2016

STATISTICAL PROPERTIES OF PSEUDORANDOM SEQUENCES

Ting Gu
University of Kentucky, gutinglizzy@gmail.com
Digital Object Identifier: http://dx.doi.org/10.13023/ETD.2016.159

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Ting Gu, Student

Dr. Andrew Klapper, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

STATISTICAL PROPERTIES OF PSEUDORANDOM SEQUENCES

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in the College of Engineering
at the University of Kentucky

By
Ting Gu
Lexington, Kentucky

Director: Dr. Andrew Klapper, Professor of Computer Science
Lexington, Kentucky

2016

ABSTRACT OF DISSERTATION

STATISTICAL PROPERTIES OF PSEUDORANDOM SEQUENCES

Random numbers (in one sense or another) have applications in computer simulation, Monte Carlo integration, cryptography, randomized computation, radar ranging, and other areas. It is impractical to generate random numbers in real life, instead sequences of numbers (or of bits) that appear to be "random" yet repeatable are used in real life applications. These sequences are called pseudorandom sequences. To determine the suitability of pseudorandom sequences for applications, we need to study their properties, in particular, their statistical properties. The simplest property is the minimal period of the sequence. That is, the shortest number of steps until the sequence repeats. One important type of pseudorandom sequences is the sequences generated by feedback with carry shift registers (FCSRs). In this dissertation, we study statistical properties of N-ary FCSR sequences with odd prime connection integer q and least period $(q-1)/2$. These are called half-ℓ-sequences. More precisely, our work includes:

- The number of occurrences of one symbol within one period of a half-ℓ-sequence;
- The number of pairs of symbols with a fixed distance between them within one period of a half-ℓ-sequence;
- The number of triples of consecutive symbols within one period of a half-ℓ-sequence.

In particular we give a bound on the number of occurrences of one symbol within one period of a binary half-ℓ-sequence and also the autocorrelation value in binary case. The results show that the distributions of half-ℓ-sequences are fairly flat. However, these sequences in the binary case also have some undesirable features as high autocorrelation values. We give bounds on the number of occurrences of two symbols with a fixed distance between them in an ℓ-sequence, whose period reaches the maximum and obtain conditions on the connection integer that guarantee the distribution is highly uniform.

In another study of a cryptographically important statistical property, we study a generalization of correlation immunity (CI). CI is a measure of resistance to Siegenthaler's divide and conquer attack on nonlinear combiners. In this dissertation, we

present results on correlation immune functions with regard to the q-transform, a generalization of the Walsh-Hadamard transform, to measure the proximity of two functions. We give two definitions of q-correlation immune functions and the relationship between them. Certain properties and constructions for q-correlation immune functions are discussed. We examine the connection between correlation immune functions and q-correlation immune functions.

KEYWORDS: FCSRs, half-ℓ-sequences, autocorrelation, correlation immunity, q-transform, q-correlation immune functions

Author's signature: Ting Gu

Date: May 3, 2016

STATISTICAL PROPERTIES OF PSEUDORANDOM SEQUENCES

By
Ting Gu

Director of Dissertation: Andrew Klapper

Director of Graduate Studies: Miroslaw Truszczynski

Date: May 3, 2016

To my family

ACKNOWLEDGMENTS

My sincere thanks to my advisor, Dr. Andrew Klapper, for his guidance and patience. He not only teaches me cryptography and sequences, but also reveals the beauty of knowledge. I am fortunate to learn from him during my doctoral studies. This dissertation would not have been possible without his inspiration and encouragement.

I would like to thank Dr. Judy Goldsmith for providing detailed comments on my presentation and advice during my job search. My special thanks to Dr. Greg Wasilkowski who shows great care to me and my family. My sincere thanks to Dr. Edgar Enochs for teaching me abstract algebra and clarifying all my questions on finite fields in great patience. Many thanks to Dr. Peter Hislop for serving as my outside examiner.

There are several other people in the computer science department who were supportive. I would like to thank Dr. Raphael Finkel for his support and help. Thanks Dr. Yi Pike and Dr. Debby Keen for their advice during my job search. Thanks Dr. Miroslaw Truszczyński for his supportive work during my graduation.

I would like to thank Dr. Zhixiong Chen for his guidance and encouragement during his visiting at University of Kentucky. Special thanks to my friend Yaowei Zhang, Department of Mathematics, University of Kentucky, for his helpful discussions on abstract algebra. Thanks the people in Crypto Seminar group for the wonderful time we spent together.

At last, I would like to thank my family. I am very grateful to my parents for providing me a relaxed and supportive environment for all these years' studies. My love to my husband and son, who always encourage me to pursue what I want and bring me laugh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Nowadays, information technologies and communications fit into every corner of our lives. People use the internet, banking systems, and mobile phones for social and business interactions. Their data transactions and footprints can leave clues about their most sensitive information and can cause harm in a variety of ways such as unauthorized illegal actions. The issues of guaranteeing secure transmission over public channels has been widely studied in cryptology [58]. Cryptographic techniques are used to provide confidentiality, authentication and data integrity during communication services such as email, banking or shopping on the internet. Historically, cryptography was developed for the purpose of protecting secret information for military and government organizations. Now it has become an indispensable tool used to protect information in modern digital society. This dissertation concerns statistical properties of certain high speed pseudorandom sequence generators. These have been suggested for use as building blocks for stream ciphers.

