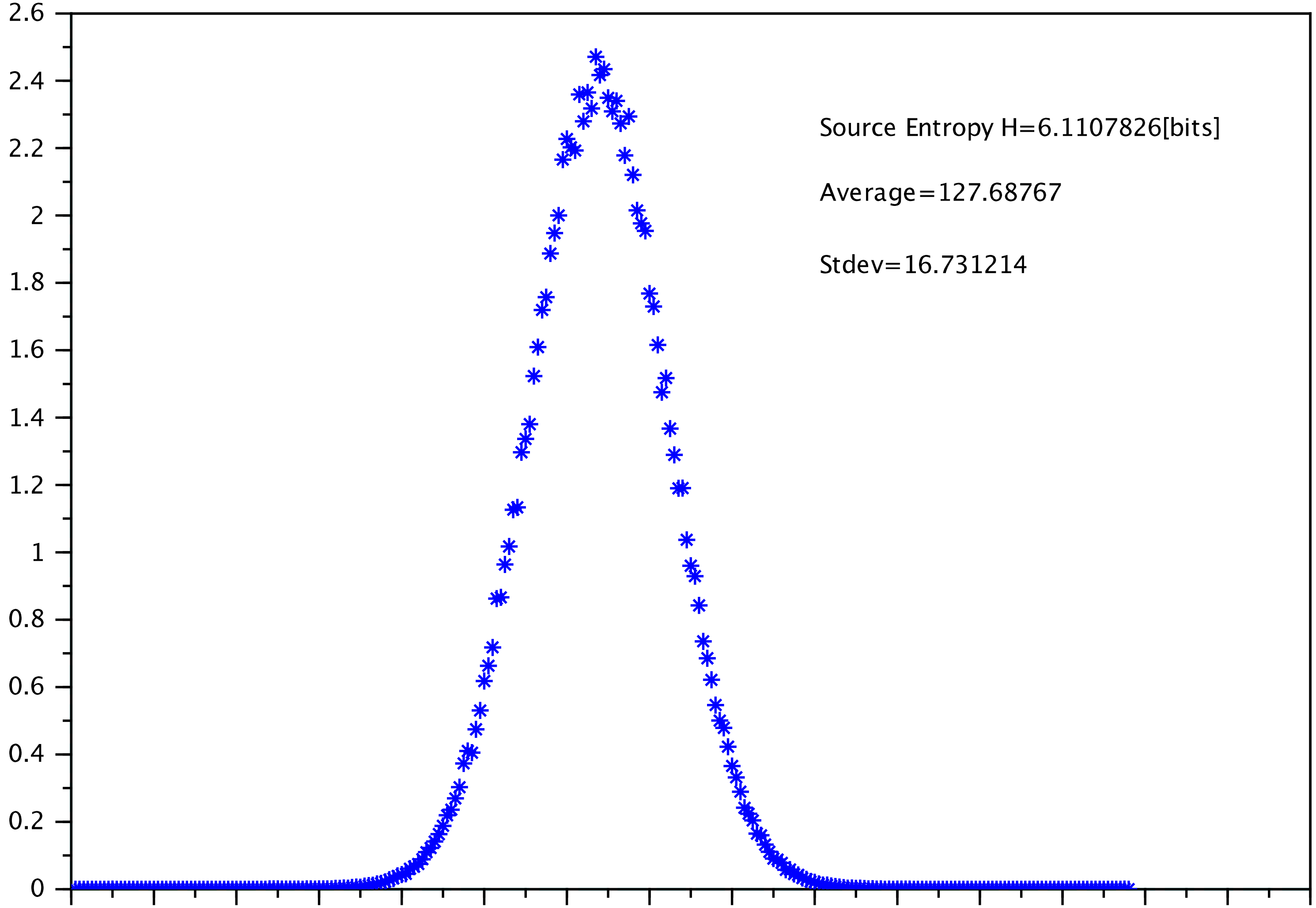
Died 2001 aged 85 years

Radio Astronomy data sources

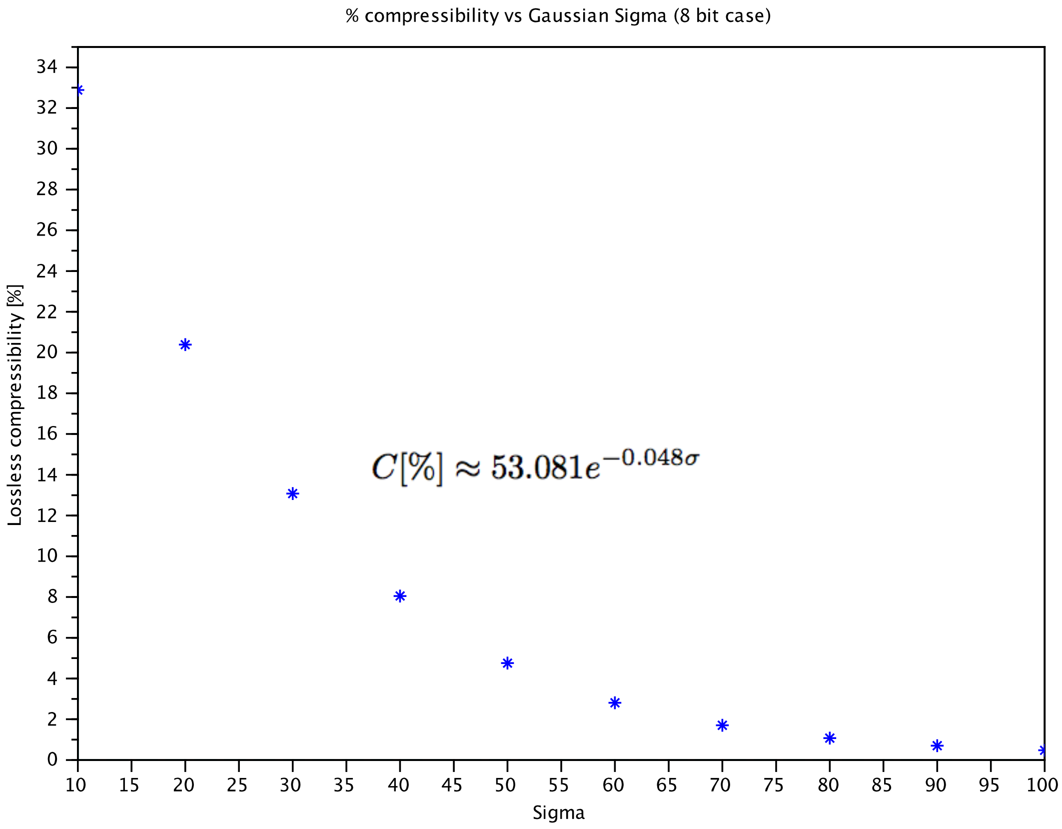
- Typically have a gaussian probability distribution- non equiprobable distribution of states!
- 8 bit example (real)



