Shift synchronization codes

R. Ahlswede[*], B. Balkenhol[†] and T. Partner[‡]

Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld
Germany

Keywords: synchronization errors, deBruijn network

AMS subject classification: 94B50, 94A45

[*]email: ahlswede@mathematik.uni-bielefeld.de
[†]email: bernhard@mathematik.uni-bielefeld.de
[‡]email: tpartner@mathematik.uni-bielefeld.de

1 Introduction

In this paper some aspects of binary codes capable of correcting or detecting shift synchronization errors are considered. Questions of synchronization always have been popular in Coding Theory (see e.g. [1], [2], [4], [5]).

First we introduce three models for decoding synchronization errors.

In all these models codewords $c \in \mathcal{X}^n$ of length n are considered. The decoder receives a word of n consecutive symbols from a transmission of codewords, which may have been shifted. This word is assumed to contain at least $n-t$ consecutive symbols of the codeword to be decoded.

1. The receiver (decoder) has to recover the transmitted codeword.

2. The decoder has access to the whole (infinite) sequence of transmitted codewords. In this case starting at the received position he can shift to the left and to the right in order to look at the bits around the actual n bits.

3. The decoder has the information of 2, but he has not only to return the transmitted codeword, he has also to correct a shift error.

The codes for model 3 are called here t error correcting shift synchronization codes. Their maximal cardinality is denoted by $S(t,n)$ and is a quality of prime interest.

Now, the first situation is like in the usual theory of error correcting codes but instead of the Hamming distance we have to define another one. The second one is different from the first one, because if the receiver looks at a sequence $x_2,\dots,x_n,y$ and he knows that this is a 1-shift-error (to the right), then in the first situation there is only one codeword $x_1,\dots,x_n$ whereas by moving our sliding window of length n one step left in situation 2, the receiver is getting the correct bit, i.e. it is possible that both sequences $(0,x_2,\dots,x_n$ and $1,x_2,\dots,x_n)$ are codewords. The third one is again different because the all–zero-sequence for example has the property, that it is neighbor of itself. Therefore if there is another codeword starting with 0 or ending with 0, then there is a run of zeros of length greater or equal to $n+1$ which means the decoder returns the correct codeword, but he is not necessarily able to decide which one is the correct position.

2 Basic definitions

For a finite set $\mathcal{X}=\{0,\dots,a-1\}$, called alphabet, we form $\mathcal{X}^n=\prod_1^n \mathcal{X}$, the words of length n, with letters from $\mathcal{X}$ and $\mathcal{X}^*=\bigcup_{n=0}^{\infty}\mathcal{X}^n$, the set of all finite length words including the empty word e from $\mathcal{X}^0=\{e\}$, $\mathcal{X}^*$ is equipped with an associative operation, called concatenation, defined by

$$(x_1, \ldots, x_n)(y_1, \ldots, y_m) = (x_1, \ldots, x_n, y_1, \ldots, y_m).$$

We skip the brackets whenever this results in no confusion, in particular we write the letter x instead of (x). We also write $\mathcal{X}^+ = \mathcal{X}^* \setminus \{e\}$ for the set of non–empty words.

The **length** $\mid x^n \mid$ of the word $x^n = x_1 \ldots x_n$ is the number n of letters in x^n.

A word $w \in \mathcal{X}^*$ is a **factor** of a word $x \in \mathcal{X}^*$ if there exist $u, v \in \mathcal{X}^*$ such that $x = uwv$. A factor w of x is **proper** if $w \neq x$.

For subsets $\mathcal{Y}, \mathcal{Z}$ of $\mathcal{X}^*$ and a word $w \in \mathcal{X}^*$, we define

$$\mathcal{Y}w = \{yw \in \mathcal{X}^* : y \in \mathcal{Y}\},$$

$$\mathcal{Y}\mathcal{Z} = \{yz \in \mathcal{X}^* : y \in \mathcal{Y}, z \in \mathcal{Z}\}$$

and

$$\mathcal{Y}w^{-1} = \{z \in \mathcal{X}^* : zw \in \mathcal{Y}\}.$$

A set of words $\mathcal{C} \subset \mathcal{X}^*$ is called a **code**.

Recall that a code is called prefix–free (resp. suffix–free), if no codeword is beginning (resp. ending) of another one.

