Is Entropy Guesswork?

David Malone (CNRI DIT)

19 December 2001

Entropy in Information Theory

A source which produces symbols $a \in \mathbb{A}$ with probability p_a has entropy

$$h(p) = \sum_{a \in \mathbb{A}} p_a \lg \frac{1}{p_a}.$$

Entropy is often interpreted as the amount of information or uncertainty associated with a source. It is the average number of bits required to encode a message from that source. It adds for independent sources.

Asymptotic Equipartition

One important place where entropy arises is in the Asymptotic Equipartition Property (AEP). If we have n independent identical sources and we look at their combined output in $\mathbb{A}^n$ then the set

$$T_\epsilon^{(n)} = \left\{a \in \mathbb{A} : |\mathbb{P}(a) - 2^{-nh(p)}| < \epsilon\right\}$$

has the following properties:

- $\mathbb{P}(T_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
- $|T_\epsilon^{(n)}| \approx 2^{nh(p)}$.

These elements are considered 'typical'.

Guessing and Cryptography

Encryption requires selecting an algorithm and a key. Great care is invested in designing algorithms and so it may be easier to attack the key.

- A *brute force attack* involves trying every key one after another. Your key space must be big to make this impractical.
- A *dictionary attack* uses the fact that people are more likely to choose real words as keys.

Pseudo-random numbers used by computers can be subject to dictionary-like attacks if seeded badly.

Entropy and Guessing

Entropy is a measure of uncertainty. Does it capture how hard it is to guess a number? From the `sci.crypt` FAQ:

> We can measure how bad a key distribution is by calculating its entropy. This number E is the number of "real bits of information" of the key: a cryptanalyst will typically happen across the key within 2^E guesses. E is defined as the sum of $-p_K \log_2 p_K$, where p_K is the probability of key K.

The quickest way to guess a symbol is to first guess the most likely value, and then proceed towards the least likely value. Label p_k in decreasing order with integers then the expected amount of guessing time or *guess work* is

$$G(p) = \sum_k p_k k.$$

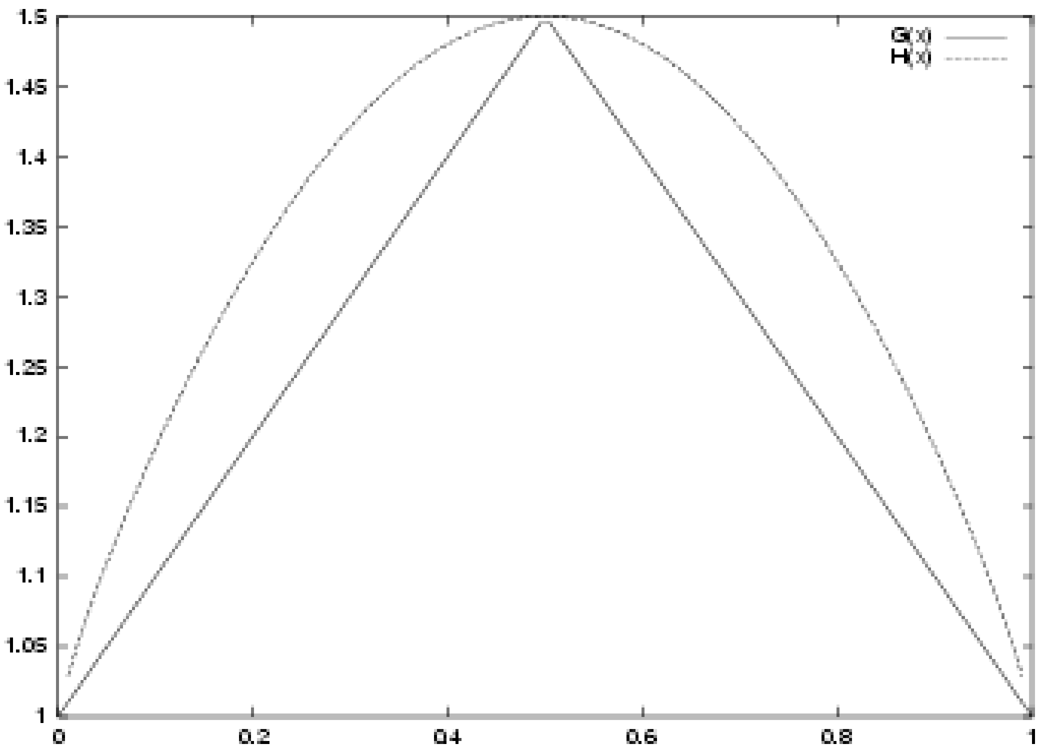
We want to compare this to an entropy based estimate,

$$H(p) = \frac{2^{h(p)} + 1}{2},$$

because guessing from r equally likely options takes $(r + 1)/2$ guesses.