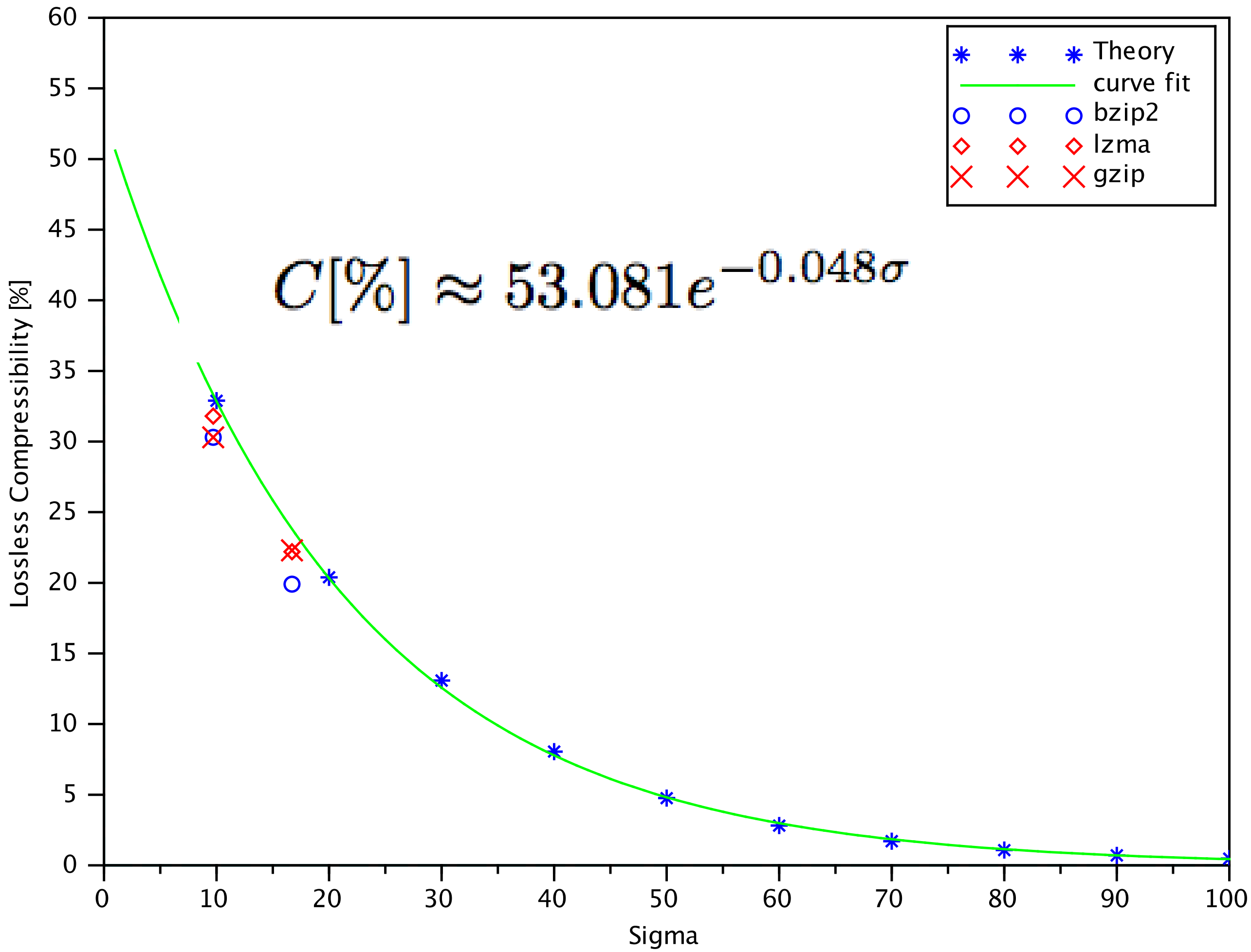
Another file



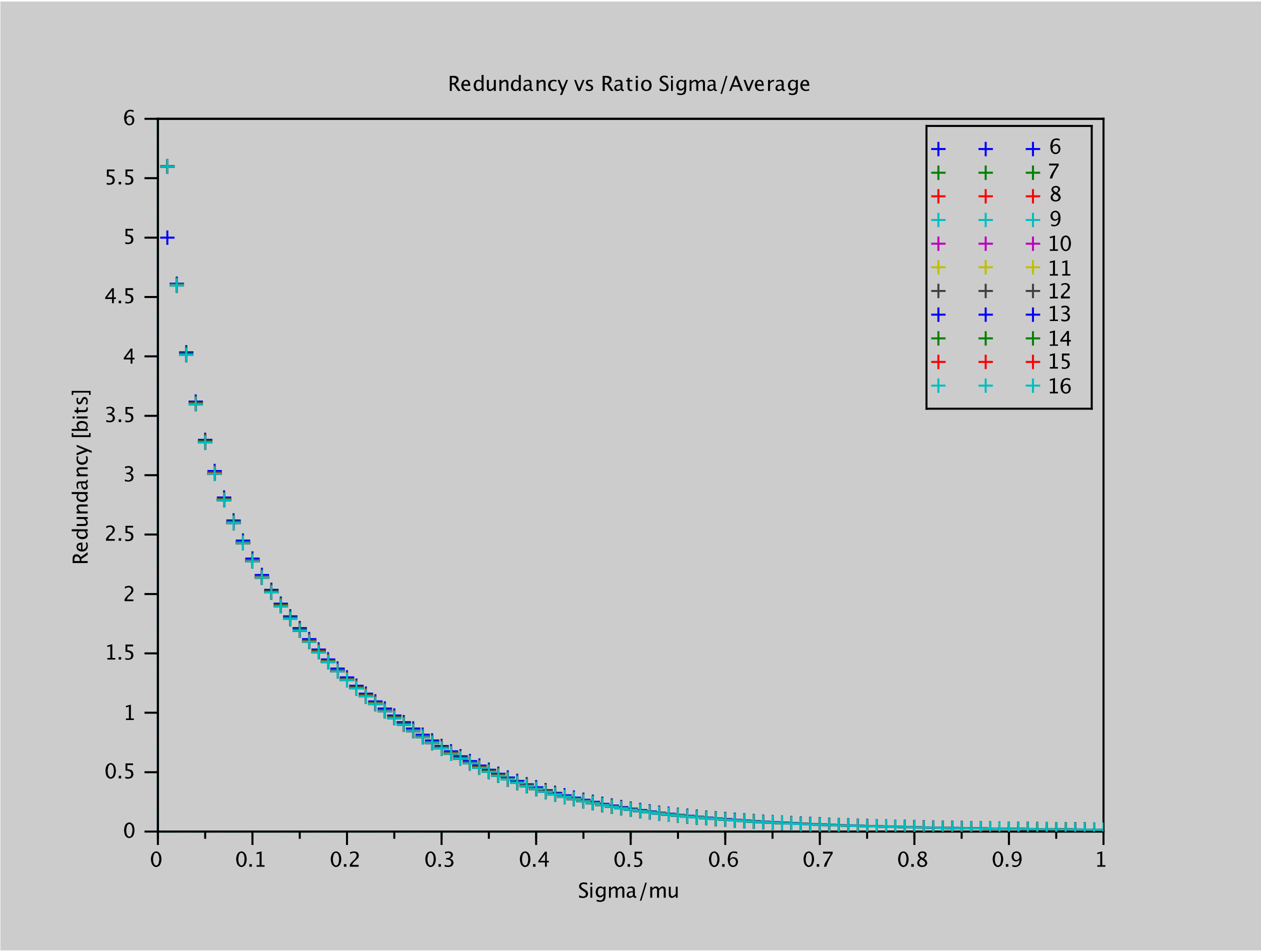
Results from simulations of 8 bit gaussian distributed data

