The Disparity between Work and Entropy in Cryptology

John O. Pliam[*]

February 1, 1999

Contents

Abstract

A brief theory of work is developed. In it, the work-factor and guesswork of a random variable are linked to intuitive notions of time complexity in a brute-force attack. Bounds are given for a specific work-factor called the minimum majority. Tight bounds are given for the guesswork in terms of variation distance. Differences between work-factor, guesswork and the entropy of a random variable are pointed out, calling into question a common misconception about entropy indicating work.

The purpose of this document is to discuss measures of work and uncertainty relevant to cryptology. In [Pli99], these ideas are used to develop the necessary machinery for quantifying the fundamental limits facing a cryptanalytic adversary in a cipher attack. This machinery plays a role analogous to that of the basic inequalities of information theory [CT91] (e.g. Jensen's inequality, the information inequality, etc.).

1 Definitions of Work

It is important to note at the outset that entropy, the classical measure of uncertainty in cryptology and information theory, is inappropriate as a measure of work in a cipher attack. This assertion is justified in section 2.3, but the idea behind it is quite simple. Due to the *asymptotic equipartition property (AEP)*, entropy gives the effective number of outcomes when a long sequence of values of a random variable are drawn in succession. By contrast, in a cipher attack, the adversary typically faces the problem of guessing the outcome of a single value of a random variable. In this case, uncertainty is more about the shape of the distribution. This shape is known in economics as the the *Lorenz curve* [Pea92], where it is used to study inequality in income distribution. Lorenz's idea [Lor05] of plotting cumulative fraction of total income as a function of percentile, intuitively captures many of the same issues that would concern a cryptologist. Cumulative probability in a brute-force attack represents the chance of success as a function of the number of trials. Lorenz's approach to studying uncertainty can be historically viewed as the starting point for the theory of *majorization* (cf. [MO79]). Not surprisingly, the inequalities of majorization play a crucial role in [Pli99].

The notions of work discussed here have appeared before. In a broad sense they are intimately connected to Lorenz's theory of wealth. A treatment in the context of cryptology is given in Cachin's Ph.D. thesis [Cac97]. In particular, the thesis cites Massey [Mas94] for the first formulation in cryptology of *guessing entropy* which is what we shall call the *guesswork* of a random variable.[1]

[*] Department of Control Science & Dynamical Systems, University of Minnesota, E-Mail: `pliam@ima.umn.edu`

[1] We resist calling guesswork "guessing entropy" because theorem 3 below is so closely analogous to Shannon's first theorem that the natural analogue of guesswork is really the expected codeword length of Shannon's theorem. It perhaps makes more sense to call variation distance a kind of entropy, because *it* appears in the upper and lower bounds of theorem 3, just as entropy does in Shannon's theorem. Thus to call guesswork "guessing entropy" might lead to confusion.

1.1 Guesswork in Optimal Brute-Force Attacks

Let $\mathscr{X}$ be a finite set and suppose that X is the $\mathscr{X}$-valued random variable determined by probability distribution p. We may arrange $\mathscr{X}$ so that the probabilities $p_i = p(x_i)$ satisfy

$$p_1 \geq p_2 \geq \ldots \geq p_{|\mathscr{X}|}. \tag{1}$$

Many situations in cryptology and computer security force an adversary to conduct a brute-force attack in which the values of $\mathscr{X}$ are enumerated and tested for a certain success condition. The only possible luxury afforded the adversary is that he may know which events are more likely. For example, UNIX passwords are routinely guessed with the aid of a public-domain software package called `crack` [Muf]. The technique used goes far beyond the traditional *dictionary attack* in which all words from a list are tested in no particular order. In fact, `crack` constructs a sequence of generalized "dictionaries" by means of formulas for combining other more primitive dictionaries. The sequence of formulas is carried out in order of nonincreasing probability, where empirical observations have been used to determine likelihood. The most probable passwords are in fact, words in a dictionary. The next most probable passwords are of the form

word`[0-9]`

and so on. This process must still be considered *brute-force* because nothing about the UNIX password encryption function is used. The formulas are sufficiently general so that `crack` can be configured to be as close to optimal as desired. The safest bet for the cryptographer is to assume that the adversary has complete knowledge of p and will conduct any brute-force attack *optimally*, i.e. in the order given by equation (1). This suggests the following definitions.

Definition 1 *Let X be an $\mathscr{X}$-valued random variable whose probabilities are arranged according to equation (1). The α-**work-factor** or sometimes simply the **work-factor** of X is given by*

$$\mathrm{wf}_\alpha(X) = \min\left\{|\mathscr{Y}| \,\middle|\, \mathscr{Y} \subseteq \mathscr{X},\ p(\mathscr{Y}) \geq \alpha\right\}.$$

*The **guesswork** of X is given by*

$$W(X) = \sum_{i=1}^{|\mathscr{X}|} i p_i.$$

Since

$$\mathrm{wf}_\alpha(X) = i, \quad \sum_{j=1}^{i-1} p_j < \alpha \leq \sum_{j=1}^{i} p_j,$$

it is readily verified that

$$W(X) = \int_0^1 \mathrm{wf}_\alpha(X)\, d\alpha.$$

An adversary who must be guaranteed at least probability α of discovering the value of X has at least $\mathrm{wf}_\alpha(X)$ work to do. One who tries values of $\mathscr{X}$ in order of nonincreasing probability can expect to do $W(X)$ work. The next two simple algorithms formalize these notions. The adversary is assumed to have access to the necessary optimal enumerators and oracles which tell whether they have guessed correctly. The first demonstrates the computational meaning behind the α-work-factor.

Algorithm 1 *Optimal brute-force attack against X with chance of success α and complexity $O(\mathrm{wf}_\alpha(X))$.*

```
input: (i). an enumerator of the values of $\mathscr{Y} \subseteq \mathscr{X}$, with $p(\mathscr{Y}) \geq \alpha$. (ii). an
    oracle which answers whether $X = x$.

output: Either the value of $X$, or if unsuccessful, $\emptyset$.

for $x \in \mathscr{Y}$ do
  if $X = x$ then
    return $x$.
  endif
done
return $\emptyset$.
```

The next algorithm demonstrates the computational meaning behind guesswork.