1.1 Cryptography

Cryptography involves the processes of encryption of a plaintext (or message) and decryption of the ciphertext (or encrypted message). Figure 1.1 shows the encryption and decryption process, where k_e is the encryption key and k_d is the decryption key.

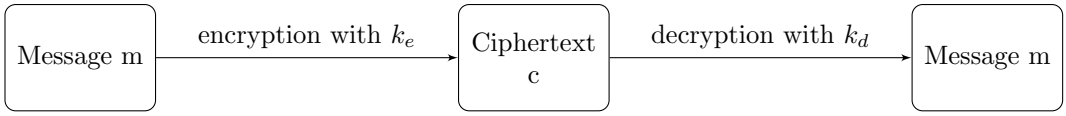

Figure 1.1: Encryption and Decryption

There are mainly two types of cryptographic algorithms. One is asymmetric cryptographic algorithms where $k_e \neq k_d$. These algorithms include the well known

public key crypto-system RSA. The other type of cryptographic algorithm is symmetric cryptographic algorithms where $k_e = k_d$ such as stream ciphers. The main goal of cryptography is to design systems for securely transmitting information over insecure channels. This goal is particularly challenging for transmission of large volumes of information when the encryption and decryption must be fast. Since the public key cryptographic algorithms are still far too slow for most practical needs, many applications use public key cryptographic tools for key exchange and symmetric key crypto-systems to provide confidentiality of data communication. The only symmetric crypto-system that is guaranteed to be secure is the one-time pad. In this system the message is first encoded as a binary sequence $M = m_0, m_1, m_2, \cdots$. The sender and receiver of the message each has a copy of a random binary sequence $K = k_0, k_1, k_2, \cdots$(the key). The message is encrypted by adding the key to the message bitwise modulo two, forming the cipher

$$C = c_0, c_1, c_2, \cdots = (m_0 + k_0 \pmod 2), (m_1 + k_1 \pmod 2), (m_2 + k_2 \pmod 2), \cdots .$$

The message is recovered by the receiver by the same operation. This method is provably secure against an adversary who knows part of the key. Moreover, it is extremely fast, requiring a single exclusive or for each transmitted bit. However, it requires the sender and receiver to securely share a key that is as large as the message, hence is impractical for most situations.

Stream ciphers are practical solutions that use the same methods of encryption and decryption as the one-time pad uses. They are typically used in digital telephones, video on demand, and other applications where high volume data is transmitted. For example, a stream cipher called $A5/1$ [11] is used to ensure the confidentiality of phone conversations. Another stream cipher $RC4$ [54] is adopted by Wired Equivalent Privacy (WEP), a part of the IEEE 802.11 wireless networking standard [20].

Other examples include stream cipher $E0$ in Bluetooth [8], an industry standard for short distance wireless networks. However, weakness and attacks on existing stream ciphers are discovered, often using sophisticated mathematical concepts. For example, the widely-used $RC4$ has serious flaws. As a result, several competitions were undertaken by the cryptographic community to address the lack of secure stream cipher standards that could be used by industry. These include the New European Schemes for Signature, Integrity and Encryption (NESSIE) project [52] organized by EU, the Cryptographic Research and Evaluation Committee (Cryptrec) [36] initiated by Japan and the European Network of Excellence for Cryptology (ECRYPT) Stream Cipher Project [55].

In a stream cipher the sequence K, however, is a pseudorandom sequence, which is a sequence of numbers (or of bits) that appears to be "random" yet repeatable. These sequences are often generated by a simple device called a keystream generator whose initial state has a short description and is thus more easily shared than the message. The goal of stream cipher design is to find keystream generators that have short descriptions, are fast, and whose security approaches that of the one-time pad.

1.2 Statistical Properties of Pseudorandom Sequences

Pseudorandom sequences play an important role in many areas of communications and computing such as the keystream sequences in stream ciphers, spread spectrum communications, error correcting codes, and quasi-Monte Carlo integration. To determine the suitability of pseudorandom sequences for application use, we need to study their properties.

1.2.1 Period

One property is the minimal period length of the sequence. That is, the shortest number of steps until the sequence begins to repeat.

Definition 1.2.1 *([30, p. 15]) A sequence* $\mathbf{a} = (a_0, a_1, \cdots)$ *is eventually periodic if there exists a* $p \in \mathbb{Z},\ p > 0$ *and* $u \geq 0$ *such that*

$$a_{i+p} = a_i \text{ for all } i > u \tag{1.1}$$

If $u = 0$, $\mathbf{a}$ *is strictly periodic or just periodic. The smallest* p *that satisfies the eq. (1.1) is called the period of* $\mathbf{a}$.

It is seen from the previous section that stream cipher encryption is based on addition modulo two, so all of the strength of the encryption is derived from the secrecy of the pseudorandom sequence. Since the pseudorandom sequence has a period, an attacker knows that any terms separated by the minimum period length were encrypted by addition modulo two with the same term and this is true for all terms after the sequence begins to repeat. As a result, information may be leaked, security may be lost. Therefore it is a basic requirement that a pseudorandom sequence that is used for a stream cipher encryption should have large period and in particular much longer than the message length.

1.2.2 Randomness

Pseudorandom sequences are required to satisfy several randomness properties; otherwise attacks may be launched based on the statistical deviation between the pseudorandom sequences and truly random sequences. In 1967 Golomb proposed three postulates for the appearance of binary periodic pseudorandom sequences.