Definition 1. *A code, which is simultaneously prefix–free and suffix–free, is called biprefix or fix–free. This can be expressed by the equations*

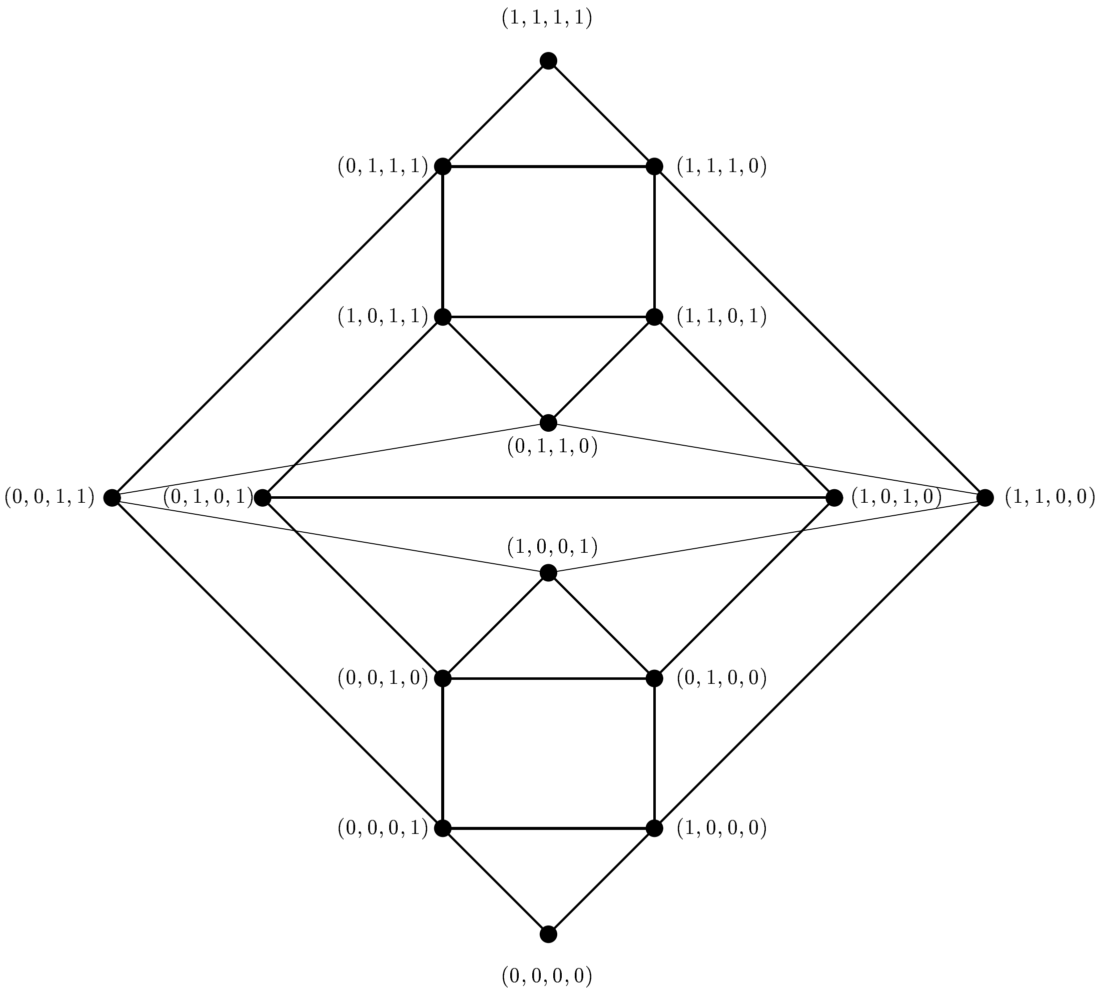
$$\mathcal{C}\mathcal{X}^+ \cap \mathcal{C} = \phi \text{ and } \mathcal{X}^+\mathcal{C} \cap \mathcal{C} = \phi.$$

3 Relations to the deBruijn Network

The **binary deBruijn Network** of order n is an undirected graph $\mathcal{B}^n = (\mathcal{V}^n, \mathcal{E}^n)$, where $\mathcal{V}^n = \mathcal{X}^n$ is the set of vertices and $(u^n, v^n) \in \mathcal{E}^n$ is an edge iff

$$u^n \in \{(b, v_1, \ldots, v_{n-1}), (v_2, \ldots, v_n, b) : b \in \{0, 1\}\}.$$

