Dense and universal sets of words

Narad Rampersad

June 8, 2014

Let $w = w_1 w_2 \cdots w_n$ be a word of length n. For a subset $J = \{j_1, j_2, \ldots, j_k\} \subseteq \{1, 2, \ldots, n\}$, let w_J denote the subsequence $w_{j_1} w_{j_2} \cdots w_{j_k}$. Similarly, if A is a set of words of length n, let A_J denote the set $\{w_J : w \in A\}$. A set of words $A \subseteq \{\mathtt{0}, \mathtt{1}\}^n$ is (n, k)-*dense* if there exists a k-element subset $J \subseteq \{1, 2, \ldots, n\}$ such that $A_J = \{\mathtt{0}, \mathtt{1}\}^k$. The set A is (n, k)-*universal* if for every k-element subset J we have $A_J = \{\mathtt{0}, \mathtt{1}\}^k$.

Example 1. *Let A be the set of words*

$$\begin{matrix} 0 & 0 & 0 & 0, \\ 0 & 1 & 0 & 1, \\ 1 & 0 & 1 & 0, \\ 1 & 1 & 1 & 1. \end{matrix}$$

Then A is $(4, 2)$-dense, since if $J := \{1, 2\}$, we have $A_J = \{\mathtt{00}, \mathtt{01}, \mathtt{10}, \mathtt{11}\}$. However, A is not $(4, 2)$-universal, since if $J := \{1, 3\}$, we have $A_J = \{\mathtt{00}, \mathtt{11}\} \neq \{\mathtt{0}, \mathtt{1}\}^2$.

Let B be the set of words

$$\begin{matrix} 0 & 0 & 0 & 0, \\ 0 & 0 & 1 & 1, \\ 0 & 1 & 0 & 1, \\ 1 & 0 & 1 & 0, \\ 1 & 1 & 0 & 0, \\ 1 & 1 & 1 & 1. \end{matrix}$$

Then B is $(4, 2)$-universal, as one may verify that for any $i < j$, we have $B_{\{i,j\}} = \{\mathtt{0}, \mathtt{1}\}^2$.

Clearly for a set A to be (n, k)-dense it must contain a word that contains at least k $\mathtt{1}$'s. The set

$$X_{n,k} := \{w \in \{\mathtt{0}, \mathtt{1}\}^n : w \text{ does not contain } k\ \mathtt{1}\text{'s}\}$$

is thus not (n, k)-dense. Define

$$t(n, k) := |X_{n,k}| = \sum_{i=0}^{k-1} \binom{n}{i}.$$

Theorem 2. *Let $A \subseteq \{\mathtt{0}, \mathtt{1}\}^n$. If $|A| > t(n, k)$ then A is (n, k)-dense.*

Proof. The proof is by induction on n and k. If $k = 1$ then for all n we have $t(n,k) = 1$ and every set A of more than 1 element is clearly $(n,1)$-dense. Suppose then that $k > 1$. Define

$$B := \{x \in \{\mathtt{0},\mathtt{1}\}^{n-1} : x \text{ is a prefix of } w \text{ for some } w \in A\}$$

and

$$C := \{x \in \{\mathtt{0},\mathtt{1}\}^{n-1} : x\mathtt{0} \in A \text{ and } x\mathtt{1} \in A\}.$$

Then $|A| = |B| + |C|$. To see this, let $w \in A$ and write $w = xa$ for some letter $a \in \{\mathtt{0},\mathtt{1}\}$. If $x\overline{a} \notin A$, then the x in B accounts for w in A. If $x\overline{a} \in A$, then the x in B accounts for one of $xa, x\overline{a}$ in A and the x in C accounts for the other. Thus, $|A| = |B|+|C|$. If $|B| > t(n-1,k)$, then by induction B is $(n-1,k)$-dense, which implies that A is (n,k)-dense. Suppose then that $|B| \le t(n-1,k)$. We then have

$$\begin{aligned}
|C| &= |A| - |B| \\
&> t(n,k) - t(n-1,k) \\
&= \sum_{i=0}^{k-1} \binom{n}{i} - \sum_{i=0}^{k-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{k-1} \left(\binom{n}{i} - \binom{n-1}{i} \right) \\
&= \sum_{i=1}^{k-1} \binom{n-1}{i-1} \\
&= \sum_{i=0}^{k-2} \binom{n-1}{i} \\
&= t(n-1,k-1),
\end{aligned}$$

where we have used the "Pascal's triangle" identity $\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}$. Since $|C| > t(n-1,k-1)$, we can apply the induction hypothesis to conclude that C is $(n-1,k-1)$-dense. However, $C\{\mathtt{0},\mathtt{1}\} \subseteq A$, and so A is (n,k)-dense. ☐