Algorithm 2 *Optimal brute-force attack against X which will always succeed and has average complexity $O(W(X))$.*

```
input: (i). An enumerator of the values of $\mathscr{X}$ in order of nonincreasing
    probability. (ii). An oracle which answers whether $X = x$.

output: The value of $X$.

for $x \in \mathscr{X}$ do
  if $X = x$ then
    return $x$.
  endif
done
```

Clearly, the probability of success of Algorithm 1 is

$$\sum_{x \in \mathscr{Y}} p(x) = p(\mathscr{Y}) \geq \alpha,$$

and the average computation time of Algorithm 2 is

$$\sum_{i=1}^{|\mathscr{X}|} i p_i = W(X).$$

Thus our definitions of $\mathrm{wf}_\alpha(X)$ and $W(X)$ formally meet the expectations of their everyday meanings.

As alluded to earlier, these two notions of work are intimately connected to the measure of economic inequality by means of the Lorenz curve. Consider the plot of the cumulative probability of success in a brute-force guessing attack as a function of the number of trials.[2] After transposing axes, this is the plot of $\mathrm{wf}_\alpha(X)$ as a function of α as shown in figure 1. The area under $\mathrm{wf}_\alpha(X)$ is $W(X)$, and the area between $\mathrm{wf}_\alpha(X)$ and $\mathrm{wf}_\alpha(U)$ would seem to be an reasonable measure of the "distance to uniformity." In fact, we shall show in section 2.2 that this area is within a factor of 2 of the variation distance between p_X and $u = p_U$.

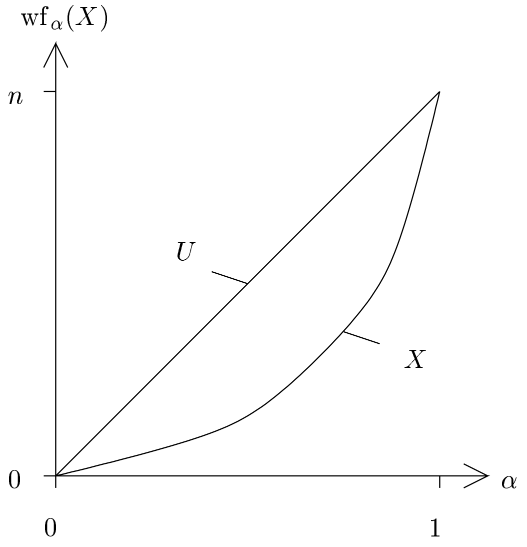

Figure 1: Lorenz-like curve of $\mathrm{wf}_\alpha(X)$ as function of α in comparison to the similar curve for the uniform random variable U. The area under the $\mathrm{wf}_\alpha(X)$ curve is $W(X)$.

[2] The Lorenz curve is actually a plot of cumulative wealth as a function of percentile group given in *increasing* (or nondecreasing) order. Despite this difference, the Lorenz curve and the probability curve both measure a clustering of the distribution.

1.2 Majorization and Strict Majorization

We will occasionally work with vectors of real numbers which are not probability distributions (their sum isn't 1 or their component values aren't positive). For $x = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$, recall definitions of the 1-*norm*, 2-*norm* and ∞-*norm* (or *infinity-norm*) respectively given by

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^n |x_i|, \\ \|x\|_2 &= \sum_{i=1}^n x_i^2, \\ \|x\|_\infty &= \max_{1 \le i \le n} |x_i|. \end{aligned}$$

Following [MO79], when $x \in \mathbb{R}^n_+$ it may always be rearranged into the *decreasing rearrangement* of x given by

$$x_\downarrow = (x_{[1]}, x_{[2]}, \ldots, x_{[n]}),$$

where

$$x_{[1]} \ge x_{[2]} \ge \ldots \ge x_{[n]}.$$

Similarly we have the *increasing rearrangement* of x given by

$$x^\uparrow = (x_{(1)}, x_{(2)}, \ldots, x_{(n)}),$$

where

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}.$$

Given vectors $x, y \in \mathbb{R}^n_+$, we say that x is *majorized* by y, or that y *majorizes* x if

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \tag{2}$$

and for each k between 1 and n,

$$\sum_{i=1}^k x_{[i]} \le \sum_{i=1}^k y_{[i]}. \tag{3}$$

That x is majorized by y will be written (cf. [MO79])

$$x \preceq y.$$

Remark 1 *Be warned that here we stray from the ubiquitous notation for majorization introduced by Hardy, Littlewood and Pólya (see [HLP52] and [MO79]). We reserve the usual notation $x \prec y$ for strict majorization, which is now introduced.*

Definition 2 *If for $x, y \in \mathbb{R}^n_+$, $x \preceq y$ while $y \npreceq x$, then we say that x is* **strictly majorized** *by y, and write*

$$x \prec y.$$

Of fundamental importance are two theorems which characterize majorization "algebraically."

Theorem 1 (Hardy-Littlewood-Pólya, 1929) *For $x, y \in \mathbb{R}^n_+$,*

$$x \preceq y \iff x = Dy,$$

for some $n \times n$ doubly stochastic matrix D.

Theorem 2 (Birkhoff, 1946) *Every doubly stochastic matrix D is a convex sum of permutation matrices, i.e.,*

$$D = \sum_i p_i \Pi_i, \quad \text{where} \quad \sum_i p_i = 1,$$

and where each Π_i is a permutation matrix.

The Hardy-Littlewood-Pólya theorem originally appeared in 1929 [HLP29]. The Birkhoff theorem first appeared in 1946 [Bir46], but was independently discovered in 1953 by von Neumann [von53] who applied it to a problem in game theory. It is sometimes called the Birkhoff-von Neumann theorem [BR97]. Recall that the *convex hull* of a set of vectors $B = \{v_1, v_2, \ldots, v_k\}$ is defined by

$$\Omega(B) = \left\{ \sum_{i=1}^{k} p_i v_i \,\middle|\, \sum_{i=1}^{k} p_i = 1 \right\}.$$

Thus Birkhoff's theorem identifies the set of doubly stochastic matrices with the convex hull of permutation matrices. It is a standard observation that $x \preceq y$ and $y \preceq x$ iff $x = \Pi y$, for some permutation matrix Π. In light of this, the Hardy-Littlewood-Pólya theorem and the Birkhoff theorem, we immediately have the following consequence of strict majorization.

