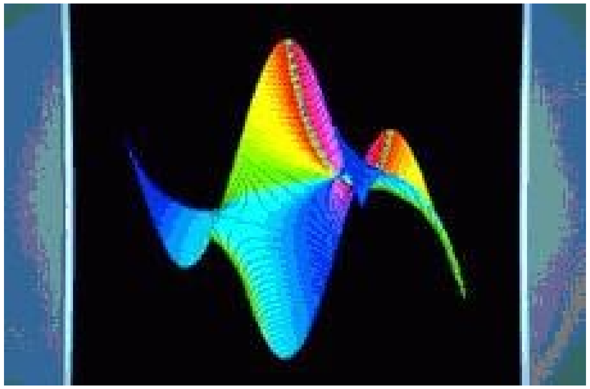
A One-to-One Code and Its Anti-Redundancy*

W. Szpankowski
Department of Computer Science,
Purdue University

July 4, 2005



*This research is supported by NSF, NSA and NIH.

Outline of the Talk

1. Prefix Codes
2. Redundancy
3. Our One-to-One Code
4. Asymptotic Results for Anti-redundancy
5. Sketch of Proof
   - Sums involving the floor function
   - Saddle point method
   - Sums and sequences modulo 1

Some Definitions

A block code

$$C_n : \mathcal{A}^n \to \{0,1\}^*$$

is an injective mapping from the set $\mathcal{A}^n$ of all sequences $x_1^n = x_1 \ldots x_n$ of length n over the alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of binary sequences.

For a **given source P**, the pointwise redundancy and the average redundancy are defined as respectively

$$\begin{aligned} R_n(C_n, P; x_1^n) &= L(C_n) + \lg P(x_1^n) \\ \bar{R}_n(C_n, P) &= \mathbf{E}_{X_1^n}[R_n(C_n, P; X_1^n)] \\ &= \mathbf{E}[L(C_n, X_1^n)] - H_n(P) \end{aligned}$$

where $L(C_n, x_1^n)$ is the code length,
$H_n(P) = -\sum_{x_1^n} P(x_1^n)$ the source entropy,
and $\mathbf{E}$ denotes the expectation,

Prefix Codes

Usually, we deal with prefix codes which are defined as those in which there is no codeword being a prefix of another codeword.
Prefix codes do satisfy Kraft's inequality:
$\sum_{x_1^n} 2^{-L(x_1^n)} \leq 1.$

Shannon Lower Bound:
For any prefix code

$$\mathbf{E}[L(C_n, X_1^n)] \geq H_n(P).$$

Indeed, let $K = \sum_{x_1^n} 2^{-L(x_1^n)} \overset{Kraft}{\leq} 1.$

$$\begin{aligned}
\mathbf{E}[L(C_n, X_1^n)] &- H_n(P) = \\
&= \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) L(x_1^n) + \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log P(x_1^n) \\
&= \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log \frac{P(x_1^n)}{2^{-L(x_1^n)}/K} - \log K \\
&\geq 0
\end{aligned}$$

since the first term is a divergence and cannot be negative (or $\log x \leq x - 1$ for $0 < x \leq 1$).

Redundancy for Prefix Codes

Throughout this talk we assume that the source P is given and is binary memoryless with probability p for transmitting a 0. That is, $P(x_1^n) = p^k(1-p)^{n-k}$ where k is the number of 0's.

Let

$$\alpha = \log_2\left(\frac{1-p}{p}\right), \quad \beta = \log_2\left(\frac{1}{1-p}\right).$$

and $\langle x \rangle = x - \lfloor x \rfloor$ be the fractional part of x.

Redundancy of the Shannon-Fano Code:

