"All-One" Blocks in a Binary String

Rana Barua
Stat-Math Division
Indian Statistical Institute
203, B.T. Road
Calcutta 700035
India
e-mail: rana@isical.ac.in

Palash Sarkar*
Centre for Applied Cryptographic Research
Department of Combinatorics and Optimization
University of Waterloo
200 University Avenue West
Waterloo, Ontario
Canada N2L 3G1
e-mail: psarkar@cacr.math.uwaterloo.ca

Abstract

We study the distribution of "all-one" blocks in a binary string. More specifically, we compute the expected number of all-one blocks and the expected length of an all-one block in a randomly chosen binary string. This is done under two probability distributions - the uniform distribution and the Kolmogorov complexity based universal distribution.

Keywords : combinatorial problems, binary strings, all-one block, uniform distribution, Kolmogorov complexity, universal distribution.

1 Introduction

Binary strings have an ubiquitous presence throughout the whole of computer science. Here we study patterns in a binary string. The simplest possible pattern is a contiguous sequence of 0's or a contiguous sequence of 1's. It is clear that results about a contiguous sequence of 1's will also hold for contiguous sequence of 0's. Thus it is enough to concentrate on sequences with a contiguous sequence of 1's. We call a contiguous sequence of 1's an "all-one" block and study the distribution of this pattern in a binary string of length n. (As an example the string 011110011000 has two all-one blocks of lengths 4 and 2). Here we would like to point out that similar studies have been made on the weight of different representations of numbers (see [1, 6]).

We study two parameters about the all-one blocks - the expected number of all-one blocks and the expected length of an all-one block in a randomly chosen binary string. We consider two possible

*On leave from the Indian Statistical Institute, Calcutta, INDIA. e-mail: palash@isical.ac.in

distributions over the binary strings - the uniform and the universal distributions. It is common to assume the uniform distribution in the lack of any other information about the scenario. However, it has been argued [4] that the Kolmogorov complexity based universal distribution is a more natural choice in the absence of any other information.

The Kolmogorov complexity of a string measures its inherent complexity and is the length of the shortest program whose output is the string when run on an fixed universal computer. For reasons described in [5] the Kolmogorov complexity is not directly suitable for defining a probability distribution on the set of all binary strings. Instead a different measure called prefix complexity is used to define a probability distribution on the set of all binary strings. The prefix complexity is the shortest length of a description whose output is the string when we restrict ourselves to the set of prefix free description, i.e., no description is a proper prefix of any other description. Denote by $K(x)$ the prefix complexity of a binary string x. Then the universal distribution $\mathbf{m}(x)$ assigns probability $2^{-K(x)}$ to the string x. A surprising application of the universal distribution shows that under the universal distribution the average case complexity of programs is same as the worst case complexity of programs [4]. This is one of our motivations for considering the behaviour under the universal distribution.

2 Behaviour Under Uniform Distribution

Let $N(n, b)$ be the number of n-bit strings having exactly b many "all-one" blocks.

Proposition 2.1 *The following holds.*

1. $N(n, 0) = 1$ *for all* $n \geq 0$.
2. $N(n, 1) = \frac{n(n+1)}{2}$ *for all* $n \geq 1$.
3. $N(n, i) = 0$ *for* $i > \lceil \frac{n}{2} \rceil$.
4. *If* n *is odd, then* $N(n, \lceil \frac{n}{2} \rceil) = 1$.

Proof : (1) is obvious.
(2) Since there is only one all-one block, its length can be between 1 and n. If the length is i, then there are $\binom{n-i}{1}$ binary strings having exactly one all-one block of length i. Hence

$$N(n, 1) = n + (n-1) + \ldots + 1 = \frac{n(n+1)}{2}.$$

(3) The length of each block must be at least 1 and there must be at least one 0 between two block of length 1. Hence there cannot be binary string of length n having more than $\lceil \frac{n}{2} \rceil$ all-one blocks.
(4) For odd n, the only string having $\lceil \frac{n}{2} \rceil$ blocks is of the form $101010\ldots01$. ■

We now obtain a recurrence for $N(n, b)$.