Bernoulli Source

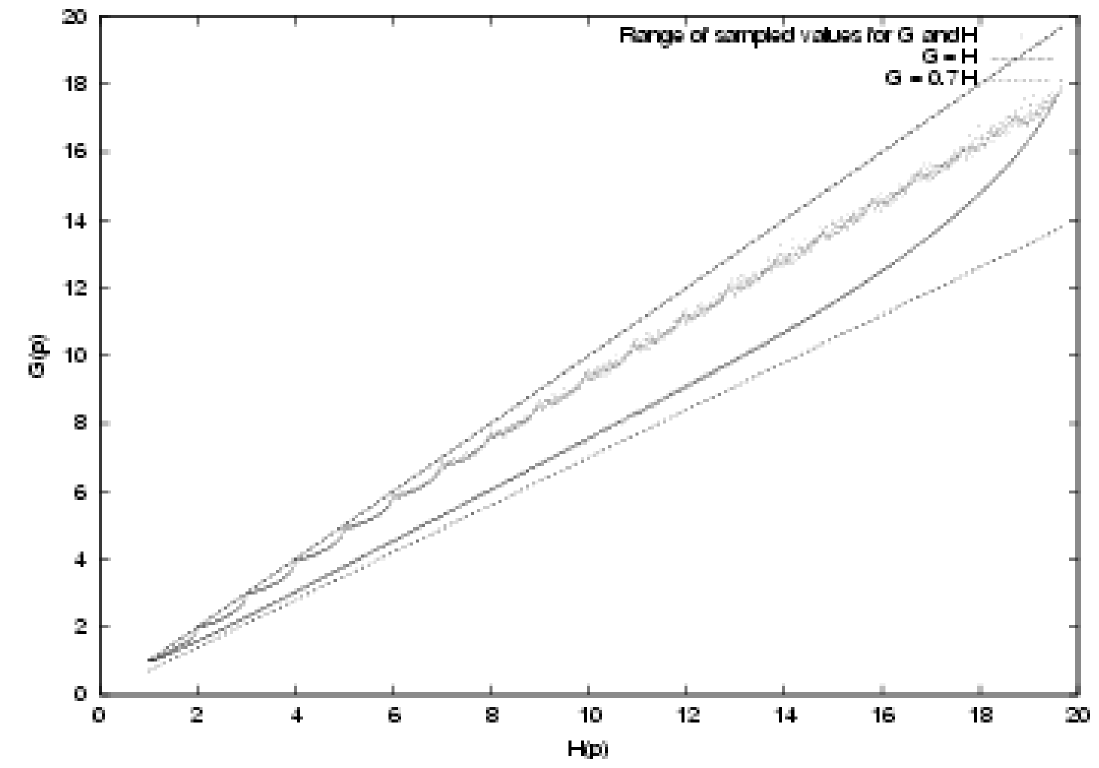
Here $\mathbb{A} = \{0, 1\}$ and
$\mathbb{P}(0) = p, \mathbb{P}(1) = q = 1 - p$.



$$G(p) = \begin{cases} 1p + 2(1-p) & p \geq 0.5 \\ 1(1-p) + 2p & p < 0.5 \end{cases}.$$