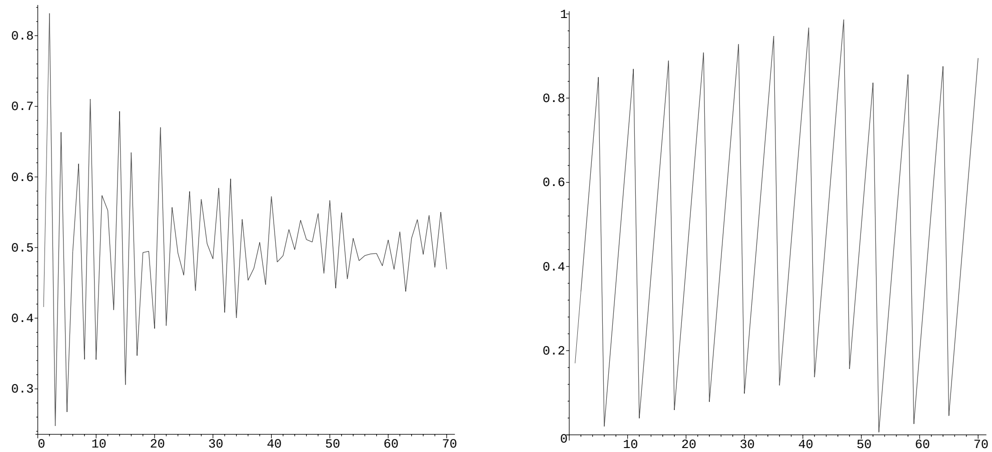
$$\bar{R}_n^{SF} = \begin{cases} \frac{1}{2} + o(1) & \alpha \quad \text{irrational} \\ \\ \frac{1}{2} - \frac{1}{M}\left(\langle Mn\beta \rangle - \frac{1}{2}\right) + O(\rho^n) & \alpha = \frac{N}{M}, \quad \gcd(N,M) = 1 \end{cases}$$

Redundancy of the Huffman Code:

$$\bar{R}_n^{H} = \begin{cases} \frac{3}{2} - \frac{1}{\log 2} + o(1) \approx 0.057304 & \alpha \quad \text{irrational} \\ \\ \frac{3}{2} - \frac{1}{M}\left(\langle \beta M n \rangle - \frac{1}{2}\right) - \frac{1}{M(1-2^{-1/M})} 2^{-\langle n\beta M \rangle / M} + O(\rho^n) & \alpha = \frac{N}{M} \end{cases}$$

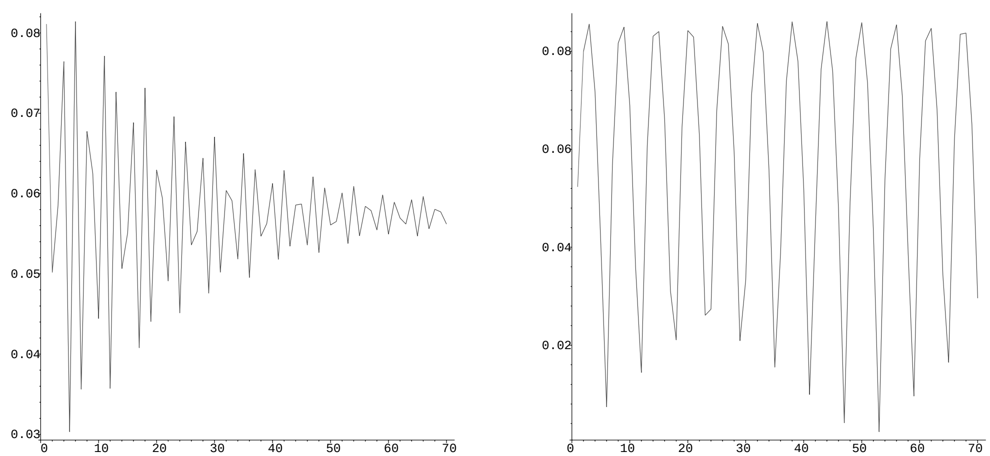
where N, M are integers such that $\gcd(N,M) = 1$ and $\rho < 1$.

Oscillations



(a) (b)

Figure 1: Shannon–Fano code redundancy



(a) (b)

Figure 2: Huffman's code redundancy versus block size n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

One-to-One Codes

One-to-One codes are not prefix codes.

In one-to-one codes a distinct codeword is assigned to each source symbol and unique decodability is not required. Such codes are usually one shot codes and there is one designated an "end of message" channel symbol.

Wyner in 1972 proved that

$$L \leq H(X),$$

which was further improved by Alon and Orlitsky who showed

$$L \geq H(X) - \log(H(X) + 1) - \log e.$$

Can we establish more precise bounds?
Where are the oscillations observed in prefix codes?

Block One-to-One Codes

We consider a block one-to-one code for $x_1^n = x_1 \ldots x_n \in \mathcal{A}^n$ generated by a memoryless source with p being the probability of generating a 0 and $q = 1 - p$.

We write $P(x_1^n) = p^k q^{n-k}$, where k is the number of 0s. Throughout we assume $p \leq q$.