Theorem 2.1 *For $b > 1$,*

$$N(n,b) = 2N(n-1,b) - N(n-2,b) + N(n-2,b-1). \tag{1}$$

Proof : Since $b > 1$ there must be at least one zero in the string. Any binary string of length n having at least one zero can be written in the form $1^i 0 w$, where w is a string of length $n-1-i$. Note that the number of all-one blocks must be b. Thus if $i = 0$, then w must have b all-one blocks. The number of such w is $N(n-1,b)$. If $i > 0$, then w must have $(b-1)$ all-one blocks. the length of w being $(n-i-1)$, the number of such blocks is $N(n-1-i, b-1)$. Thus we have

$$N(n,b) = N(n-1,b) + N(n-2,b-1) + N(n-3,b-1) + \ldots + N(1,b-1).$$

Changing n to $(n-1)$ we get

$$N(n-1,b) = N(n-2,b) + N(n-3,b-1) + N(n-4,b-1) + \ldots + N(1,b-1).$$

Thus we get

$$N(n,b) = 2N(n-1,b) - N(n-2,b) + N(n-2,b-1). \qquad \blacksquare$$

The generating function for $N(n,b)$ involves two variables and is defined below.

$$G(x,y) = \sum_{n \geq 0} \sum_{b=0}^{b=\lceil \frac{n}{2} \rceil} N(n,b) x^n y^b. \tag{2}$$

A tedious but straightforward computation provides an expression for the generating function.

Proposition 2.2

$$G(x,y) = \frac{1}{1-x} + \frac{xy}{(1-x)(1-2x+x^2-x^2y)}. \tag{3}$$

An expression for $N(n,b)$ is obtained from $G(x,y)$ as follows.

Theorem 2.2 *For all $n \geq 0$ and $0 \leq b \leq \lceil \frac{n}{2} \rceil$, $N(n,b) = \binom{n+1}{2b}$.*

Proof : We write $G(x,y)$ as

$$G(x,y) = \frac{1}{1-x} + \frac{\sqrt{y}}{2(1-x)} \left(\frac{1}{1-x-x\sqrt{y}} - \frac{1}{1-x+x\sqrt{y}} \right).$$

The term $\frac{1}{(1-x)}$ expands to $1 + x + x^2 + x^3 + \cdots$. Hence the coefficient of $x^n y^0$ in $G(x,y)$ is 1 which agrees with $N(n,0) = 1$ for all $n \geq 0$.

Let $A = \frac{1}{1-x-x\sqrt{y}}$ and $B = \frac{1}{1-x+x\sqrt{y}}$. We have

$$\begin{aligned} A &= 1 + x(1+\sqrt{y}) + x^2(1+\sqrt{y})^2 + x^3(1+\sqrt{y})^3 + \cdots. \\ B &= 1 + x(1-\sqrt{y}) + x^2(1-\sqrt{y})^2 + x^3(1-\sqrt{y})^3 + \cdots. \end{aligned}$$

Let $c(n,b)$ denote the coefficient of $x^n y^b$ in $\frac{\sqrt{y}(A-B)}{2}$. It is not difficult to check that $c(n,b) = \binom{n}{2b-1}$. For $b > 0$ the coefficient of $x^n y^b$ in $G(x,y)$ is the coefficient of $x^n y^b$ in $\frac{\sqrt{y}(A-B)}{2(1-x)}$. Thus we have

$$\begin{aligned} N(n,b) &= c(n,2b-1) + c(n-1,2b-1) + c(n-2,2b-1) + \cdots + c(0,2b-1) \\ &= \binom{n}{2b-1} + \binom{n-1}{2b-1} + \binom{n-2}{2b-1} + \cdots + \binom{0}{2b-1} \\ &= \binom{n+1}{2b} \end{aligned}$$

■

We now turn to computing the expected number of all-one blocks in a binary string. We do this directly from the recurrence for $N(n,b)$.

Let $p(n,b)$ be the probability that an n-bit binary string (chosen uniformly at random from the set of all n-bit strings) has exactly b all-one blocks. Then $p(n,b) = \frac{N(n,b)}{2^n}$. Let $E(n)$ be the expected number of all-one blocks in a binary string of length n. Then we have

$$E(n) = \sum_{b=0}^{b=\lceil \frac{n}{2} \rceil} bp(n,b) = \sum_{b=0}^{b=\lceil \frac{n}{2} \rceil} \frac{bN(n,b)}{2^n}. \tag{4}$$

We now compute $E(n)$.