- It should be balanced. The difference between the number of 1s and the number of 0s must be at most one.
- It should have the run property. About half of the runs (sequences of the same bit) must be of length one, one quarter of length two, one eighth of length three,

etc. Moreover, there must be equally many runs of 1s and of 0s for each of these lengths.

- It should have an ideal autocorrelation function.

A binary sequence which satisfies Golomb's randomness postulates is called a pseudo-noise sequence or a *pn*-sequence [48, p.181]. The properties stated in the above postulates can be empirically measured by various statistical tests. Pseudo-random sequences are supposed to pass all the statistical tests of randomness that can be found. Another important property is the unpredictability. Given the first k bits of the sequence, it should be computationally infeasible to predict the next $k+1$ bit unless the secret seed is given. The best way to generate unpredictable random numbers is by using some physical process such as thermal noise or radioactive decay. However, these methods are extremely inefficient. In practice, pseudorandom sequence generators based on deterministic algorithms are used to generate sequences of bits. A random seed is used for these generators. An attacker must not be able to make any correct predictions with probability significantly better than guessing without the seed. This should hold even with the knowledge of design detail of the generators. A pseudorandom sequence generator should have the following properties:

- good randomness properties of output sequences;
- speed and efficiency;
- reproducibility;
- large period.

1.3 Contribution

Feedback with carry shift registers (FCSRs), first introduced by Klapper and Goresky in 1993 [42], are high speed pseudorandom sequence generators based on integer ad-

dition. They are important building blocks of pseudorandom sequence generators in stream ciphers. Since speed and efficiency are important for pseudorandom sequence generators, it is vital to have efficient hardware or software implementations for FCSRs. Lee and Park [43] proposed software implementations for word-based FCSRs in 2011. They improved the efficiency of software implementation of FCSRs by increasing the size of register cells from 1 bit to k bits, where k is the size of words in a given CPU (e.g. $k = 32$). Sequences of k bits are produced at every clocking. Their implementations with simple arithmetic operators (such as shifts, maskings, xors, modular additions, etc.) over variables of size 2^{32} or 2^{16} are claimed to have better efficiency than usual methods using conditional operators (such as 'if' statements) to handle the carry in FCSRs. It is nice to see efficient software implementations for FCSRs, but we also need the generated FCSR sequences to have good statistical properties.

In this dissertation, we call the FCSR sequences generated by the FCSRs with the word-based software implementations as described by Lee and Park [43] half-ℓ-sequences [33] and investigate their statistical properties. These properties include:

- The number of occurrences of one symbol within one period of a half-ℓ-sequence;
- The number of pairs of symbols with a fixed distance between them within one period of a half-ℓ-sequence;
- The number of triples of consecutive symbols within one period of a half-ℓ-sequence.

In particular we give a bound on the number of occurrences of one symbol within one period of a binary half-ℓ-sequence and also the autocorrelation value in binary case. The results show that the distributions of half-ℓ-sequences are fairly flat. However, these sequences in the binary case also have some undesirable features such as high autocorrelation values. We give bounds on the number of occurrences of two symbols with a fixed distance between them in an ℓ-sequence, whose period reaches

the maximum and obtain conditions on the connection integer that guarantee the distribution is highly uniform.

In another study of a cryptographically important statistical property, we study a generalization of correlation immunity (CI). CI is a measure of resistance to Siegenthaler's divide and conquer attack on nonlinear combiners. In this dissertation, we present results on correlation immune functions with regard to the q-transform, a generalization of the Walsh-Hadamard transform, to measure the proximity of two functions. We give two definitions of q-correlation immune functions and the relationship between them. Certain properties and constructions for q-correlation immune functions are discussed. We examine the connection between correlation immune functions and q-correlation immune functions.

1.4 Organization

The remainder of this dissertation is structured as follows. Chapter 2 provides basics of linear feedback shift registers (LFSRs) and FCSRs. The structure of keystream generators is given. Useful mathematical tools and preliminaries on Boolean functions are listed. Chapter 3 and 4 discuss some statistical properties of pseudorandom sequences. More precisely, Chapter 3 investigates distribution properties of half-ℓ-sequences. We show that this type of sequences has a fairly flat distribution. In Chapter 4, a combination of topics is addressed, including the distribution properties of combined half-ℓ-sequences, autocorrelation of half-ℓ-sequences and distribution features of ℓ-sequences. Chapter 5 presents some results on correlation immune functions with regard to q-transforms. We discuss certain properties and construction for this type of function. In Chapter 6 we discuss future research directions on distribution properties of pseudorandom sequences.

Chapter 2 Background and Preliminaries

Shift register sequences are used in both cryptography and coding theory and many other areas [28]. Because shift registers can be easily implemented in digital hardware, many stream ciphers are made up of shift registers. Stream ciphers based on shift registers have been used for military cryptography since the beginning of electronics [58]. A feedback shift register consists of two parts: a shift register and a feedback function. The shift register is a sequence of bits, whose length is the number of bits it contains. Each time the new left-most bit is computed as a function of the bits in the register. The right-most bit becomes the output bit and all of the remaining bits in the shift register shift one bit to the right. The period of a shift register is the length of the output sequence before it starts to repeat.