Proposition 1 *If $x \prec y$, then $x = Dy$, where D is a nontrivial convex sum of permutation matrices.*

However, the converse is not in general true, because for any doubly stochastic D,

$$e = De, \quad \text{but} \quad e \not\prec e,$$

where $e = (1, 1, \ldots, 1)$. Strict majorization leads to useful strict inequalities, for example (see [MO79] or [BR97])

$$x \prec y \Rightarrow \|x\|_2 < \|y\|_2.$$

It is natural to seek a kind of converse to proposition 1. This is taken up in [Pli99].

1.3 Work and Majorization

It is useful to generalize the definition of work-factor and guesswork for arbitrary vectors in $\mathbb{R}_+^n$.

Definition 3 *For $x \in \mathbb{R}_+^n$ and real $0 \leq \alpha \leq 1$, the α-**work-factor** of x is given by*

$$\mathrm{wf}_\alpha(x) = \min \left\{ k \,\middle|\, \sum_{i=1}^{k} x_{[i]} \geq \alpha \|x\|_1 \right\}$$

*and the **guesswork** of x is given by*

$$W(x) = \sum_{i=1}^{n} i x_{[i]}.$$

Again it is readily verified that

$$W(x) = \|x\|_1 \int_0^1 \mathrm{wf}_\alpha(x)\, d\alpha.$$

The following simple observation pertains to uniformly distributed random variables.

Proposition 2 *If $x_i = \frac{1}{n}$, for every component of $x \in \mathbb{R}_+^n$, then*

$$W(x) = \frac{n+1}{2},$$

and

$$\mathrm{wf}_\alpha(x) = \lceil \alpha n \rceil.$$

Proof: Evidently,

$$W(x) = \sum_{i=1}^{n} i x_{[i]} = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

Also,

$$\sum_{i=1}^{\lceil \alpha n \rceil} x_{[i]} = \frac{\lceil \alpha n \rceil}{n} \geq \alpha,$$

while

$$\sum_{i=1}^{\lceil \alpha n\rceil - 1} x_{[i]} = \frac{\lceil \alpha n\rceil - 1}{n} < \alpha,$$

so that $\mathrm{wf}_\alpha(x) = \lceil \alpha n \rceil$. □

Some important inequalities involving work and majorization are elementary.

Proposition 3 *If $x \preceq y$, then*

$$W(x) \geq W(y),$$

and for each α

$$\mathrm{wf}_\alpha(x) \geq \mathrm{wf}_\alpha(y).$$

Proof: From the definitions of α-work-factor (above) and majorization (equations 2 and 3), it follows that

$$\begin{aligned} \sum_{i=1}^{\mathrm{wf}_\alpha(x)} y_{[i]} &\geq \sum_{i=1}^{\mathrm{wf}_\alpha(x)} x_{[i]} \\ &\geq \alpha \|x\|_1 \\ &= \alpha \|y\|_1. \end{aligned}$$

Therefore $\mathrm{wf}_\alpha(y)$ is at most $\mathrm{wf}_\alpha(x)$. Now using this we have

$$\begin{aligned} W(x) &= \|x\|_1 \int_0^1 \mathrm{wf}_\alpha(x)\, d\alpha \\ &\geq \|y\|_1 \int_0^1 \mathrm{wf}_\alpha(y)\, d\alpha \\ &= W(y), \end{aligned}$$

which completes the proof. □

Strict inequalities of this form are also possible.

Proposition 4 *If $x \prec y$, then*

$$W(x) > W(y)$$

and for some λ,

$$\mathrm{wf}_\lambda(x) > \mathrm{wf}_\lambda(y).$$

Proof: $x \prec y$ means that $x_\downarrow \neq y_\downarrow$ and so there must be an $m < n$ for which

$$\sum_{i=1}^{m} x_{[i]} < \sum_{i=1}^{m} y_{[i]}.$$

If we take

$$\lambda = \frac{1}{\|y\|_1} \sum_{i=1}^{m} y_{[i]},$$

then

$$\sum_{i=1}^{m} x_{[i]} < \lambda \|y\|_1 = \lambda \|x\|_1,$$

which means that $\mathrm{wf}_\lambda(x)$ is strictly more than m. But $\mathrm{wf}_\lambda(y)$ is at most m, so that

$$\mathrm{wf}_\lambda(x) > \mathrm{wf}_\lambda(y).$$

Clearly for some $\varepsilon > 0$ we must have

$$\mathrm{wf}_\alpha(x) > \mathrm{wf}_\alpha(y), \quad \lambda \leq \alpha \leq (\lambda + \varepsilon).$$

From this it follows that

$$\begin{aligned} W(x) &= \|x\|_1 \int_0^1 \mathrm{wf}_\alpha(x)\, d\alpha \\ &> \|y\|_1 \int_0^1 \mathrm{wf}_\alpha(y)\, d\alpha \\ &= W(y), \end{aligned}$$

which completes the proof. □

There is a converse to proposition 4 for guesswork. First we need the following.

Proposition 5 *The α-work-factor and guesswork are permutation invariant.*

Proof: The quantities $\mathrm{wf}_\alpha(x)$ and $W(x)$ only involve the components of $x_\downarrow$, namely $\{x_{[i]}\}$. From the fact that for any permutation σ, $x_\downarrow = (\sigma x)_\downarrow$, it immediately follows that $W(\sigma x) = W(x)$, and $\mathrm{wf}_\alpha(\sigma x) = \mathrm{wf}_\alpha(x)$, which was to be proved. □

Now, we have following useful characterization of strict majorization.