$$H(p) = 2^{-p \lg p - (1-p)\lg(1-p)} = p^{-p} q^{-q}.$$

Simulation



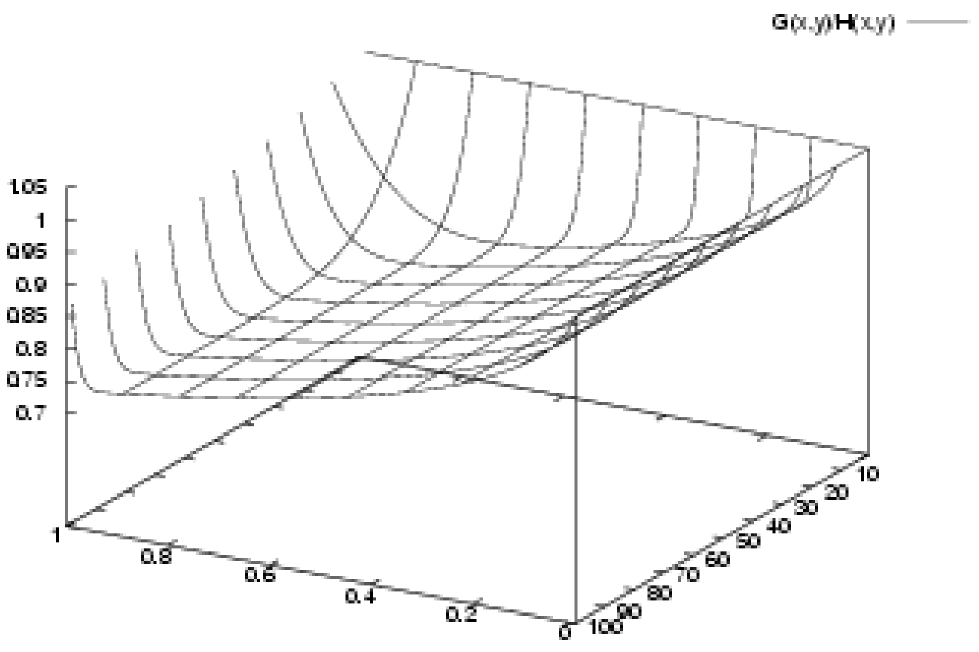
Simulated by choosing up a random distribution on up to 20 symbols. Hypothesis:

$$0.7H(p) \leq G(p) \leq H(p).$$

Show $0.7H(p) \leq G(p)$ with Lagrange Multipliers: Fix $G(p)$ and find extrema of $H(p)$ at

$$p_k = C\lambda^k,$$

(luckily, a decreasing sequence).



Then evaluate G and H explicitly.

$$\lim_{n\to\infty,\lambda\to 1} \frac{G}{H} = \lim_{\lambda\to 1} \frac{2}{1-\lambda+\lambda^{\lambda/(\lambda-1)}} \to \frac{2}{e}$$

Massey also shows that the upper bound $G(p) < H(p)$ isn't, using the sequence

$$p_k = \begin{cases} 1 - \dfrac{\beta}{n} & k = 1 \\ \\ \dfrac{\beta}{n^2 - n} & 2 \leq k \leq n \end{cases},$$

and letting n become large. This sequence has an entropy tending to zero, but a constant guess work.

So, entropy is a lower bound on guess work, but not an upper bound. Luck for those cryptologists...

How did this incorrect idea get into the folklore?

AEP and Guessing

Plaim suggests that the link between guesswork and entropy may have arisen via the AEP. Remember, the AEP says that we can find a set of words $T_\epsilon^{(n)}$ so that the probability of each word is about $2^{-nh(p)}$ and by making n big enough we can make $\mathbb{P}(T_\epsilon^{(n)})$ close to 1. Ignoring the atypical words,

$$G(p) = \sum_k p_k k = \sum_{T_\epsilon^{(n)}} 2^{-nh(p)} k = \frac{2^{nh(p)} + 1}{2}.$$

Setting $n = 1$ then produces folklore...

Reminiscent of replica formalism?

Can we salvage a result if n large?

Look at sets of symbols $(a_1, \dots a_n)$ in $\mathbb{A}^n$ with probability $p_{a_1} \dots p_{a_n}$. Guess in the same was as before and only stop if all symbols correct.

To evaluate $G_n(p)$ calculate all the products $p_{a_1} \dots p_{a_n}$ and sort them, then

$$G_n(p) = \sum_k p_{a_{k,1}} \dots p_{a_{k,n}} k.$$

Evaluating $H_n(p)$ is much easier 'cos the entropy of independent sources adds:

$$H_n(p) = \frac{2^{h_n(p)} + 1}{2} = \frac{2^{nh(p)} + 1}{2} = \frac{H(p)^n + 1}{2}.$$

Is $G_n(p) \approx H_n(p)$?

Product Bernoulli Source

Most cases are hard: have to sort product of probabilities. In Bernoulli case, if $0 \leq p \leq 0.5$, we know $p^k q^{n-k}$ is in non-increasing order. Thus,