Example 3. *For $n = 4$ and $k = 2$, we have*

$$\begin{aligned}
t(4,2) &= \sum_{i=0}^{1} \binom{4}{i} \\
&= \binom{4}{0} + \binom{4}{1} \\
&= 1 + 4 \\
&= 5.
\end{aligned}$$

Thus every set A of at least 6 binary words of length 4 is $(4,2)$-dense.

Next we show the existence of small (n, k)-universal sets. We begin with the special case $k = 2$.

Theorem 4. *For $n \geq 2$ there is an $(n, 2)$-universal set of size $2\lceil \log_2 n \rceil + 2$.*

Proof. We define such a set A as follows. Let $t := \lceil \log_2 n \rceil$. Define M to be the $t \times n$ matrix whose columns consist of the binary representations of the integers $0, 1, \ldots, n-1$. Let B denote the set of words of length n obtained by taking each row of M as a word. We define

$$A := \{\mathtt{0}^n, \mathtt{1}^n\} \cup B \cup B',$$

where B' is the set of words formed by taking the binary complements of the words in B.

Consider any pair of indices $J := \{i < j\}$. Since $\mathtt{0}^n \in A$, we have $\mathtt{00} \in A_J$, and similarly $\mathtt{11} \in A_J$. Furthermore, since the columns of M were distinct, there must be a word $w \in B$ whose i-th and j-th symbols are different (say $w_i = \mathtt{0}$ and $w_j = \mathtt{1}$), so that $\mathtt{01} \in A_J$. Since A also contains the complements of the words in B, we also have $\mathtt{10} \in A_J$. Thus A_J consists of all binary words of length 2, as required. □

Example 5. *For $n = 4$ and $k = 2$, set $t := 2$ and define the 2×4 matrix*

$$M := \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Then $B = \{\mathtt{0011}, \mathtt{0101}\}$, $B' = \{\mathtt{1100}, \mathtt{1010}\}$, and $A = \{\mathtt{0000}, \mathtt{1111}\} \cup B \cup B'$. We have already seen in Example 1 that A is $(4, 2)$-universal.

For larger values of k, we use the probabilistic method.

Theorem 6. *For $n \geq 2$ and $k \geq 3$ there is an (n, k)-universal set of size $k2^k\lceil \log n + 1 \rceil$.*

Proof. Let t be an integer satisfying

$$\binom{n}{k} 2^k (1 - 2^{-k})^t < 1 \tag{1}$$

and let A be a set of t random words of length n. Each word in A is chosen by selecting each letter from $\{\mathtt{0}, \mathtt{1}\}$ independently at random with probability $1/2$. For every set J of k indices and every word $x \in \{\mathtt{0}, \mathtt{1}\}^k$, the probability that $x \notin A_J$ is $(1 - 2^{-k})^t$. Since there are $\binom{n}{k}$ choices for J and 2^k choices for x, the probability that $A_J \neq \{\mathtt{0}, \mathtt{1}\}^k$ is

$$\binom{n}{k} 2^k (1 - 2^{-k})^t,$$

which is less than 1 by our choice of t. Thus with positive probability A is (n, k)-universal. Using the inequalities $\binom{n}{k} < (ne/k)^k$ and $(1 - 2^{-k})^t \leq e^{-t/2^k}$, one verifies that when $k \geq 3$, (1) holds for $t = k2^k\lceil \log n + 1 \rceil$. □

Theorem 2 is due (independently) to Perles and Shelah (see Shelah [1972]), Sauer [1972], and Vapnik and Chervonenkis [1971]. Theorem 4 is due to Chandra, Kou, Markowsky, and Zaks [1983]. Theorem 6 is due to Kleitman and Spencer [1973].