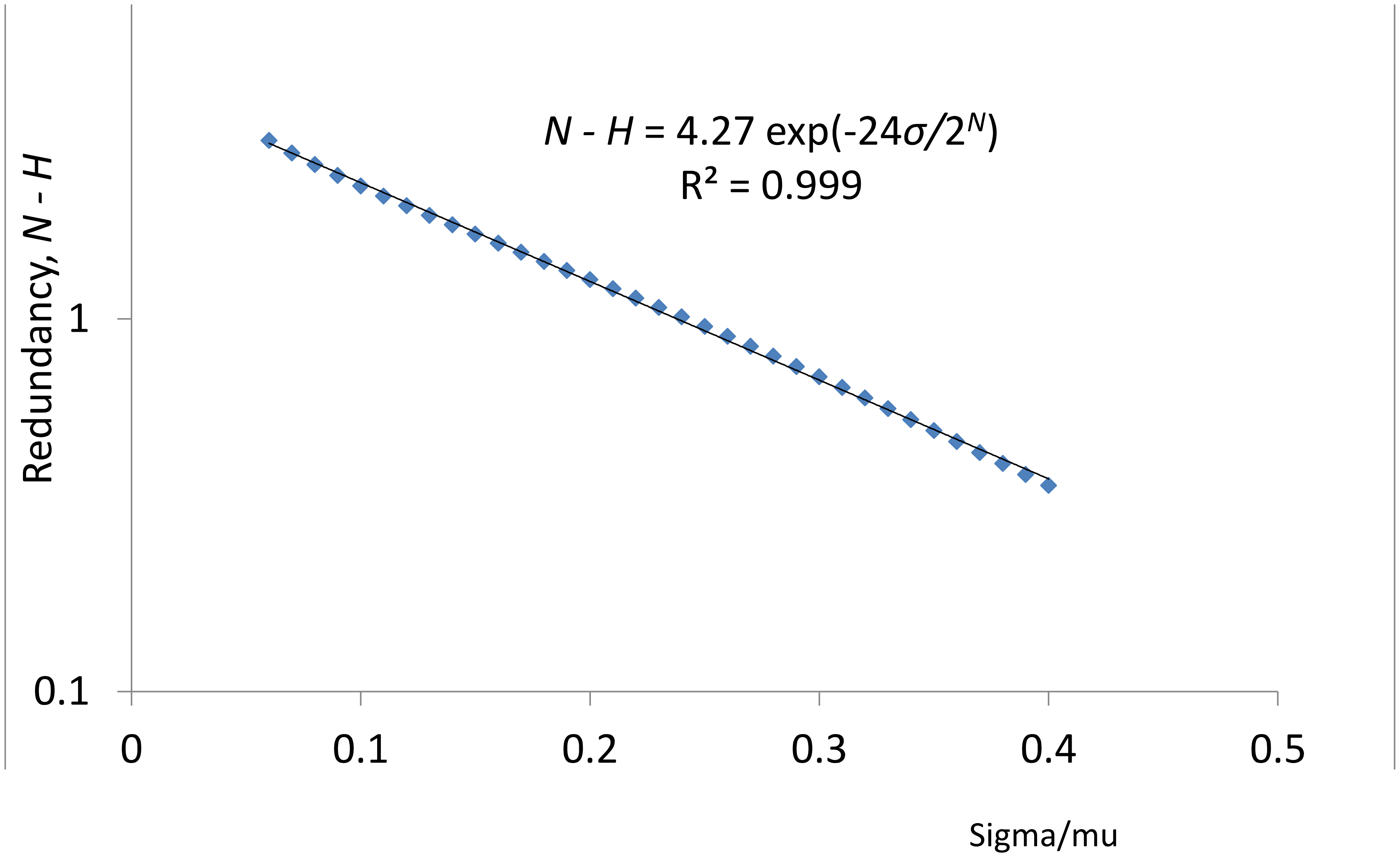


Theory vs practice (8 bit case)



Generalisation to N bit digitisation





N - H = 4.27 exp(-24σ/2^N)
R^2 = 0.999
Redundancy, N - H
1
0.1
0
0.1
0.2
0.3
0.4
0.5
Sigma/mu

Proviso's / assumptions

- Gaussian distribution
- Memoryless source; probability of observing symbol a_i is independent of any previously emitted symbols
- Non memoryless state requires a different definition of entropy and offers yet more scope for compression

$$H(A,P) = \sum_{i=1}^{N} P(a_i) \sum_{j=1}^{N} P(a_j|a_i) log_2(P(a_j|a_i)$$