Proposition 6 *If D is doubly stochastic,*

$$Dx \prec x \iff W(Dx) > W(x).$$

Proof: ($\Rightarrow$:) Follows from proposition 4. ($\Leftarrow$:) Assume that $Dx \not\prec x$. Then D acts as a permutation on x, so $W(Dx) \not> W(x)$. □

2 Bounds on Work

In this section we establish bounds on measures of work in terms of more fundamental statistical quantities such as vector space norms. An upper bound to a measure of work is relevant to cryptanalytic circumstances when one wants to say that no more than a certain amount of work is necessary to guess the key of a cipher. A lower bound to a measure of work is relevant to cryptographic circumstances when one wants to say that at least a certain much work is necessary.

The situation is analogous to Shannon's first theorem in which the average codeword length (the thing you want to know) is bounded above and below by expressions involving entropy (the thing you can often compute). The analogy is in fact quite strong, and it will be shown that the variation distance in a sense "takes the place of" the Kullback-Leibler distance.

2.1 Bounds on the Minimum Majority Work-Factor

When considering the α-work-factor of a random variable, the value $\alpha = \frac{1}{2}$ occupies a special place. One reason is that when α is $\frac{1}{2}$, algorithm 1 has the smallest complexity of all such algorithms with at least an even chance of success. For a random variable X, the work-factor $\mathrm{wf}_{\frac{1}{2}}(X)$ is known as the *minimum majority* [MO79] and has its origins in mathematical political science [Akl65], where it is used to address the problem of targeting districts in an election. For example, as the name suggests, the minimum majority gives the minimum number of districts required to win an election in an electoral college system. In that case, we take p_i to be the i-th district's fraction of electoral votes.

Again let $\mathscr{X}$ be a finite set with measures p and q defined on it. Recall ([CT91] or [Dia88]) the *variation distance* between p and q defined by

$$\|p - q\| = \max_{\mathscr{Y} \subseteq \mathscr{X}} |p(\mathscr{Y}) - q(\mathscr{Y})|. \tag{4}$$

It is a standard observation that

$$\|p - q\| = \frac{1}{2}\|p - q\|_1,$$

and that the maximum in equation (4) is achieved on the set

$$\mathscr{Y} = \{x \in \mathscr{X} \mid p(x) \geq q(x)\}\ .$$

Furthermore, if u is the uniform measure on $\mathscr{X}$ and

$$A = \left\{ x \in \mathscr{X} \,\middle|\, p(x) \geq \frac{1}{|\mathscr{X}|} \right\},$$

then

$$\|p - u\| = p(A) - \frac{|A|}{|\mathscr{X}|} = \frac{|\overline{A}|}{|\mathscr{X}|} - p(\overline{A}).$$

Upper and lower bound on the minimum majority work-factor can now be given.

Proposition 7 *Let X be the $\mathscr{X}$-valued random variable defined by probability distribution p. Then*

$$\left\lfloor \frac{1}{2\|p\|_\infty} \right\rfloor \leq \mathrm{wf}_{\frac{1}{2}}(X) \leq \lceil (1 - \|p - u\|)|\mathscr{X}| \rceil .$$

Proof: The easier lower bound is dispensed with first. Let U be the first

$$\left\lfloor \frac{1}{2\|p\|_\infty} \right\rfloor - 1$$

elements of $\mathscr{X}$ in order of nonincreasing probability. Then

$$\begin{aligned} p(U) &\leq |U|\, \|p\|_\infty \\ &\leq \left(\frac{1}{2\|p\|_\infty} - 1 \right) \|p\|_\infty \\ &= \frac{1}{2} - \|p\|_\infty \\ &< \frac{1}{2}. \end{aligned}$$

Thus,

$$\left\lfloor \frac{1}{2\|p\|_\infty} \right\rfloor - 1 < \mathrm{wf}_{\frac{1}{2}}(X),$$

or

$$\left\lfloor \frac{1}{2\|p\|_\infty} \right\rfloor \leq \mathrm{wf}_{\frac{1}{2}}(X).$$

Consider now the upper bound, and let W be the first $\lceil (1 - \|p - u\|)|\mathscr{X}| \rceil$ elements of $\mathscr{X}$ in order of nonincreasing probability. Establishing

$$p(W) \geq \frac{1}{2}, \text{ or equivalently, } p(\overline{W}) \leq \frac{1}{2}$$

will prove the lemma. Now,

$$\begin{aligned} |W| &= \lceil (1 - \|p - u\|)|\mathscr{X}| \rceil \\ &\geq (1 - \|p - u\|)|\mathscr{X}| \\ &= \left(1 - p(A) + \frac{|A|}{|\mathscr{X}|} \right) |\mathscr{X}| \\ &= |A| + p(\overline{A})|\mathscr{X}|. \end{aligned}$$

Thus $|W| \geq |A|$ with equality iff $\overline{A}$ is empty. That can only happen if $p = u$, in which case the upper bound holds trivially. From this point on, we assume that X is not uniformly distributed. Because A is a proper subset of W, we may define nonempty

$$V = W - A.$$

For any real number x, $x \leq \lceil x \rceil < (x+1)$, and thus

$$\begin{array}{rcccl} (1 - \|p-u\|)|\mathscr{X}| & \leq & |W| & < & (1 - \|p-u\|)|\mathscr{X}| + 1, \\ 1 - \|p-u\| & \leq & \frac{|W|}{|\mathscr{X}|} & < & 1 - \|p-u\| + \frac{1}{|\mathscr{X}|}, \\ 1 - p(A) + \frac{|A|}{|\mathscr{X}|} & \leq & \frac{|A|}{|\mathscr{X}|} + \frac{|V|}{|\mathscr{X}|} & < & 1 - p(A) + \frac{|A|}{|\mathscr{X}|} + \frac{1}{|\mathscr{X}|}, \\ p(\overline{A}) & \leq & \frac{|V|}{|\mathscr{X}|} & < & p(\overline{A}) + \frac{1}{|\mathscr{X}|}, \\ p(\overline{W}) + p(V) & \leq & \frac{|V|}{|\mathscr{X}|} & < & p(\overline{W}) + p(V) + \frac{1}{|\mathscr{X}|}, \\ p(\overline{W}) & \leq & \frac{|V|}{|\mathscr{X}|} - p(V) & < & p(\overline{W}) + \frac{1}{|\mathscr{X}|}, \end{array}$$