In this chapter, we recall LFSRs and FCSRs. Detailed information on LFSRs and FCSRs can be found in the books by Golomb [28] and Goresky and Klapper [30]. We show some statistical properties of two important types of pseudorandom sequences: m-sequences and ℓ-sequences, which are the maximum period sequences of LFSRs and FCSRs respectively. We discuss different ways to represent Boolean functions and the structure of keystream generators. At the end of the chapter we present some useful mathematical tools that are used throughout this dissertation.

2.1 Linear Feedback Shift Registers

The simplest kind of feedback shift registers is the LFSR (see Figure 2.1). LFSRs have been studied intensively for over fifty years [28]. They are widely used in areas of cryptography and coding theory and other areas. They are simple and fast in hardware implementations. Various statistical properties of the output sequences of an LFSR can be fully analyzed by using efficient algebraic tools (principally using the

Galois theory of finite fields and the algebra of power series rings). Many fast devices use LFSRs as building blocks to generate sequences whose statistical properties are good. There are two modes of LFSRs: Fibonacci mode LFSR and Galois mode LFSR. In general, it is preferred to use Fibonacci mode LFSR if the hardware for implementation is effective at shifts and to use Galois mode LFSR if parallelisms can be exploited. In this section, for simplicity, we only describe the Fibonacci mode LFSR. Details on Galois mode LFSR can be found in Golomb's book [28]. We introduce an important type of pseudorandom sequence, the maximum period LFSR sequences, and discuss their statistical properties in this section.

Definition 2.1.1 *([30, p. 23]) A* (Fibonacci mode) *linear feedback shift register of length m over a commutative ring R, with coefficients*

$$q_1, q_2, \cdots, q_m \in R$$

is a sequence generator whose state is an element

$$s = (a_0, a_1, \cdots, a_{m-1}) \in R^m,$$

whose output is $\mathbf{out}(\mathbf{s}) = a_0$*, and whose state change operation τ is given by*

$$(a_0, a_1, \cdots, a_{m-1}) \to (a_1, a_2, \cdots, a_{m-1}, \sum_{i=1}^{m} q_i a_{m-i}).$$

Figure 2.1 shows the structure of a linear feedback shift register.

The output sequence $\mathbf{a} = a_0, a_1, \cdots$ of an LFSR of length m satisfies a linearly recurrence relation for all $n \geq m$

$$a_n = q_1 a_{n-1} + \cdots + q_m a_{n-m}. \tag{2.1}$$

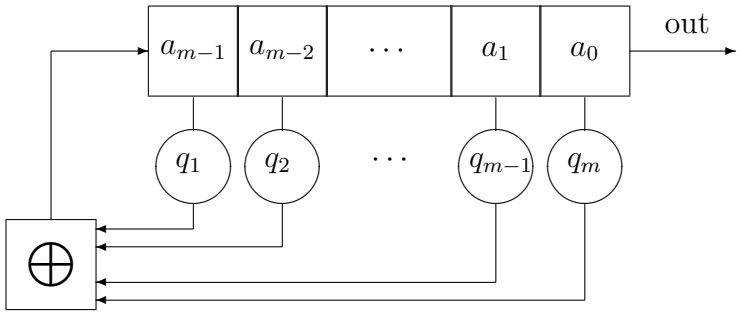

Figure 2.1: A Linear Feedback Shift Register of Length m.

The integer m is called the degree of the recurrence relation. The elements $q_1, \cdots, q_m$ are called the coefficients of the recurrence relation. This recurrence relation can be expressed by the polynomial

$$q(x) = -1 + \sum_{i=1}^{m} q_i x^i \in R[x].$$

This polynomial is called the connection polynomial or the feedback polynomial. It is reciprocal to the characteristic polynomial of the above linearly recurrence relation. The characteristic polynomial can be calculated by

$$f(x) = q(x^{-1}) \times x^m = -x^m + q_1 x^{m-1} + \cdots + q_{m-1} x + q_m.$$

If R is finite, then the number of possible states for an LFSR of a fixed length m is $|R|^m$, hence the state of the LFSR must repeat after $|R|^m$ steps. Therefore, the output sequence is *eventually periodic*. Conversely, any eventually periodic sequence $\mathbf{a}$ can be generated by an LFSR. The number of cells in the shortest LFSR that can generate $\mathbf{a}$ is called the linear complexity or linear span of $\mathbf{a}$. An LFSR with a primitive feedback polynomial is called a maximum-length LFSR and its output sequence is called an m-sequence. A formal definition of m-sequence is shown below.

Definition 2.1.2 *([30, p. 208]) The sequence* **a** *is an m-sequence (over finite ring R) of rank r (or degree r or span r) if it can be generated by an LFSR of length r, and if every nonzero block of length r occurs exactly once in each period of* **a**.

In other words, the sequence **a** is the output sequence of an LFSR that cycles through all possible nonzero states before it repeats and it has the maximum period $|R|^m - 1$. An m-sequence is balanced, is equidistributed to the order of the size of the corresponding LFSR, and has the run property. A sequence is equidistributed to order r if for every k $(1 \leq k \leq r)$ the number of occurrences of a block of length k is in the inclusive range between $\lfloor T/|A|^k \rfloor$ and $\lceil T/|A|^k \rceil$ where T is the period of the sequence and A is the finite set from which each symbol in the sequence comes. An m-sequence is a pn-sequence from the definition in Section 1.2.2. More results related to m-sequences can be found in Golomb's book [28].