The binary deBruijn Network $\mathcal{B}^4$ is given as an example:



A subset $\mathcal{A} \subset \mathcal{V}^n$ is called independent, if no two vertices of $\mathcal{A}$ are connected, and we denote by $\mathcal{I}(\mathcal{B}^n)$ the set if all independent subsets of the deBruijn network. We note, that for all $b \in \{0,1\}$, $(b,b,\dots b) \notin \mathcal{A} \in \mathcal{I}(\mathcal{B}^n)$, because $(b,b,\dots b)$ is dependent on itself. The independence number $f_{\mathcal{B}}(n)$ of $\mathcal{B}^n$ is $f_{\mathcal{B}}(n) = \max\limits_{\mathcal{A} \in \mathcal{I}(\mathcal{B}^n)} \mid \mathcal{A} \mid$.

Independent sets are of special interest also in Coding Theory. In [1] it is shown, that for a fix – free code , where all codewords have a length out of $\{k, k+1, k+2\}$, the set of codewords of length k is a member of $\mathcal{I}(\mathcal{B}^n)$.

4 Distance properties

Now for model 1 (as described in the introduction) our distance d between two sequences x and y is the minimal distance between the two sequences in the deBruijn

Network. For the other models we have to define a direction on each edge of the graph as follows.

Definition 2. *A* **right-shift** f *on* $\{0,1\}^n$ *is defind by* $f(x_1,\dots,x_n)=\{x_2,\dots,x_n,b : b\in\{0,1\}\}$*. The* **right-distance** d_r *between two sequences* x *and* y *is defined as the minimal number of right-shifts, which are necessary to shift from* x *to* y*, i.e. it is the minimal number* m *such that* $z_1,\dots z_m\in\{0,1\}^n$ *with* $z_1=x$*,* $z_m=y$ *and* $z_{i+1}\in f(z_i)$ *for all* $1\le i<m$*. We define the* **distance** $d(x,y)$ *as*

$$d(x,y)=\min\{d_r(x,y),d_r(y,x)\}.$$

For a code with codewords having pairwise a distance greater or equal to t we like to solve the problem of constructing a maximal set of codewords for model 3, which is asymptotically optimal.

Let $S(t,n)\subset\{0,1\}^n$ be a maximal set such that for all $x,y\in S(t,n)$ $d(x,y)\ge t$.

Conjecture 1.

$$\lim_{n\to\infty}\frac{|S(t,n)|}{2^n}=\frac{1}{t}.$$

5 Construction for $t=2^k$ and $n=t\cdot r$

For the special case where $t=2^k$ and $n=t\cdot r$ for a $k,r\in\mathbb{N}$ we prove Conjecture 1.

Let

$$S(t,n)\triangleq\left\{x_1,\dots,x_{4r}:\sum_{i=1}^{\frac{n}{2^j}}x_{2^ji}>\sum_{i=1}^{\frac{n}{2^j}}x_{2^ji-2^{j-1}},\ \text{for all } 1\le j\le k\right\}.$$

With $n=2^k\cdot r$ all the terms $\frac{n}{2^j}$ are integers.

Before we are able to calculate the size of the sets $S(t,n)$ we need some further definitions and notations.

Let $\mu_m,\dots,\mu_1$ with $\mu_i\in\{<,>,=\}$ be a vector of relationships and

$$S_{\mu_m,\dots,\mu_1}=\left\{x_1,\dots,x_n\in\{0,1\}^n:\sum_{i=1}^{\frac{n}{2^j}}x_{2^ji}\quad\mu_j\quad\sum_{i=1}^{\frac{n}{2^j}}x_{2^ji-2^{j-1}},\ \text{for all } 1\le j\le m\right\}$$

for all $\mu_m,\dots,\mu_1$ and all $1\le m\le k$.