We now list all 2^n probabilities in a nonincreasing order and assign code lengths as follows

$$
\begin{array}{ccccccc}
q^n \left(\frac{p}{q}\right)^0 & \geq & q^n \left(\frac{p}{q}\right)^1 & \geq & \ldots & \geq & q^n \left(\frac{p}{q}\right)^n \\
\lfloor \log_2(1) \rfloor & & \lfloor \log_2(2) \rfloor & & \ldots & & \lfloor \log_2(2^n) \rfloor
\end{array}
$$

Average Code Length

There are $\binom{n}{k}$ equal probabilities $p^k q^{n-k}$.
Define

$$A_k = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}, \quad A_{-1} = 0.$$

Starting from the position A_{k-1} the next $\binom{n}{k}$ probabilities $P(x_1^n)$ are the same.
The average code length is

$$\begin{aligned} L_n &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{j=A_{k-1}+1}^{A_k} \lfloor \log_2(j) \rfloor \\ &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor. \end{aligned}$$

Our goal is to estimate L_n asymptotically for large n.

An Ugly Sum

To evaluate the inner part of the sum for L_n we apply the following identity (cf. Knuth Ex. 1.2.4-42)

$$\sum_{j=1}^{N} a_j = N a_n - \sum_{j=1}^{N-1} (a_{j+1} - a_j)$$

for any sequence a_j. Then

$$\begin{aligned}
L_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor \\
&- \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} 2^{-\langle \log_2 A_k \rangle} \\
&+ \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{1 + A_{k-1}}{\binom{n}{k}} \left(\log_2 \left(1 + \binom{n}{k} A_{k-1}^{-1} \right) \right. \\
&\quad \left. + \langle \log_2 A_{k-1} \rangle - \langle \log_2 A_k \rangle \right) \\
&- 2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{A_{k-1}}{\binom{n}{k}} \left(2^{-\langle \log_2 A_k \rangle} - 4 \cdot 2^{-\langle \log_2 A_{k-1} \rangle} \right)
\end{aligned}$$

where $\langle x \rangle = x - \lfloor x \rangle$ is the fractional part of x.

Main Result

Theorem 1. *Consider a binary memoryless source and the one-to-one block code described above. Then for* $p < \frac{1}{2}$

$$\begin{aligned} L_n &= nH(p) - \frac{1}{2}\log_2 n - 1 - \frac{1}{2\ln 2} + \log_2 \frac{1-p}{(1-2p)\sqrt{pq\pi}} \\ &+ \frac{1-p}{1-2p}\log_2\frac{2-3p}{1-p} + \frac{5-4p}{1-2p}\left(\frac{1}{2\ln 2} + G(n)\right) \\ &+ F(n) + o(1) \end{aligned}$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$, *and* $G(n) = F(n) = 0$ *if* $\log_2 \frac{1-p}{p}$ *is irrational. If* $\log_2\frac{1-p}{p} = N/M$ *for some integers* M, N *such that* $\gcd(N, M) = 1$, *then* $G(n)$ *and* $F(n)$ *are oscillating functions of complicated nature. For example,* $F(n)$ *is equal to*

$$\frac{1}{M\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\left(\left\langle M\left(n\beta - \log\left(\frac{1-2p}{1-p}\sqrt{2\pi pqn}\right) - \frac{x^2}{2\ln 2}\right)\right\rangle - \frac{1}{2}\right)dx$$

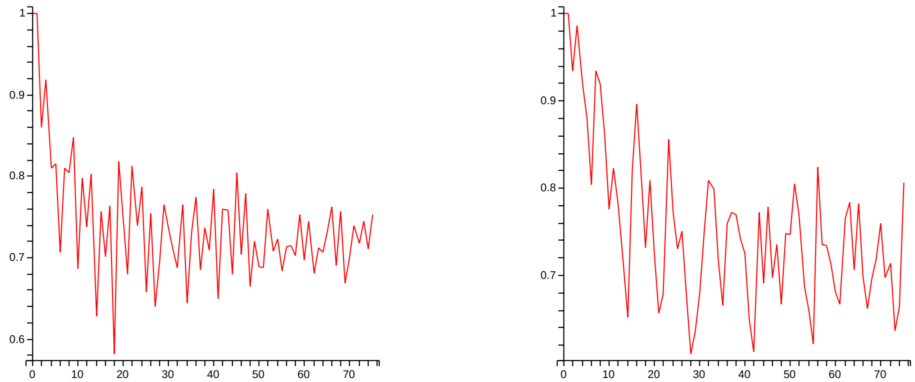
where $\beta = -\log_2(1-p)$.