2.2 Feedback with Carry Shift Registers

An FCSR is another type of feedback shift register that is similar to an LFSR, but with extra memory parts. More precisely, an FCSR is an arithmetic or with carry analog of an LFSR. It was first introduced by Klapper and Goresky in 1993 [39]. The idea was motivated by the cryptanalysis of the summation combiner [56]. They were also invented by Marsaglia [45, 46, 47] and Couture and L'Écuyer [23, 46] in the context of quasi-Monte Carlo simulation.

The analysis of FCSR sequences is based on N-adic numbers, which have been studied since at least the early 1900s. An N-adic number a is an infinite expression

$$a = \sum_{i=0}^{\infty} a_i N^i$$

where $a_i \in \{0, 1, \cdots, N-1\}$.

Let

$$\mathbb{Z}_N = \left\{ \sum_{i=0}^{\infty} a_i N^i, \quad a_i \in \{0, 1, \cdots, N-1\} \right\}.$$

By defining addition and multiplication operations of N-adic numbers, $\mathbb{Z}_N$ forms a ring and the additive inverse of the multiplicative identity element is

$$-1 = (N-1) \sum_{i=0}^{\infty} N^i.$$

This ring $\mathbb{Z}_N$ is an arithmetic or with carry analog of the ring of power series over $\mathbb{Z}/(N)$. Let $S = \{0, 1, \cdots, N-1\}$. Figure 2.2 shows an N-ary Fibonacci mode FCSR of length m with multipliers (or taps) $q_1, \cdots, q_m$ whose state is a collection

$$(a_0, \cdots, a_{m-1}; z) \quad \text{where} \quad a_i \in S \text{ and } z \in \mathbb{Z}.$$

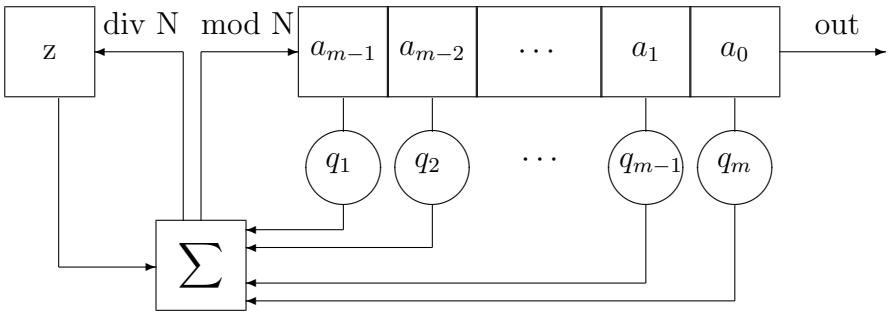

Figure 2.2: A Feedback with Carry Shift Register of Length m.

The state change of a Fibonacci mode FCSR is as follows. First we compute the linear combination

$$\sigma = \sum_{i=1}^{m} q_i a_{m-i} + z \in \mathbb{Z}.$$

Then the contents of the state cells shift one step to the right and output the element

in the right most cell. The left most cell and the next stage memory is updated by

$$a_m = \sigma \ mod \ N, \tag{2.2}$$

$$z^{'} = \sigma \ Div \ N = \frac{\sigma - a_m}{N}. \tag{2.3}$$

In other words, the new state is

$$(a_1, \cdots, a_{m-1}, \sigma \ (mod \ N); \sigma \ (div \ N)).$$

An FCSR can also be described in its Galois mode [30, 32, p. 154]. A Galois mode FCSR is preferred for parallelism consideration. One of the stream cipher candidates in eStream project, the F-FCSR stream ciphers [2, 3], uses an FCSR in Galois mode.

Let N be a positive integer and $S = \{0, 1, \cdots, N-1\}$. Denote by the symbol $\sum$ an integer adder (mod N) with carry. Figure 2.3 shows a Galois mode FCSR of length m with multipliers (or taps) $q_1, \cdots, \ q_m$ whose state is a collection

$$(a_0, a_1, \cdots, a_{m-1}; c_1, c_2, \cdots, c_m) \ \text{ where } \ a_i \in S \ \text{ and } \ c_i \in \mathbb{Z}.$$

The state change of a Galois mode FCSR is as follows. First we compute the linear combination