For example

$$S_{=>}=\left\{x_1\dots x_n\in\{0,1\}^n:\sum_{i=1}^{\frac{n}{4}}x_{4i}=\sum_{i=1}^{\frac{n}{4}}x_{4i-2}\text{ and }\sum_{i=1}^{\frac{n}{2}}x_{2i}>\sum_{i=1}^{\frac{n}{2}}x_{2i-1}\right\}$$

and

$$S_> = \left\{ x_1 \dots x_n \in \{0,1\}^n : \sum_{i=1}^{\frac{n}{2}} x_{2i} > \sum_{i=1}^{\frac{n}{2}} x_{2i-1} \right\}.$$

Lemma 1. *For all* $\mu_m, \dots \mu_1 \in \{<,>,=\}^m$

$$S_{>\mu_m,\dots\mu_1} \cup S_{<\mu_m,\dots,\mu_1} \cup S_{=\mu_m,\dots\mu_1} = S_{\mu_m,\dots,\mu_1} \tag{1}$$

and

$$S_{>\mu_m,\dots\mu_1} \cap S_{<\mu_m,\dots} = \emptyset, \quad S_{>\mu_m,\dots\mu_1} \cap S_{=\mu_m,\dots} = \emptyset, \quad S_{<\mu_m,\dots\mu_1} \cap S_{=\mu_m,\dots} = \emptyset \tag{2}$$

Proof: The statements are obvious.

Lemma 2. *For all* $1 \leq m < k$ *and for all* $\mu_m, \dots, \mu_1 \in \{<,>,=\}^m$

$$|S_{>\mu_m,\dots,\mu_1}| = |S_{<\mu_m,\dots,\mu_1}| \tag{3}$$

Proof: For all $1 < j \leq k$

$$\sum_{i=1}^{\frac{n}{2^j}} x_{2^j i} + \sum_{i=1}^{\frac{n}{2^j}} x_{2^j i - 2^{j-1}}, = \sum_{i=1}^{\frac{n}{2^{j-1}}} x_{2^{j-1} i}. \tag{4}$$

Therfore for each sequence $x_1, \dots, x_n \in S_{>\mu_m,\dots,\mu_1}$ we exchange the positions $x_{2^j i}$ and $x_{2^j i - 2^{j-1}}$ for all $1 \leq i \leq \frac{n}{2^j}$ and keep all other positions unchanged. The corresponding sequence is in $|S_{<\mu_m,\dots,\mu_1}|$ and visa versa.

Now we conclude that $|S(t,n)| = |S_{\mu_k,\dots,\mu_1}|$ with $\mu_i \in \{>\}$.

Case $k = 1$

This is solved in [1].

Theorem 1. (see [1])

$$\lim_{n \to \infty} \frac{|S(2,n)|}{2^n} = \frac{1}{2}.$$

The authors of [1] have further proved

Lemma 3.

$$|S_=| = \sum_{i=0}^{\frac{n}{2}} \binom{\frac{n}{2}}{i}^2 = \binom{n}{\frac{n}{2}} \tag{5}$$

$$|S_>| = |S_<| = \frac{1}{2}(2^n - |S_=|) = \frac{1}{2}\left(2^n - \binom{n}{\frac{n}{2}}\right) \tag{6}$$

Case $k = 2$

Here the codewords have pairwise distance greater or equal to 4, i.e. our code is a 1-shift-error-correcting 2-shift-error-detecting code.

Lemma 4. *For all* $n = 2r, r \in \mathcal{N}$ *we have* $|S_{=>}| = |S_{=<}|$.

Proof: For each sequence $x_1, \dots x_n \in S_{=>}$ we exchange in all odd positions all zeroes by ones and all ones by zeroes. The first relation depends only on the even positions. Therefore with these exchanges we are getting a new sequence but the second relation is in the other direction, i.e. it is a sequence in $S_{=<}$ and visa versa.

Theorem 2. *For all* $r \in \mathbb{N}$ *and* $n = 4r$ *we have*

$$|S(4,n)| = \frac{1}{4}\left(2^n - \binom{n}{2r} - \sum_{i=0}^{r}\binom{r}{i}^2\left(2^{2r} - \binom{2r}{2i}\right)\right).$$

Proof: Notice that $|S(4,n)| = |S_{>>}|$. From Lemma 2 it follows that $|S_{>>}| = |S_{<>}|$. From Lemma 1 we get

$$|S_{>>}| + |S_{<>}| + |S_{=>}| = |S_>|.$$

and with equation 6 it follows that

$$\begin{aligned} |S_{>>}| &= \frac{1}{2}\left(|S_>| - |S_{=>}|\right) \\ &= \frac{1}{2}\left(\frac{1}{2}\left(2^n - \binom{n}{\frac{n}{2}}\right) - |S_{=>}|\right) \end{aligned} \tag{7}$$