Theorem 2.3 *The expected number $E(n)$ of all-one blocks in a binary string of length n is $\frac{n+3}{4}$.*

Proof : For $b \geq 2$, from Theorem 2.1, we have

$$N(n,b) = 2N(n-1,b) - N(n-2,b) + N(n-2,b-1).$$

Multiplying both sides by b and summing from $b=2$ to $b = \lceil \frac{n}{2} \rceil$, we get,

$$\begin{aligned} \sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} bN(n,b) &= 2\sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} bN(n-1,b) - \sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} bN(n-2,b) \\ &+ \sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} (b-1)N(n-2,b-1) + \sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} N(n-2,b-1). \end{aligned}$$

Using $\sum_{b=0}^{b=\lceil \frac{n}{2} \rceil} bN(n,b) = 2^n E(n)$ and $\sum_{b=2}^{b=\lceil \frac{n}{2} \rceil} N(n-2,b-1) = 2^{n-2} - 1$, we get,

$$\begin{aligned} 2^n E(n) - N(n,1) &= 2\left(2^{n-1}E(n-1) - N(n-1,1)\right) - \left(2^{n-2}E(n-2) - N(n-2,1)\right) \\ &+ 2^{n-2}E(n-2) + 2^{n-2} - 1. \end{aligned}$$

We simplify this using $N(n,1) = \frac{n(n+1)}{2}$, to get

$$2^n E(n) = 2^n E(n-1) + 2^{n-2}.$$

This gives $E(n) - E(n-1) = \frac{1}{4}$, and so $E(n) = \frac{n-2}{4} + E(2)$. Since $E(2) = \frac{5}{4}$, we get $E(n) = \frac{n+3}{4}$. ■

It is known that the expected number of ones in a binary string of length n is $\frac{n}{2}$. Thus the expected number of all-one blocks is roughly half this value. Accordingly we would expect the expected length of an all-one block is small. We proceed to find this quantity in the next section.

2.1 Expected Length of All-One Block

The computation of the expected length of an all-one block is different from the computation of the expected number of all-one blocks. In this case we have to consider the ensemble Ω_n of all strings of length n. Let $S(n)$ be the total number of all-one blocks present among all strings of Ω_n. Also let $M(n,i)$ be the total number of all-one blocks of length i present among all the strings in Ω_n. Then the probability $q(n,i)$ such that a randomly chosen string from Ω_n will have an all-one block of length i is $q(n,i) = \frac{M(n,i)}{S(n)}$. So the expected length of an all-one block in a randomly chosen string from Ω_n is

$$E_1(n) = \sum_{i=0}^{i=n} iq(n,i) = \sum_{i=0}^{i=n} \frac{iM(n,i)}{S(n)}. \tag{5}$$

Theorem 2.4 *The expected length $E_1(n)$ of an all-one block in a binary string of length n is $\frac{2n}{n+1}$.*

Proof : First note that the quantity $\sum_{i=0}^{i=n} iM(n,i)$ is equal to the number of ones in all the strings in Ω_n. There are 2^n strings in Ω_n and each string is n bits long. Thus the total number of bits in Ω_n is $n2^n$. Clearly the total number of ones is exactly half this value. Hence we have $\sum_{i=0}^{i=n} iM(n,i) = n2^{n-1}$. We now compute $S(n)$.

Let $x(n,i)$ be the total number of all-one blocks that are present in n-bit strings of the form 1^i0w, for $0 \leq i \leq n-1$. Then $x(n,0) = S(n-1)$ and for $i > 0$, $x(n,i) = S(n-1-i) + 2^{n-1-i}$. We give an explanation for the last expression. The number of strings of the form 1^i0w is equal to 2^{n-1-i}. Each of these strings certainly have the all-one block 1^i. The other all-one blocks are from all the possible w's. Since length of w is $n-1-i$, this is equal to $S(n-1-i)$. Thus we get

$$\begin{aligned} S(n) &= x(n,0) + \ldots + x(n,n-1) + 1 \\ &= S(n-1) + \left(S(n-2) + 2^{n-2}\right) + \ldots + \left(S(0) + 2^0\right) + 1 \\ &= S(n-1) + S(n-2) + \ldots + S(0) + 2^{n-1}. \end{aligned}$$