For $p = \frac{1}{2}$, *then*

$$L_n = nH(1/2) - 1 + 2^{-n}(n-2)$$

for every $n \geq 1$.

Oscillations



(a) (b)

Figure 3: The "constant" part of the average anti-redundancy versus n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

Anti-redundancy $R_n = L_n - nH(p)$ for our one-to-one code is

$$\bar{R}_n = -\frac{1}{2}\log n + O(1)$$

where the $O(1)$ terms contains oscillations.

Sketch of Proof

1. We only deal with the sum

$$\begin{aligned} S_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k \\ &\quad - \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle \\ &= a_n + b_n \end{aligned}$$

where

$$\begin{aligned} a_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k, \\ b_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle. \end{aligned}$$

Asymptotics of A_n

2. We need the saddle point approximation of A_n.

Lemma 1. *For large n and $p < 1/2$*

$$A_{np} = \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}2^{nH(p)}\left(1+O(n^{-1/2})\right).$$

More precisely, for an $\varepsilon > 0$ and $k = np + \Theta(n^{1/2+\varepsilon})$ we have

$$\begin{aligned} A_k &= \frac{1-p}{1-2p}\frac{1}{\sqrt{2\pi np(1-p)}}\left(\frac{1-p}{p}\right)^k \frac{1}{(1-p)^n} \\ &\times \exp\left(-\frac{(k-np)^2}{2p(1-p)n}\right)\left(1+O(n^{-\delta})\right) \end{aligned}$$

for some $\delta > 0$.

Proof. Notice that

$$A_n(z) = \sum_{k=0}^{n} A_k z^k = \frac{(1+z)^n - 2^n z^{n+1}}{1-z}.$$

and apply the saddle point method to the Cauchy formula.

Binomial Distribution Approximation

3. Using Stirling's approximation we find a good approximation for the binomial distribution.

Lemma 2. *Let* $p_n(k) = \binom{n}{k} p^k q^{n-k}$ *where* $q = 1 - p$ *be the binomial distribution. Then for* $|k - pn| \leq n^{1/2+\varepsilon}$ *we have*

$$p_n(k) = \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{(k-pn)^2}{2p(1-p)n}\right) + O(n^{-\delta})$$

uniformly as $n \to \infty$. *Furthermore*

$$\sum_{|k-np| > \sqrt{np} n^{1/2+\varepsilon}} p_n(k) < 2n^{-\varepsilon} e^{-n^{2\varepsilon}/2}$$

for large n.

4. From above lemmas we find

$$\log A_k = \log A_{np} + \alpha(k - np) - \frac{(k - np)^2}{2pqn \ln 2} + O(n^{-\delta}).$$

and then

$$a_n = \log A_{np} - \frac{1}{2 \ln 2} + O(n^{-\delta})$$

which is the desired result.

Returning to b_n

5. Recall we need asymptotics of

$$b_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle .$$

From previous lemmas we conclude that

$$\log A_k = \alpha k + n\beta - \log_2 \omega\sqrt{n} - \frac{(k-np)^2}{2pqn \ln 2} + O(n^{-\delta})$$

for some $\omega > 0$.
Thus we need asymptotics of the following sum

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left\langle \alpha k + n\beta - \log_2 \omega\sqrt{n} - \frac{(k-np)^2}{2pqn \ln 2} \right\rangle .$$

Final Lemma

6. To complete we need the following lemma.

Lemma 3. *Let $0 < p < 1$ be a fixed real number and $f : [0,1] \to \mathbf{R}$ be a Riemann integrable function.*

(i) *If α is irrational, then*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right)$$

$$= \int_0^1 f(t)\,dt,$$

where the convergence is uniform for all shifts $y \in \mathbf{R}$.

Continue ...

(ii) *Suppose that* $\alpha = \frac{N}{M}$ *is a rational number with integers* N, M *such that* $\gcd(N, M) = 1$. *Then uniformly for all* $y \in \mathbf{R}$

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right)$$
$$= \int_0^1 f(t)\,dt + H_M(y)$$

where

$$H_M(y) \quad := \quad \frac{1}{M}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\left(\left\langle M\left(y - \frac{x^2}{2\ln 2}\right)\right\rangle - \int_0^1 f(t)\,dt\right)\,dx$$

is a periodic function with period $\frac{1}{M}$.