$$G_n(p) = \sum_{k=0}^{n} f(k,n) p^k q^{n-k} \binom{n}{k}$$

where

$$f(k,n) = \sum_{j=0}^{k-1} \binom{n}{j} + \frac{1}{2}\binom{n}{k}.$$

$H_n(p)$ grows exponentially so consider

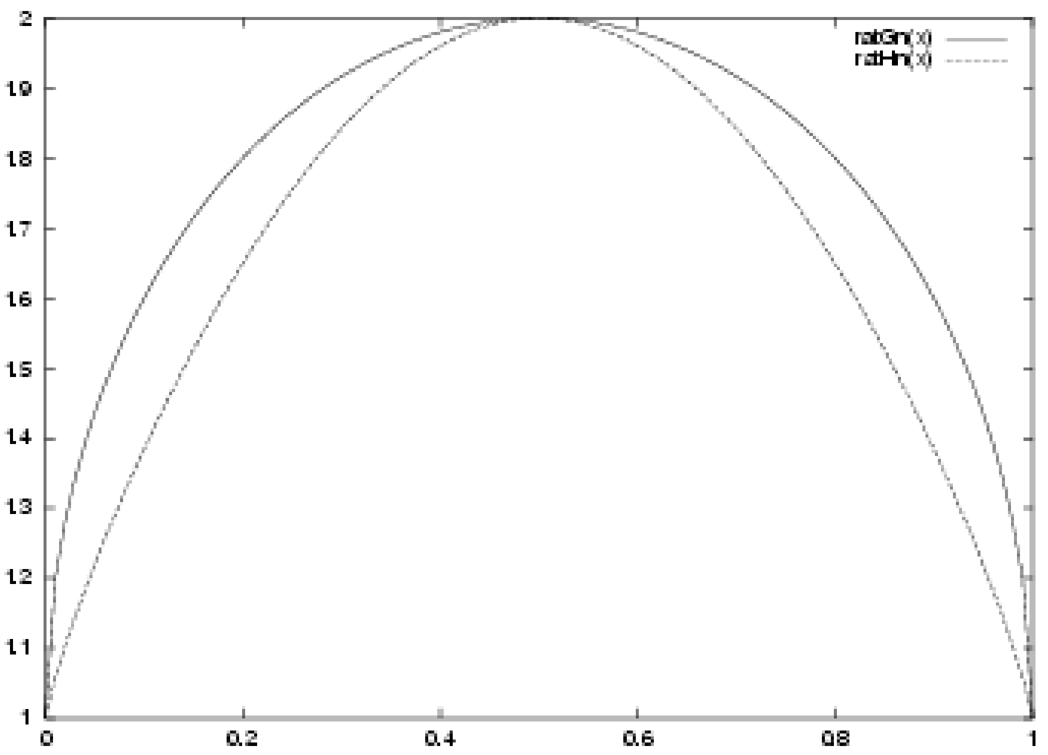
$$\lim_{n\to\infty} \frac{1}{n} \log G_n(p).$$