Changing n to $n-1$ we get,

$$S(n-1) = S(n-2) + S(n-3) + \ldots + S(0) + 2^{n-2}.$$

Combining the two expressions we get,

$$S(n) = 2S(n-1) + 2^{n-2}.$$

This gives $S(n) = (n+1)2^{n-2}$. Using $\sum_{i=0}^{i=n} iM(n,i) = n2^{n-1}$ and (5) we get $E_1(n) = \frac{2n}{n+1}$. ■

3 Behaviour Under the Universal Distribution

We now consider the behaviour of the number of all-one blocks and the expected length of an all-one block under the universal distribution. The universal distribution has been used in [4] to show that the average case complexity is same as the worst case complexity. This is in sharp contrast to the behaviour under the uniform distribution. Our motivation for studying the behaviour under the universal distribution partially originates from the above mentioned result. Intuitively strings that have longer all-one blocks should be more compressible and should have a higher probability of occurence under the universal distribution. Thus the expected length of an all-one block under the universal distribution should be higher. But our analysis shows that this is not the case. The behaviour under the uniform and the universal distribution is asymptotically the same. Thus our result provides an example where both the distributions agree.

We follow [4, 5] in our notation and idea of the proof. Let x be a binary string and $K(x)$ be the prefix complexity of x. The universal distribution $\mathbf{m}$ is defined by $\mathbf{m}(x) = 2^{-K(x)}$. It is known [5] that $\mathbf{m}(x)$ is enumerable and multiplicatively dominates all enumerable distributions. Let $a(x)$ be the number of all-one blocks in a binary string x. Also by $l(x)$ we denote the length of a binary string x. We define an enumerable probability distribution $P(x)$ over the set of all binary strings as follows.

1. For each $n = 0, 1, \ldots$, set $a_n = \sum\limits_{l(x)=n} \mathbf{m}(x)$.
2. If $l(x) = n$ and x is lexicographically least (under some fixed enumeration of all binary strings) with $a(x) = \lceil \frac{n}{2} \rceil$, then set $P(x) = a_n$ else $P(x) = 0$.

If n is odd, then there is only one string with $a(x) = \lceil \frac{n}{2} \rceil$. If n is even, then there are more than one string with $a(x) = \lceil \frac{n}{2} \rceil$. The lexicographically least string gets a nonzero probability which is equal to the sum of the probabilities that $\mathbf{m}(x)$ assigns to all binary strings of length n. Since $\mathbf{m}(x)$ is enumerable, a_n is also enumerable and hence $P(x)$ is enumerable. Using [5, Theorem 4.1, Example 4.5], we have $c_P \mathbf{m}(x) \geq P(x)$, where $c_P = K(P) + O(1)$ is a constant depending on $P()$ but not on x. Let $E(n|\mathbf{m})$ denote the expected number of all-one blocks in a string of length n under the universal distribution. Thus we get

$$\begin{aligned} E(n|\mathbf{m}) &= \frac{\sum_{l(x)=n} \mathbf{m}(x) a(x)}{\sum_{l(x)=n} \mathbf{m}(x)} \\ &\geq \frac{\frac{1}{c_P} \sum_{l(x)=n} P(x) a(x)}{\sum_{l(x)=n} \mathbf{m}(x)} \\ &= \frac{1}{c_P} \frac{\lceil \frac{n}{2} \rceil a_n}{\sum_{l(x)=n} \mathbf{m}(x)} \\ &= \frac{1}{c_P} \left\lceil \frac{n}{2} \right\rceil . \end{aligned}$$

Moreover, since the maximum number of all-one blocks in a string of length n is $\lceil \frac{n}{2} \rceil$ we certainly have

$$E(n|\mathbf{m}) \quad \leq \quad \left\lceil \frac{n}{2} \right\rceil$$

and so

$$\frac{1}{c_P} \left\lceil \frac{n}{2} \right\rceil \leq E(n|\mathbf{m}) \leq \left\lceil \frac{n}{2} \right\rceil .$$

Thus we get

Theorem 3.1 *The expected number of all-one blocks $E(n|\mathbf{m})$ under the universal distribution is $\theta(\lceil \frac{n}{2} \rceil)$.*

This "agrees" with the expected number of all-one blocks under the uniform distribution.