Now it remains to calculate the size of $S_{=>}$. Let $l = \sum\limits_{i=1}^{r} x_{4i}$. Then we have $\binom{r}{l}$ possibilities to choose a binary sequence of length r for the positions j with $j \equiv 0 \bmod 4$. For the positions j with $j \equiv 2 \bmod 4$ we have the same number of possibilities. Therefore

$$|S_{=>}| + |S_{=<}| + |S_{==}| = \sum_{i=0}^{r}\binom{r}{i}^2 2^{2r}. \tag{8}$$

With

$$|S_{==}| = \sum_{i=0}^{r}\binom{r}{i}^2\binom{2r}{2i} \tag{9}$$

and Lemma 4 we have

$$|S_{=>}| = \frac{1}{2}\left(\sum_{i=0}^{r}\binom{r}{i}^2 2^{2r} - \sum_{i=0}^{r}\binom{r}{i}^2\binom{2r}{2i}\right)$$

and with equation 7

$$|S_{>>}| = \frac{1}{2}\left(\frac{1}{2}\left(2^n - \binom{n}{\frac{n}{2}}\right) - \frac{1}{2}\sum_{i=0}^{r}\binom{r}{i}^2\left(2^{2r} - \binom{2r}{2i}\right)\right)$$
$$= \frac{1}{4}\left(2^n - \binom{n}{\frac{n}{2}} - \sum_{i=0}^{r}\binom{r}{i}^2\left(2^{2r} - \binom{2r}{2i}\right)\right).$$

For the asymptotic estimate we use a well–known fact, which we prove for the sake of completeness.

Lemma 5.

$$\frac{\binom{n}{\frac{n}{2}}}{2^n} \approx \frac{2}{\sqrt{2\pi n}}.$$

Proof: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ by Stirling's formula and therefore

$$\frac{\binom{n}{\frac{n}{2}}}{2^n} = \frac{n!}{(\frac{n}{2})!(\frac{n}{2})!2^n} \approx \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{\sqrt{2\pi\frac{n}{2}}(\frac{n}{2e})^{\frac{n}{2}}\sqrt{2\pi\frac{n}{2}}(\frac{n}{2e})^{\frac{n}{2}}2^n} \approx \frac{2\cdot 2^n}{n\sqrt{2\pi}2^n} \approx \frac{2}{n\sqrt{2\pi}}.$$

Lemma 6.

$$\lim_{r\to\infty}\frac{|S(4,4r)|}{2^{4r}} = \frac{1}{4}.$$

Proof: By Theorem 2

$$\frac{S(4,4r)}{2^{4r}} = \frac{1}{4}\left(1 - \frac{\binom{4r}{2r}}{2^{4r}} - \frac{1}{2^{4r}} - \sum_{i=0}^{r}\binom{r}{i}^2\left(2^{2r} - \binom{2r}{2i}\right)\right)$$

and thus by Lemma 5

$$\left|\frac{S(4,4r)}{2^{4r}} - \frac{1}{4}\right| \le \frac{\binom{4r}{2r}}{2^{4r}} + \frac{1}{2^{4r}}(r+1)\binom{r}{\frac{r}{2}}^2 2^{2r} \approx \frac{2}{4r\sqrt{2\pi}} + (r+1)\left(\frac{2}{r\sqrt{2\pi}}\right)^2$$

and thus the result.

Simular calculations give

Theorem 3. *For $t = 2^k$ and $n = t\cdot r$*

$$\lim_{r\to\infty}\frac{|S(t,n)|}{2^n} = \frac{1}{t}.$$

Remark: In [3] there are further motivations for these questions related with parallelization of algorithms on transputer systems.

References

[1] R. Ahlswede, B. Balkenhol, and L.H. Khachatrian, Some properties of fix–free codes, Proceedings First INTAS International Seminar on Coding Theory and Combinatorics, Thahkadzor, Armenia, 20–33, 6.–11. October 1996.

[2] G.H. Khachatrian and S. Sargissian, Almost optimal single–shift self–synchronized codes, Proceedings First INTAS International Seminar on Coding Theory and Combinatorics, Thahkadzor, Armenia, 122–124, 6.–11. October 1996.

[3] T. Partner, Synchronization, Diploma thesis at Bielefeld University, in preparation.

[4] J.J. Stiffler, *Theory of Synchronous Communication*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[5] W. Wesley Peterson and E.J. (Jr.) Weldon, *Error–Correcting Codes*, MIT Press, Cambridge, MA and London, England, 1972.