which can be expressed as

$$p(\overline{W}) = \frac{|V|}{|\mathscr{X}|} - p(V) - \varepsilon,$$

where

$$0 \leq \varepsilon < \frac{1}{|\mathscr{X}|}.$$

A further constraint on $p(\overline{W})$ comes from the fact that the elements of $\overline{W}$ are less probable than those of V. Thus the average probability on V exceeds that of $\overline{W}$. In other words

$$\frac{p(V)}{|V|} \geq \frac{p(\overline{W})}{|\overline{W}|}.$$

Combining these constraints we obtain

$$p(\overline{W}) \leq \frac{|V|}{|\mathscr{X}|} - |V| \frac{p(\overline{W})}{|\overline{W}|} - \varepsilon,$$

or

$$\begin{aligned} p(\overline{W}) &\leq \left(\frac{|V|}{|\mathscr{X}|} - \varepsilon \right) \frac{1}{\left(1 + \frac{|V|}{|\overline{W}|}\right)} && (5) \\ &= \frac{1}{|\mathscr{X}|} \left(\frac{|V||\overline{W}|}{|V| + |\overline{W}|} \right) - \varepsilon \frac{|\overline{W}|}{|V| + |\overline{W}|} && (6) \\ &\leq \frac{1}{|\mathscr{X}|} \left(\frac{|V| + |\overline{W}|}{2} \right) - \varepsilon \frac{|\overline{W}|}{|V| + |\overline{W}|} && (7) \\ &= \frac{1}{2} \frac{|\overline{A}|}{|\mathscr{X}|} - \varepsilon \frac{|\overline{W}|}{|A|} && (8) \\ &\leq \frac{1}{2}, && (9) \end{aligned}$$

where (7) follows from the fact that the arithmetic mean exceeds the harmonic mean. This completes the proof. □

Remark 2 *The upper bound portion of previous proposition bears a striking resemblance to the upper bound in Shannon's first theorem which we may write as (cf. [CT91] p. 88):*

$$\lfloor L^* \rfloor \leq \lceil \log |\mathscr{X}| - D(p\|u) \rceil, \tag{10}$$

where L^* is the optimal average codeword length, and where $D(p\|q)$ is the Kullback-Leibler distance (or relative entropy) between p and q defined by

$$D(p\|q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}.$$

It is a standard observation that entropy may be defined as $H(p) = \log|\mathscr{X}| - D(p\|u)$, whence equation (10) takes a slightly more familiar form. The Kullback-Leibler distance is analogous to the variation distance, and codeword length is analogous to the minimum majority work-factor. The Kullback-Leibler distance ranges from its minimum *of* 0 *when* $p = u$ *to its maximum of* $\log|\mathscr{X}|$ *when* p *is deterministic and* $H(p) = 0$. *The variation distance ranges from its minimum of* 0 *when* $p = u$ *to its supremum of* 1 *when* p *is deterministic and* $|\mathscr{X}| \to \infty$.

2.2 Tight Bounds on Guesswork

Notice that the uniform distribution is majorized by any other probability distribution. In other words, $u \preceq p$ and so by proposition 3,

$$W(X) \leq \frac{n+1}{2}.$$

The following theorem offers tight upper and lower bounds on the difference between guesswork and maximum guesswork. The analogy to Shannon's first theorem is even more pronounced.

Theorem 3 *For any random variable* X *defined by probability distribution* $p \in \mathbb{R}^n_+$,

$$\frac{n}{2}\|p-u\| \leq \frac{n+1}{2} - W(X) \leq n\|p-u\|. \tag{11}$$

Proof: Without loss of generality, we may assume $p_1 \geq p_2 \geq \ldots \geq p_n$. Define,

$$x_i = \sum_{j=1}^{i} p_j,$$

and

$$q_i = \sum_{j=1}^{i} p_j - \sum_{j=1}^{i} u_i = x_i - \frac{i}{n}.$$

It follows from the definition of equation (4) that

$$\|p-u\| = \max_i q_i.$$

Next we claim that

$$n+1-W(X) = \sum_{i=1}^{n}\sum_{j=1}^{i} p_j.$$

To see this observe that the right hand side is

$$p_1 + (p_1+p_2) + \cdots + (p_1 + \cdots + p_n) = \sum_{i=1}^{n}(n-i+1)p_i,$$

so that

$$W(X) + \sum_{i=1}^{n}(n-i+1)p_i = (n+1)\sum_{i=1}^{n} p_i = n+1.$$

Since

$$\sum_{i=1}^{n} q_i = \sum_{i=1}^{n}\sum_{j=1}^{i} p_j - \frac{n+1}{2},$$

we arrive at a useful formula relating $W(X)$ and q_i:

$$\frac{n+1}{2} - W(X) = \sum_{i=1}^{n} q_i. \tag{12}$$

The upper bound follows immediately from

$$\frac{n+1}{2} - W(X) \leq \sum_{i=1}^{n} q_{\max} = n\|p-u\|.$$

In order to establish the lower bound we shall show that x_i, and therefore q_i, are concave functions of i. For any positive integers a, b,

$$\frac{1}{b-a+1}\sum_{j=a}^{b} p_j = \text{ average probability on } \{a, \ldots, b\}.$$

Because p_i is nonincreasing, it is clear that for $a < i \leq b$,

$$\frac{1}{i-a}\sum_{j=a+1}^{i} p_j = \frac{x_i - x_a}{i-a} \geq \frac{x_b - x_a}{b-a} = \frac{1}{b-a}\sum_{j=a+1}^{b} p_j.$$