We find that

$$G_n(p) \asymp \left((\sqrt{p} + \sqrt{q})^2 \right)^n$$

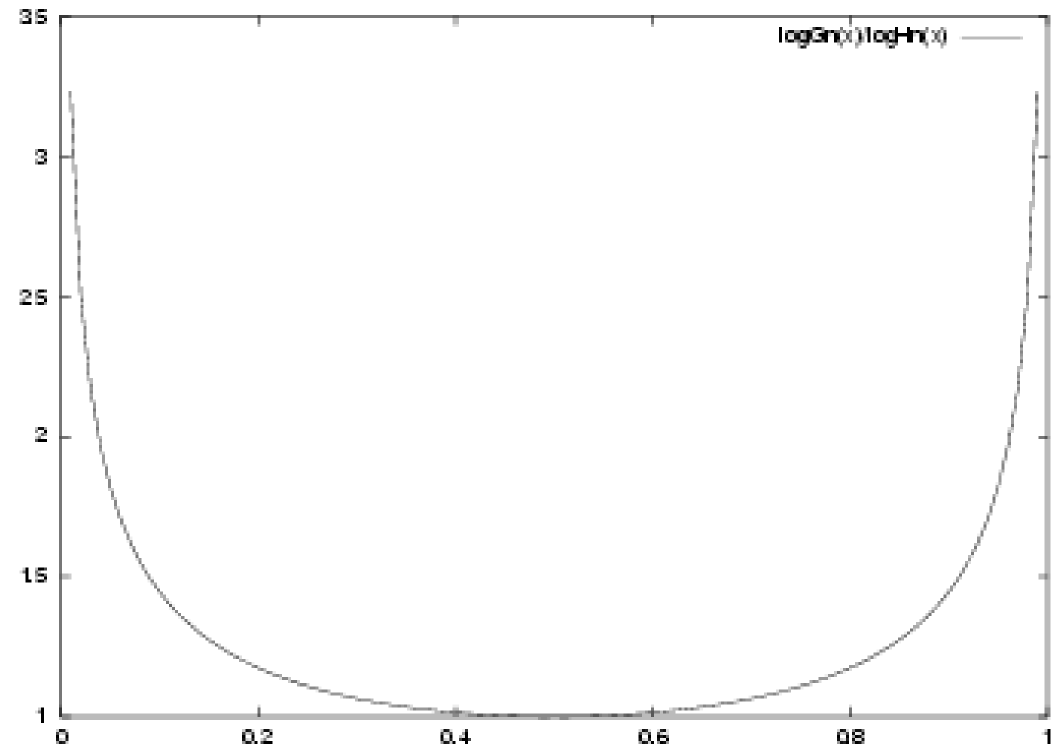
and know that

$$H_n(p) \asymp \left(p^{-p} q^{-q} \right)^n .$$



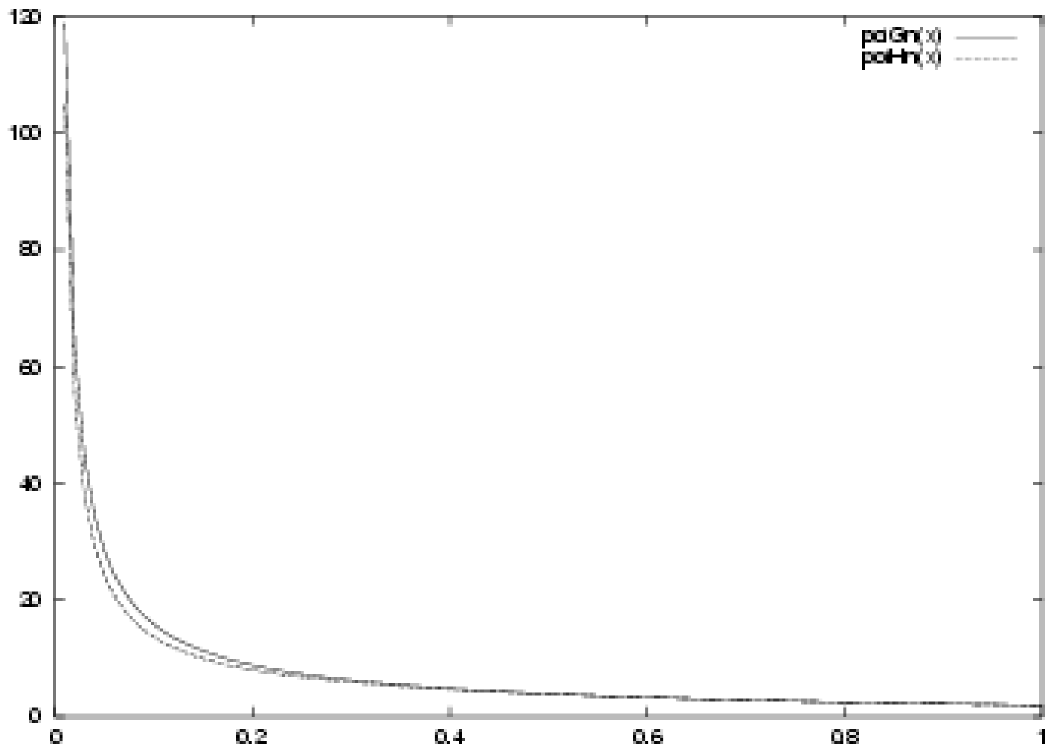
We can also look at

$$\lim_{n\to\infty} \frac{\log G_n(p)}{\log H_n(p)} = \lim_{n\to\infty} \frac{\log G_n}{n} \frac{1}{h(p)}$$



Application

Collecting randomness by measuring background radiation. Watch for time interval T, no decays $a = 0$ otherwise $a = 1$. Poisson distributed so $p = e^{-T}$. Do optimal T for long term rate of entropy and guess work collection differ?



Conclusions

1. Entropy is not guess work, it is easier than guess work. This does not seem to have been spotted until 1994.

2. Simulation didn't pick this out, but should have. Used $p = (U_1, \ldots, U_r)/\sum U_k$, with U_k uniform in $[0, 1]$.

3. Even for distributions with AEP, entropy is not guess work. We can make them similar by sticking to the typical sets and giving up if we don't guess correctly.

4. Massey's example shows that guess work may not be a good measure of guessability. Plaim also suggests work factor:

$$wf_\alpha(p) = \inf\{n : \sum_{k=1}^{n} p_k > \alpha\}.$$

5. For Bernoulli sources have guess work entropy based entropy $\left(\sqrt{p} + \sqrt{q}\right)^2$. Can this be extended to other independent cases?

6. It would be fun to calculate $H(p)$ and $G(p)$ for some human chosen passwords.