Next we proceed to estimate the expected length of an all-one block under the universal distribution. Let $E_1(n|\mathbf{m})$ be the expected length of an all-one block under the universal distribution $\mathbf{m}$. Recall that $a(x)$ is the number of all-one blocks in the string x and $M(n,b)$ is the total number of all-one blocks present among all strings in Ω_n. Let

$$B(x,b) = \text{number of times a block of length } b \text{ occur in the string } x.$$

The probability $p(n,b|\mathbf{m})$ that an n-bit string chosen randomly according to $\mathbf{m}$ will have an all-one block of length b is given by

$$p(n,b|\mathbf{m}) = \frac{\sum_{x:l(x)=n} m(x)B(x,b)}{\sum_{x:l(x)=n} m(x)a(x)}.$$

It is not difficult to check that $\sum_{b=0}^{b=n} p(n,b|\mathbf{m}) = 1$. Thus $p(n,b|\mathbf{m})$ is a proper probability distribution. The expected length $E_1(n|\mathbf{m})$ is given by the following.

$$E_1(n|\mathbf{m}) = \frac{\sum_{b=0}^{b=n} b \sum_{x:l(x)=n} m(x)B(x,b)}{\sum_{x:l(x)=n} m(x)a(x)}. \tag{6}$$

Let $K_{\max}(n) = \max\{K(x) : l(x) = n\}$ and $K_{\min}(n) = \min\{K(x) : l(x) = n\}$. Using $\mathbf{m}(x) = 2^{-K(x)}$, we can write

$$E_1(n|\mathbf{m}) \leq 2^{K_{\max}(n) - K_{\min}(n)} \frac{\sum_{b=0}^{b=n} b \sum_{x:l(x)=n} B(x,b)}{\sum_{x:l(x)=n} a(x)}. \tag{7}$$

We clearly have $M(n,b) = \sum_{x:l(x)=n} B(x,b)$ and $S(n) = \sum_{x:l(x)=n} a(x)$. Using $\sum_{b=0}^{b=n} bM(n,b) = n2^{n-1}$ and $S(n) = (n+1)2^{n-2}$, we get

$$E_1(n|\mathbf{m}) \leq 2^{K_{\max}(n) - K_{\min}(n)} \frac{2n}{n+1}. \tag{8}$$

It is known that $K_{\max}(n) = n + K(n) + O(1)$ (see [2]). However, the value of $K_{\min}(n)$ is not known. Thus it is difficult to upper bound Equation (8) for all n-bit strings. Fortunately, we can upper bound Equation (8) for most n bit strings by using a result of Chaitin [3], where it is shown that the number of strings that have the maximum prefix complexity is at least 2^{n-c} for some constant c. Let $\hat{E}_1(n|\mathbf{m})$ be an approximation of $E_1(n|\mathbf{m})$ by restricting attention to the set of all n-bit strings having the maximum prefix complexity. Then starting from Equation (6) we get

$$\hat{E}_1(n|\mathbf{m}) = 2^{O(1)} \frac{2n}{n+1}. \tag{9}$$

This gives us the following result.

Theorem 3.2 *The estimate $\hat{E}_1(n|\mathbf{m})$ of the expected length of an all-one block $E_1(n|\mathbf{m})$ is equal to $\theta(2^{O(1)}\frac{2n}{n+1})$.*

Since only a negligible fraction of all n-bit strings have been ignored, the estimate above seems to be a reasonable one. However, it would be nice to prove the upper bound holds for the set of all n-bit strings, or compute a bound on the error in the above approximation.

References

[1] S. Arno and F. S. .Wheeler. Signed digit representation of minimum Hamming weight. *IEEE Transactions on Computers*, volume 42, number 8, pages 1007-1010, 1993.

[2] G. J. Chaitin. *Algorithmic Information Theory* Cambridge University Press, 1987.

[3] G. J. Chaitin. A Note on the number of N-bit strings with maximum complexity. *Applied Mathematics and Computation*, volume 59, pages 97-100, 1993.

[4] M. Li and P. Vitanyi. Average case complexity equals worst-case complexity under the universal distribution. *Information Processing Letters*, 42(1992), 145-149.

[5] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer-Verlag, New York, 1993.

[6] H. Wu and M. A. .Hasan. Closed-form expressions for the average weight of signed-digit representations. *IEEE Transactions on Computers*, volume 48, number 9, pages 848-851, 1999.