Thus for integral $i = \alpha a + \beta b$ with $\alpha + \beta = 1$, we have

$$\begin{aligned} x_i &\geq x_a + \left(\frac{i-a}{b-a}\right)(x_b - x_a) \\ &= \left[1 - \left(\frac{i-a}{b-a}\right)\right] x_a + \left(\frac{i-a}{b-a}\right) x_b \\ &= \left(\frac{b-i}{b-a}\right) x_a + \left(\frac{i-a}{b-a}\right) x_b \\ &= \alpha x_a + \beta x_b. \end{aligned}$$

Therefore q_i is concave and hence satisfies a kind of parallelogram bound,

$$\sum_{i=a}^{b} q_i \geq \frac{q_a + q_b}{2}(b-a+1).$$

Observe that $q_1 = p_1 - 1/n$, $q_n = 0$, and if $q_k = q_{\max} = ||p - u||$, then

$$q_{k-1} = q_k - p_k + \frac{1}{n}.$$

Starting from equation (12) we have,

$$\begin{aligned} \frac{n+1}{2} - W(X) &= \sum_{i=1}^{k-1} q_i + \sum_{i=k}^{n} q_i \\ &\geq \frac{q_1 + q_{k-1}}{2}(k-1) + \frac{q_k + q_n}{2}(n-k+1) \\ &= \frac{q_k + (p_1 - p_k)}{2}(k-1) + \frac{q_k}{2}(n-k+1) \\ &\geq \frac{q_k}{2}(k-1) + \frac{q_k}{2}(n-k+1) \\ &= n\frac{q_k}{2} \quad = \quad \frac{n}{2}||p-u||, \end{aligned}$$

which completes the proof. □

When $||p - u||$ is sufficiently small, the upper and lower bounds of

$$\frac{n+1}{2} - n||p-u|| \leq W(X) \leq \frac{n+1}{2} - \frac{n}{2}||p-u||, \tag{13}$$

are both positive. We see that as $||p - u|| \to 0$, $W(X)$ approaches its maximum

$$W(X) \to \frac{n+1}{2},$$

within increasingly tight bounds.

Remark 3 *Notice again how $1 - \|p - u\|$ appears in the upper bound of equation (13), which can be rewritten*

$$W(X) \leq \frac{1}{2} + \frac{n}{2}(1 - \|p - u\|).$$

2.3 The Problem with Entropy

Recall that the entropy of an $\mathscr{X}$-valued random variable X is given by

$$H(X) = -\sum_{i=1}^{|\mathscr{X}|} p_i \log_2 p_i,$$

where $p_i = \mathsf{P}[X = x_i]$.

Proposition 7 gave upper and lower bounds for the work-factor of the job of guessing with even odds, a value drawn once from a random variable. It turns out that entropy is generally a *bad* indicator of this work-factor. This could be a surprising fact to some since our language and intuition, as well as many fundamental results about entropy seem to suggest a strong work-factor interpretation. The folklore about entropy as a measure of work is directly asserted in the Internet document "Cryptography Frequently Asked Questions (FAQ)" [Bac94]. Section §4.9 reads in its entirety:

> 4.9. What's a key-guessing attack? What's entropy?
>
> Say somebody is using the one-time pad—but isn't choosing keys randomly and uniformly from all m-bit messages, as he was supposed to for our security proof. In fact say he's known to prefer keys which are English words. Then a cryptanalyst can run through all English words as possible keys. This attack will often succeed, and it's much faster than a brute-force search of the entire keyspace.
>
> 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.

On the contrary, as we shall show, entropy is *not* a reliable measure of how bad a distribution is under precisely the paradigm described, and there may be nearly *no* chance of encountering the value of X within the best $2^{H(X)}$ guesses. What then might have justified such folklore?

The asymptotic equipartition property (AEP) tells us that for sufficiently large n, a sequence of n values drawn from a random variable X has approximately

$$(2^{H(X)})^n,$$

outcomes of non-negligible probability, each being nearly equally likely.[3] Thus in trying to find such a sequence in a brute-force guessing attack, we would justifiably expect the size, in bits, of the effective search space to be about $nH(X)$. Asymptotically, each extra random value in the sequence adds only about $H(X)$ bits to the search space. It is tempting to hope that for *short sequences*, the AEP offers a decent *approximation* to the search space size. In

[3] Formally, the AEP is analogous to the law of large numbers and can be stated as follows. Let $X_1, \ldots, X_n$ be independent and identically distributed $\mathscr{X}$-valued random variables drawn according to probability distribution $p : \mathscr{X} \longrightarrow \mathbb{R}$. Define the self-referential real valued random variables $p(X_i)$ by $\mathsf{P}[p(X_i) = p(x_j)] = \mathsf{P}[X_i = x_j] = p(x_j)$, and similarly $p(X_1, X_2 \ldots, X_n) = p(X_1)p(X_2) \cdots p(X_n)$. Then

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \to H(p) \text{ in probability,}$$

as $n \to \infty$ (see [Ash65] and cf. [CT91]). Thus for every $\varepsilon > 0$, there is an n such that

$$\mathsf{P}\left[\left|-\frac{1}{n} \log p(X_1, \ldots, X_n) - H(p)\right| < \varepsilon\right] \geq 1 - \varepsilon.$$

In essence, drawing a value of $p(X_1, \ldots, X_n)$ means drawing the probability of a sequence $(X_1, \ldots, X_n)$. For fixed n and ε, it is natural to divide the set of all $|\mathscr{X}|^n$ sequences into the *typical sequences* T_n^ε, those for which

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

and the remaining *atypical sequences*. The following immediate consequences of the AEP support its less formal interpretation.

$$\mathsf{P}[(X_1, \ldots, X_n) \in T_n^\varepsilon] \geq 1 - \varepsilon,$$

and

$$(1 - \varepsilon) 2^{n(H(p)-\varepsilon)} \leq |T_n^\varepsilon| \leq 2^{n(H(p)+\varepsilon)}.$$

other words, it is tempting to assume that the entropy of a random variable is in the same ball-park as the logarithm of its work-factor. However, that assumption is dangerously invalid; propositions 9 and 10 below show that $H(X)$ can be orders of magnitude away from $\log \mathrm{wf}_{\frac{1}{2}}(X)$, in *either* direction.

