Uniform Density Binary Words

Lionel Levine

September, 2000

Given a finite binary word $a = a_1 \dots a_n$ and an irrational number $\alpha \in [0,1]$, we say that a has uniform density α if

$$\lfloor (s-r)\alpha \rfloor \leq \sum_{i=r+1}^{s} a_i \leq \lceil (s-r)\alpha \rceil$$

for all $0 \leq r < s \leq n$; in other words, the proportion of 1's in every subword of a is close to α. Let D_α denote the set of all such words.

We will see that D_α is closely related to the infinite binary sequence whose n-th term is

$$s_n(\alpha, t) = \lfloor t + n\alpha \rfloor - \lfloor t + (n-1)\alpha \rfloor.$$

The basic relationship between D_α and $s(\alpha, t)$ is the following.

Theorem 1. D_α is the set of all finite subwords of $s(\alpha, t)$.

Proof. If a is a subword of $s(\alpha, t)$ – say it begins at the k-th place – then for any $0 \leq r < s \leq n$ we have

$$\sum_{i=r+1}^{s} a_i = \lfloor t + (k+s)\alpha \rfloor - \lfloor t + (k+r)\alpha \rfloor,$$

which is the number of integers in the half-open interval $(t + (k+r)\alpha, t + (k+s)\alpha]$. This interval has length $(s-r)\alpha$, so it contains either $\lfloor (s-r)\alpha \rfloor$ or $\lceil (s-r)\alpha \rceil$ integers. Hence $a \in D_\alpha$.

For the converse, suppose $a \in D_\alpha$. Then we need to show that a is a subword of $s(\alpha, t)$. Note that truncating the first k characters from $s(\alpha, t)$ yields $s(\alpha, t + k\alpha)$. Since the fractional parts of the multiples of α are dense in $[0, 1]$, it's enough to find an interval $I \subset [0, 1]$ such that for any $u \in I$, the sequence $s(\alpha, u)$ begins with a. Let

$$I = (\max_{r=1}^{n} m_r, 1 + \min_{r=1}^{n} m_r),$$

where $m_r = \sum_{i=1}^{r} a_i - r\alpha$. We need to check that I is in fact an interval, i.e. that its right endpoint exceeds its left endpoint. Since $a \in D_\alpha$, for any $1 \le r < s \le n$ we have

$$m_s - m_r = \sum_{i=r+1}^{s} a_i - (s - r)\alpha < 1,$$

so the min and max differ by less than 1 as desired.

Finally, let's check that $s(\alpha, u)$ begins with a for all $u \in I$. If $u \in I$, then $m_r \le u < 1 + m_r$ for every $r = 1, \ldots, n$. Hence

$$\sum_{i=1}^{r} a_i \le u + r\alpha < 1 + \sum_{i=1}^{r} a_i;$$

in other words, there is an integer between $u + (r-1)\alpha$ and $u + r\alpha$ if and only if $a_r = 1$. Thus $s_r(\alpha, u) = a_r$, $r = 1, \ldots, n$. This completes the proof.

Corollary 1. If $a \in D_\alpha$, it occurs not just once, but infinitely many times in $s(\alpha, t)$.

Proof. Wherever a occurs in $s(\alpha, t)$ – say it ends at the k-th place – we can substitute $t + k\alpha$ for t in the theorem to get a later occurrence of a.

Corollary 2. If $a_1 \ldots a_n \in D_\alpha$, then $a_1 \ldots a_n 0 \in D_\alpha$ or $a_1 \ldots a_n 1 \in D_\alpha$.

Proof. If $a_1 \ldots a_n \in D_\alpha$, then it occurs somewhere in $s(\alpha, t)$, where it is followed either by a 0 or a 1.

While Theorem 1 gives an equivalent description of D_α, it doesn't provide any means of constructing words of uniform density. For a more direct approach, we can arrange the elements of D_α in a tree (Figure 1).

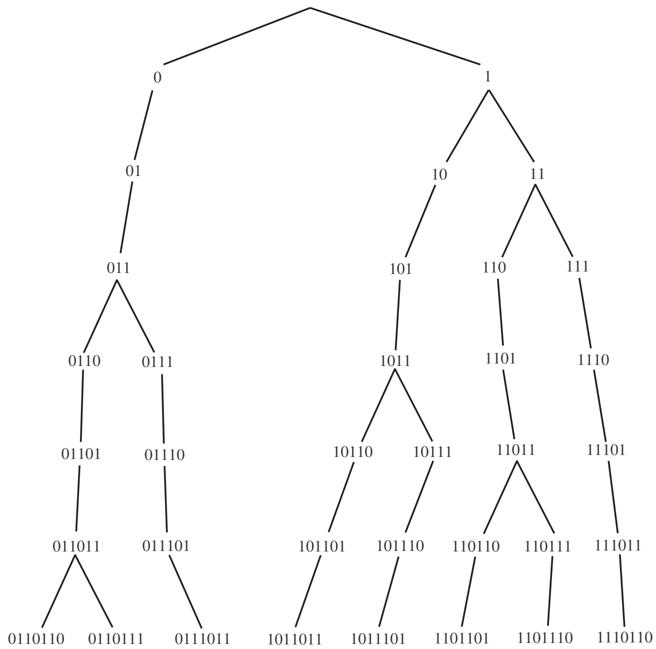

Figure 1: Words of uniform density $\sqrt{2}/2$ of length ≤ 7

We draw a branch from word a down to word b if b can be obtained by appending either 0 or 1 to a. By Corollary 2 above, this tree has no terminating branches. Thus each word $a_1 \ldots a_n$ in the tree is the top node of either one or two branches accordingly as one or both of $a_1 \ldots a_n 0$ and $a_1 \ldots a_n 1$ are in D_α. If a is the top node of two branches, call it a "splitter." From Figure 1 we see that when $\alpha = \sqrt{2}/2$, the first few splitters are