$$\sigma_m = c_m + q_m a_0$$

and

$$\sigma_j = c_j + a_j + q_j a_0, \ \ 1 \leq j < m.$$

Each of the new state cells and memory cells are updated individually by

$$a^{'}_{j-1} = \sigma_j \pmod{N}$$

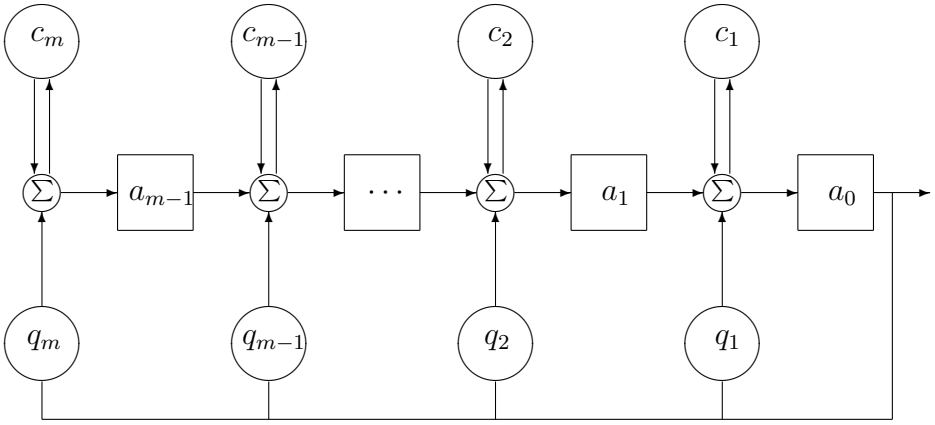

Figure 2.3: Galois FCSR

$$c_j^{'} = \sigma_j \text{ div } N = \lfloor \sigma_j / N \rfloor$$

such that

$$Nc_j^{'} + a_{j-1}^{'} = c_j + a_j + q_j a_0 \ (1 \leq j < m)$$

and

$$Nc_m^{'} + a_{m-1}^{'} = c_m + q_m a_0.$$

The output sequence $\mathbf{a} = a_0, a_1, \cdots$ satisfies a linearly recurrence relation with carry modulo N for all $n \geq m$

$$a_n + Nz_n = \sum_{i=1}^{m} q_i a_{n-i} + z_{n-1}. \tag{2.4}$$

The integer m is called the length (or span or degree) of the recurrence relation. The integers $q_1, \cdots, q_m$ are called the coefficients of the recurrence relation. The infinite sequence $z_{m-1}, z_m, z_{m+1}, \cdots$ of integers are the memory values. Set $q_0 = -1$. We

associate the recurrence relation with a connection integer

$$q = -1 + \sum_{i=1}^{m} q_i N^i = \sum_{i=0}^{m} q_i N^i. \tag{2.5}$$

The following important theorem reveals the relationship between N-adic numbers and the structure of an FCSR.

Theorem 2.2.1 *[30, 40] Let q be the connection integer of an FCSR with initial memory z_m and initial loading $a_0, a_1, \cdots, a_{m-1}$. Any output sequence* $\mathbf{a}$ *of this FCSR is the coefficient sequence of the N-adic representation of the rational number*

$$a = \sum_{i=0}^{\infty} a_i N^i = \frac{f}{q} \in \mathbb{Z}_N, \ f \in \mathbb{Z}.$$

An exponential representation is a useful tool to analyze shift register sequences. Our analyses of distribution properties of half-ℓ-sequence depend on the exponential representation of FCSR sequences, an analog of the trace representation of LFSR sequences. Let $q > 1$ and N be positive integers with $\gcd(N, q) = 1$. Throughout this dissertation, we shall consider only the case when q is an odd prime unless otherwise specified. An N-ary FCSR sequence $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ with connection integer q can be algebraically defined or exponentially represented by

$$a_i = N^{-i} h \pmod{q} \pmod{N}, \tag{2.6}$$

where $h \in \mathbb{Z}/(q)$, the residue ring modulo q. Here the notation $\pmod{q}$ $\pmod{N}$ means first that the number $N^{-i}h$ is reduced modulo q to give a number between 0 and $q-1$ and then the number is reduced modulo N to give a number in $\{0, 1, \cdots, N-1\}$.

We study the period of FCSR sequences here. The eventual period of an FCSR sequence with connection integer q is a divisor of $\mathrm{ord}_q(N)$, the multiplicative order of N modulo q. The largest possible value of $\mathrm{ord}_q(N)$ is $\phi(q)$, Euler's totient function

of q, and $\mathrm{ord}_q(N)|\phi(q)$. In the extreme case, $\phi(q) = q - 1$ when q is a prime number. In particular, when an FCSR sequence $\mathbf{s}$ achieves maximum period, in other words, $\mathrm{ord}_q(N) = \phi(q)$, $\mathbf{a}$ is referred to as an ℓ-sequence [29, 30].

Much work has been done on the statistical properties of ℓ-sequences. Goresky and Klapper showed that the number of occurrences of any two blocks in one period of an ℓ-sequence can differ at most by 1 if q is prime [41].

Theorem 2.2.2 *Let* $\mathbf{a}$ *be an* N*-ary* ℓ*-sequence based on a connection integer* $q = p^t$ *with an odd prime* p*. Then the number* $n(b)$ *of occurrences of any block* b *of size* s *within a single period of* $\mathbf{a}$ *is*

$$n_1 \leq n(b) \leq n_1 + 1 \quad \textit{if} \quad t = 1$$

$$n_1 - n_2 - 1 \leq n(b) \leq n_1 - n_2 + 1 \quad \textit{if} \quad t \geq 2,$$

where

$$n_1 = \left\lfloor \frac{q}{N^s} \right\rfloor = \left\lfloor \frac{p^t}{N^s} \right\rfloor \textit{ and } n_2 = \left\lfloor \frac{p^{t-1}}{N^s} \right\rfloor.$$

L-sequences are balanced and their arithmetic autocorrelations are zero [31, 42, 53]. The ordinary autocorrelation of binary ℓ-sequences has also been studied. Xu and Qi studied the expected value and variance of autocorrelations with prime power connection integer [66]. They showed that when the connection integer is sufficiently large, the autocorrelation for a random shift is low with high probability. Tian and Qi gave an upper bound on the autocorrelations of ℓ-sequences with prime connection integer [61]. More specifically, the autocorrelation is less than or equal to $2\left(\lceil q/6 \rceil - 1\right)$, where q is the prime connection integer. In this paper, we use the above result for the estimation of the autocorrelations of binary half-ℓ-sequences.