The paradigm of having to guess a *single value* (e.g. an encryption key) is strikingly different from that of having to guess a *long sequence*. The difference is highlighted by the following observation.

Proposition 8 *For any $\varepsilon > 0$, there exists an integer n and nonincreasing probabilities $p_1 \geq p_2 \geq \ldots \geq p_n$, whose entropy H satisfies*

$$\sum_{i=1}^{2^{\lceil H \rceil}} p_i < \varepsilon.$$

Proof: Let $a = 2^j$ and define a family of random variables $X(j,k)$ by the sequence of probabilities

$$\frac{1}{a}, \frac{1}{a^2}, \frac{1}{a^3} \ldots, \frac{1}{a^k},$$

followed by m copies of the last value

$$\frac{1}{a^k}, \ldots, \frac{1}{a^k}, \; m \text{ times},$$

where m must be chosen so that the probabilities add to 1. These probabilities come from self-similar Huffman trees discussed in figure 2 below. It is easy to show that

$$m = \frac{1 + (a-2)a^k}{a-1}$$

satisfies this condition. Now let us examine the entropies of this family.

$$\begin{aligned} H_{j,k} &= \sum_{i=1}^{k} \frac{1}{a^i} \log a^i + \left[\frac{1 + (a-2)a^k}{a-1} \right] \frac{1}{a^k} \log a^k \\ &= j \sum_{i=1}^{k} \frac{i}{a^i} + jk \left[\frac{1 + (a-2)a^k}{(a-1)a^k} \right] \\ &= j \left[\frac{a^{k+1} - (k+1)a + k}{(a-1)^2 a^k} \right] + jk \left[\frac{1 + (a-2)a^k}{(a-1)a^k} \right] \\ &= j \left[\frac{ka^{k+2} + (1-3k)a^{k+1} + 2ka^k - a}{a^k (a-1)^2} \right] \\ &= jk \left(\frac{a-2}{a-1} \right) + h_{j,k}, \end{aligned}$$

where

$$h_{j,k} = \frac{j(a^k - 1)}{a^{k-1}(a-1)^2}.$$

Let us fix $j > 2$ and thus $a = 2^j > 3$. Then $h_{j,k} > 0$ and

$$2^{\lceil H_{j,k} \rceil} \geq 2^{jk} = a^k > k.$$

so that

$$\begin{aligned} s_{j,k} &\triangleq \sum_{i=1}^{2^{\lceil H_{j,k} \rceil}} p_i \\ &= \sum_{i=1}^{k} \frac{1}{a^i} + (2^{\lceil H_{j,k} \rceil} - k) \frac{1}{a^k} \\ &= \frac{a^k - 1}{a^k (a-1)} + (2^{\lceil H_{j,k} \rceil} - k) \frac{1}{a^k} \end{aligned}$$

$$= \quad \frac{1}{a-1} + \sigma_{j,k},$$

where

$$\sigma_{j,k} = \frac{(a-1)(2^{\lceil H_{j,k} \rceil} - k) - 1}{a^k(a-1)}.$$

We claim that for fixed j, $\sigma_{j,k} \to 0$ as $k \to \infty$. If that were true, then for any $\varepsilon > 0$, we may fix a $j > 2$ such that

$$\frac{1}{a-1} < \varepsilon,$$

and find a k such that

$$s_{j,k} < \varepsilon.$$

Thus the proof would be complete.

We now turn our attention to finding the limit of $\sigma_{j,k}$ as $k \to \infty$. For some $\widehat{k}(j)$,

$$h_{j,k} < 1, \text{ for all } k \geq \widehat{k}(j),$$

because $h_{j,k} \to j/a$ as $k \to \infty$. Thus after $k = \widehat{k}(j)$,

$$\lceil H_{j,k} \rceil \leq jk\left(\frac{a-2}{a-1}\right) + 2.$$

An upper bound for $\sigma_{j,k}$ may be given as

$$\sigma_{j,k} \leq \frac{(a-1)(4\alpha^k - k) - 1}{a^k(a-1)},$$

where

$$\alpha = 2^{j\left(\frac{a-2}{a-1}\right)} < a.$$

By two applications of L'Hospital's rule, $\sigma_{j,k} \to 0$ as $k \to \infty$. □

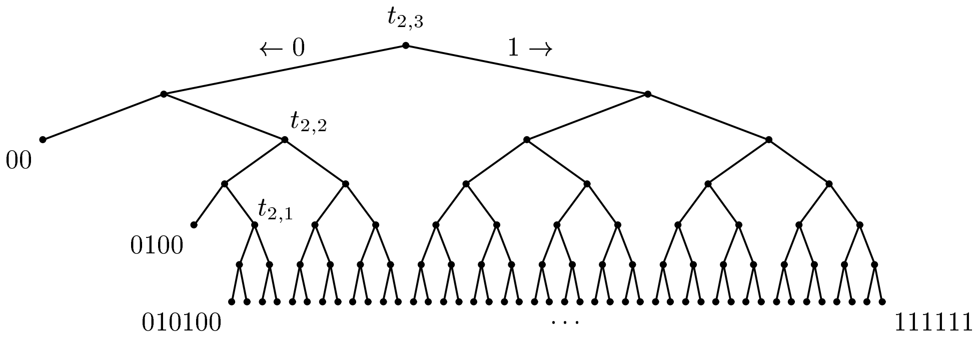

Figure 2: Each random variable $X(j,k)$ in the proof of proposition 8 corresponds to a self-similar Huffman tree $t_{j,k}$ ($t_{2,3}$ is shown above). The sample space can be thought of as the set of leaves of the tree, and to each leaf there is a codeword of 0's and 1's defining the path from the root to that leaf. The probability of each codeword w is $2^{-|w|}$. The trees are self-similar in that each $t_{j,k}$ contains subtrees isomorphic to $t_{j,i}$, for all $i < k$. The self-similarity is also characterized by a zig-zag pattern of the path starting from the root and going alternately through the maximal proper codeword prefixes and the roots of the trees isomorphic to $t_{j,i}, i < k$.