$$1, 11, 011, 1011, 11011, 011011, \ldots$$

If we understand where the splitters occur in the tree, then we understand the whole tree. First of all, notice that there is at most one splitter of each length. This is certainly true for length zero; suppose it fails for the first time at length n. Then there is a splitter of length $n-1$ (just truncate the first character from any splitter of length n) and it's unique. Call it $a_1 \ldots a_{n-1}$. Then the only possibilities for splitters of length n are $0a_1 \ldots a_{n-1}$ and $1a_1 \ldots a_{n-1}$. But these can't both be splitters, else $0a_1 \ldots a_{n-1}0$ and $1a_1 \ldots a_{n-1}1$ would both be in D_α, which is impossible.

We'll show that in fact there's *exactly* one splitter of each length. Since $a_1 \dots a_n$ is a splitter whenever $a_0 a_1 \dots a_n$ is, this means there is an infinite binary sequence $f_1, f_2, \dots$ such that for each n, the unique splitter of length n has the form $f_n f_{n-1} \dots f_1$. So the question is, what's this mystery sequence f? As it turns out, the answer is sitting right in front of us. In this context, what's the most natural infinite binary sequence we can think of? Well, let's try $s(\alpha, t)$. But for what value of t? What's the most natural real number we can think of? Well, how about α? That's right, this mystery sequence f is none other than $s(\alpha, \alpha)$! In less dramatic terms, that's $s(\alpha, 0)$ with the initial 0 truncated. Thus

Theorem 2. For each positive integer n, there is exactly one splitter of length n, namely $s_{n+1}(\alpha, 0)s_n(\alpha, 0) \dots s_2(\alpha, 0)$.

Proof. Since we've already shown that there is at most one splitter of length n, we need only show that $s_{n+1}(\alpha, 0)s_n(\alpha, 0) \dots s_2(\alpha, 0)$ is in fact a splitter. To do this, choose $\epsilon > 0$ small enough so that $s(\alpha, 0)$ and $s(\alpha, 1-\epsilon)$ coincide to the first $n + 1$ places, excluding the first place. Note that for sufficiently small ϵ, they do *not* coincide at the first place: $s_1(\alpha, 0) = 0$, while $s_1(\alpha, 1-\epsilon) = 1$. By Theorem 1, this means that both $0s_2(\alpha, 0) \dots s_n(\alpha, 0)s_{n+1}(\alpha, 0)$ and $1s_2(\alpha, 0) \dots s_n(\alpha, 0)s_{n+1}(\alpha, 0)$ are in D_α. Since D_α is preserved under reversal, $s_{n+1}(\alpha, 0)s_n(\alpha, 0) \dots s_2(\alpha, 0)$ is a splitter.

As trivial as it seems, truncating the initial 0 from $s(\alpha, 0)$ is in some strange way the key to the proof.

Since there's exactly one splitter of each length, the number of words of length n in D_α is always one more than the number of words of length $n-1$, which gives us the following rather counterintuitive result.

Corollary. The number of words of length n in D_α is $n+1$, independent of α.

What is the significance of these results? In some sense, the sequence $s(\alpha, 0)$ (or, if you like, $s(\alpha, \alpha)$) tells us everything we could ever want to know about D_α. Not only does it describe D_α as a set (Theorem 1), it also encodes the complete *structure* of D_α as a tree (Theorem 2).

We conclude with some questions that deserve further thought. Given real numbers $\alpha_1, \ldots, \alpha_n > 0$ with $\sum_{i=1}^n \alpha_i = 1$, we can say that an n-ary word $a_1 \ldots a_m$ has uniform density $(\alpha_1, \ldots, \alpha_n)$ if

$$\lfloor (s-r)\alpha_k \rfloor \leq |\{r < i \leq s : a_i = k\}| \leq \lceil (s-r)\alpha_k \rceil$$

for all $0 \leq r < s \leq m$ and all $k = 1, \ldots, n$. Can we classify such sequences? In particular, what is the n-ary counterpart to the sequence $s(\alpha, t)$?

To make this question a little more precise, we can say that an infinite binary sequence has uniform density α if all of its finite subwords do. Let U_α be the space of all such sequences, with the topology induced by the lexicographic ordering. Call two sequences in U_α equivalent if they agree in all but finitely many places, and let V_α be the corresponding quotient space. Then by an argument similar to that used in the proof of Theorem 1, it can be shown that up to equivalence, the infinite binary sequences of uniform density α are precisely the sequences $s(\alpha, t)$, $0 \leq t < 1$. Thus the space V_α is homeomorphic to the circle S^1. What is the analogous parameterization for n-ary sequences?