As in the case of linear complexity, the 2-adic complexity of a sequence is a measure of the length of FCSR required to output the sequence.

Definition 2.2.1 *Let* $\mathbf{a} = a_0, a_1, a_2, \cdots$ *be an eventually periodic binary sequence, whose* 2*-adic expression is*

$$\sum_{i=0}^{\infty} a_i 2^i = \frac{p}{q}$$

where $\gcd(p, q) = 1$*. Then the* 2*-adic complexity of* $\mathbf{a}$ *is the real number* $\log_2(\phi(p, q))$*, where* $\phi(p, q) = \max(|p|, |q|)$*.*

2.3 Boolean Functions

Boolean functions play crucial roles in the design of cryptographic primitives. In particular, they are widely used as components of pseudorandom sequence generators for stream ciphers. In this section, we recall different representations of Boolean functions and their properties. A more detailed description on Boolean functions can be found in Cusick and Stanica's book [25]. We begin with some notations that will be used throughout this dissertation.

Let $\mathbb{F}_2^n$ be the vector space of dimension n over the field $\mathbb{F}_2$ (Galois field with two elements). The lexicographical order of vectors is defined as: for $x, y \in \mathbb{F}_2^n$, $x \leq y$ if and only if there exists $i \in \{0, 1, \cdots, n-1\}$ such that $x_0 = y_0, \cdots, x_{i-1} = y_{i-1}$ and $x_i < y_i$. The Hamming weight of a vector x is denoted by $wt(x) = |\{i : x_i \neq 0\}|$, the number of nonzero positions. The support of a vector x is the set of indices of nonzero positions, i.e., $supp(x) = \{i : x_i \neq 0\}$. The Hamming distance between two vectors x and y, denoted by $d(x, y)$, is the number of positions in which x and y differ from each other. For the purpose of clarity, we use "$\oplus$" for addition modulo 2 and "$+$" for addition in $\mathbb{Z}$. The inner product or scalar product of two vectors x and y is denoted as $x \cdot y = x_1 y_1 \oplus \cdots \oplus x_n y_n$.

A Boolean function f in n variables is a map from $\mathbb{F}_2^n$ to $\mathbb{F}_2$. The set of all Boolean functions on $\mathbb{F}_2^n$ is denoted by $\mathbf{B}_n$. A Boolean function f can be uniquely represented

by its truth table as $(f(v_0), f(v_1), ..., f(v_{2^n-1}))$ where

$$v_0 = (0, ..., 0, 0), v_1 = (0, ..., 0, 1), ..., v_{2^n-1} = (1, ..., 1, 1)$$

are ordered by the lexicographical order.

Another way to uniquely represent a Boolean function f is by the polynomial

$$f(x) = \bigoplus_{\mathbf{a} \in \mathbf{F}_2^n} c_a x_1^{a_1} \cdots x_n^{a_n},$$

where $c_a \in \mathbb{F}_2$ and $\mathbf{a} = (a_1, \cdots, a_n)$ with $a_i \in \{0, 1\}$. This is called the algebraic normal form (ANF) of f. The number of variables in the highest order monomial with a nonzero coefficient is called the algebraic degree, or simply the degree of f.

The Boolean functions with degree less than or equal to 1 are called affine functions. They take the form $f_{\mathbf{a},c}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} \ \oplus \ c = a_1 x_1 \oplus \cdots \oplus a_n x_n \oplus c$, where $\mathbf{a} = (a_1, \cdots, a_n) \in \mathbf{F}_2^n$ and $c \in \mathbb{F}_2$. If $c = 0$, then $f_{\mathbf{a},0}$ $(= f_{\mathbf{a}})$ is a linear function.

A Boolean function can also be uniquely determined by its Walsh-Hadamard transform. The Walsh-Hadamard transform plays an important role in cryptography. For a Boolean function of n variables $f(x) = f(x_1, x_2, \cdots, x_n)$, the Walsh-Hadamard transform of $f(x)$ at $\omega \in \mathbb{F}_2^n$ is defined as

$$W(f)(\omega) = \sum_{x \in \mathbf{F}_2^n} (-1)^{f(x) \oplus \omega \cdot x}$$

The Walsh-Hadamard transform is a useful tool to analyze the properties of Boolean functions. For example, let $\bar{0} = (0, 0, \cdots, 0)$ be the zero vector, a Boolean function f is balanced if $W(f)(\bar{0}) = 0$. It can also be used to measure the nonlinearity of a Boolean function, the minimum distance from the set of all affine functions. The

nonlinearity of $f(x)$ is defined by

$$N_f = \min_{a \in \mathbb{F}_2^n / \{\bar{0}\}, b \in \mathbb{F}_2} |\{x \in \mathbb{F}_2^n : f(x) \neq a \cdot x \oplus b\}|.$$

For cryptographic applications, we generally require that $f(x)$ must be 'far' from the affine map $l(x) = a \cdot x \oplus b$, that is, N_f needs to be large. The Walsh-Hadamard transform of $f(x)$ can measure N_f by

$$N_f = 2^{n-1} - \frac{1}{2} \max_{\omega \in \mathbb{F}_2^n} |W(f)(\omega)|.$$

Due to the Parseval identity on $W(f)(\omega)$, i.e., $\sum_{\omega \in \mathbb{F}_2^n} W(f)(\omega)^2 = 2^{2n}$, we have

$$N_f \leq 2^{n-1} - 2^{n/2-1}.$$

$f(x)$ is called a bent function if the inequality above achieves equality. Indeed, $f(x)$ is bent if and only if $|W(f)(\omega)| = 2^{n/2}$ for all $\omega \in \mathbb{F}_2^n$.