The next proposition formalizes the notion that there are random variables for which $\log \mathrm{wf}_{\frac{1}{2}}(X) \gg H(X)$, i.e. that the effective search space size measured in bits can be arbitrarily more than the entropy.

Proposition 9 *For each $N > 0$, there is a random variable satisfying*

$$\log \mathrm{wf}_{\frac{1}{2}}(X) - H(X) > N.$$

Proof: Again let $H = H(X)$ and the probabilities p_i of X be in nonincreasing order. Choose the ε of proposition 8 satisfying

$$\varepsilon = \frac{1}{2^{N+1}}.$$

Then

$$\sum_{i=1}^{2^N 2^{\lceil H \rceil}} p_i \leq 2^N \sum_{i=1}^{2^{\lceil H \rceil}} p_i < 2^N \varepsilon = \frac{1}{2}.$$

We conclude that

$$\mathrm{wf}_{\frac{1}{2}}(X) > 2^N 2^{\lceil H \rceil} \geq 2^{N+H},$$

or

$$\log \mathrm{wf}_{\frac{1}{2}}(X) - H > N.$$

□

There are also random variables for which $\log \mathrm{wf}_{\frac{1}{2}}(X) \ll H(X)$.

Proposition 10 *For each $N > 0$, there is a random variable X satisfying*

$$H(X) - \log \mathrm{wf}_{\frac{1}{2}}(X) > N.$$

Proof: Consider the sequence of probabilities given by $p_1 = \frac{1}{2}$ followed by

$$p_i = \frac{1}{2^{k+1}}, \quad 2 \leq i \leq 2^k + 1,$$

in other words the same smallest probability is repeated 2^k times. This corresponds to a Huffman tree as in figure 3. It is easy to see that

$$\mathrm{wf}_{\frac{1}{2}}(p) = 1,$$

and

$$H(p) = \frac{k+2}{2}.$$

The desired result is obtained when $H(p) > N$, which is easily achieved if we choose

$$k > 2(N-1).$$

□

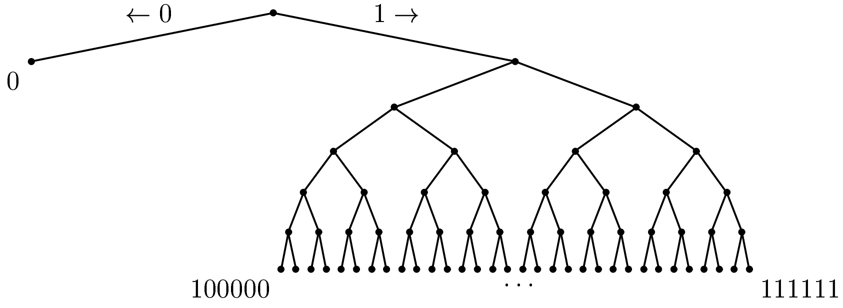

Figure 3: An unbalanced Huffman tree demonstrating $H(X) \gg \log \mathrm{wf}_{\frac{1}{2}}(X)$, as in proposition 10.

In conclusion, entropy is not in general a good indicator of work. As cryptologists, we should avoid using the term “entropy” to connote “work,” or as an all-encompassing synonym for “uncertainty” or “randomness.” In fact, any function of the probabilities of a random variable which satisfies propositions 9 and 10 should be disqualified from the everyday usage which “entropy” enjoys today.[4] We have remarked that $1 - \|p - u\|$ appears in both the upper bound for $\mathrm{wf}_{\frac{1}{2}}(X)$ (in proposition 7) and in the increasingly tight upper bound for $W(X)$ (in equation (13)).

[4] Undoubtedly, this usage began with the thermodynamic meaning of entropy and was not introduced into the language by cryptologists or information theorists.

Evidently,

$$1 - \|p - u\|,$$

is a better measure of the uncertainty of guessing than

$$H(p) = \log|\mathscr{X}| - D(p\|u).$$

References

[Akl65] Hayward R. Akler. *Mathematics and Politics.* Macmillan, New York, 1965.

[Ash65] Robert B. Ash. *Information Theory.* Dover, New York, 1965.

[Bac94] Bach and *et al.* Cryptography Frequently Asked Questions (FAQ), 1994. URL: `news:sci.crypt`.

[Bir46] G. Birkhoff. Tres observaciones sobre el algebra lineal. *University Nac. Tucuman Rev. Ser. A*, 5:147–150, 1946.

[BR97] R. B. Bapat and T.E.S. Raghavan. *Nonnegative Matrices and Applications*, volume 64 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 1997.

[Cac97] Christian Cachin. *Entropy Measures and Unconditional Security in Cryptography.* PhD thesis, ETH Zürich, 1997.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[Dia88] Persi Diaconis. *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics, Hayward, CA, 1988.

[HLP29] Godfrey H. Hardy, John E. Littlewood, and George Pólya. Some simple inequalities satisfied by convex functions. *Messenger Math.*, 58:145–152, 1929.

[HLP52] Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities.* Cambridge University Press, New York and London, second edition, 1952.

[Lor05] M. O. Lorenz. Methods of measuring concentration of wealth. *J. Amer. Statist. Assoc.*, 9:209–219, 1905.

[Mas94] James L. Massey. Guessing and entropy. *Proc. 1994 IEEE Int'l Symp. on Information Theory*, page 204, 1994.

[MO79] Albert W. Marshall and Ingram Olkin. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, San Diego, 1979.

[Muf] Alec Muffett. *Crack Version 5.0a User Manual.* URL: `ftp://ftp.cert.org/pub/tools/crack/`.

[Pea92] David W. Pearce. *The MIT Dictionary of Modern Economics.* The MIT Press, Cambridge, MA, fourth edition, 1992.

[Pli99] John O. Pliam. *Ciphers and their Products: Group Theory in Private Key Cryptography.* PhD thesis, University of Minnesota, 1999. in preparation, URL: `http://www.ima.umn.edu/~pliam/doc`.

[von53] John von Neumann. A certain zero-sum two-person game equivalent to an optimal assignment problem. *Ann. Math. Studies*, 28:5–12, 1953.