Another notion of interest is the autocorrelation of a Boolean function. The autocorrelation of a Boolean function $f \in \mathbb{F}_2^n$ is a real-valued function defined as

$$r_f(\omega) = \sum_{x \in \mathbf{F}_2^n} (-1)^{f(x) \oplus f(\omega \oplus x)}$$

for all $\omega \in \mathbb{F}_2^n$.

Boolean functions used in cryptographic applications are required to satisfy certain properties to resist existing attacks. One such property is the correlation immunity of a Boolean function, which comes from the relation between $f(x)$ and linear functions. We explain this notion in more detail in Chapter 5. One can find details on other properties of Boolean functions in many books and papers [15, 16, 26, 57].

2.4 Keystream Generators

In a stream cipher, a keystream generator (see Definition 2.4.1) produces a pseudo-random sequence of bits. This sequence is added bit by bit with the plaintext to get the ciphertext [30, p. 9]. Stream ciphers are used for transmitting large amounts of data. They are extremely fast and are often implemented in hardware for added speed. The keystream generator must be designed to produce a pseudorandom sequence with enormous period using a relatively simple algorithm. The keystream sequence should be unpredictable from the knowledge of a (relatively small) segment of the sequence; otherwise, it is vulnerable to a known-plaintext attack. In a known-plaintext attack, the attacker attempts to recover a large section of keystream with knowledge of a relatively small section of plaintext.

Definition 2.4.1 *[30, p. 20] A keystream generator, or a discrete state machine with output*

$$F = (U, \Sigma, f, g)$$

consists of a discrete (i.e., finite or countable) set U of states, a discrete alphabet Σ of output values, a state transition function $f : U \to U$, and an output function $g : U \to \Sigma$.

Given an initial state $\mathbf{s} \in U$, such a keystream generator outputs an infinite sequence

$$F = g(\mathbf{s}), g(f(\mathbf{s})), g(f^2(\mathbf{s})), \cdots$$

with elements in Σ. Fig. 2.4 shows a keystream generator with initial state s. The state $\mathbf{s}$ is periodic of period T if starting from $\mathbf{s}$, after T steps, the generator returns to the state $\mathbf{s}$. That is, if $f^T(\mathbf{s}) = \mathbf{s}$. The least period of such a periodic state is the least such $T \geq 1$. A state $\mathbf{s}$ is eventually periodic if, starting from $\mathbf{s}$, after a finite number of steps, the generator arrives at a periodic state.

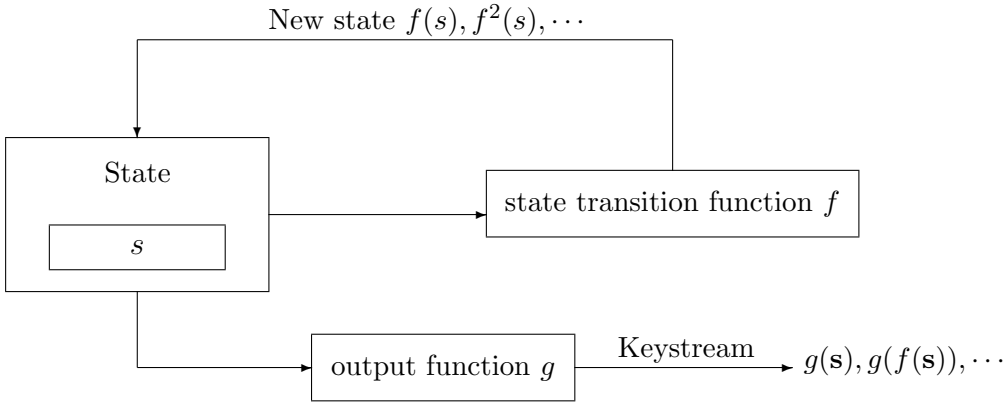

Figure 2.4: Keystream Generator

LFSRs are pseudorandom generators by themselves, but they have some unpleasant features. For an LFSR of length n, the internal state is the next n output bits of the generator. By observing the first n bits of output, the attacker can obtain the initial internal state. Since the state transition function of an LFSR is linear, all the generated sequence of this LFSR can be retrieved by solving a linear system of equations of the initial internal state. This property makes an LFSR by itself inappropriate for encryption. Also, the feedback coefficients of an LFSR with length n can be determined from only $2n$ output bits of the generator by using Berlekamp-Massey algorithm [7]. As a result, keystream generators based on LFSRs often employ nonlinear output functions. There are two widely used generator structures based on LFSRs. One structure employs several LFSRs with a combiner function as shown in Figure 2.5. The output bits of each LFSR serve as the input of a nonlinear Boolean function $f(x)$. The keystream sequences are the output bits of $f(x)$. Another structure is an LFSR with a nonlinear filter function as shown in Figure 2.6. The internal states of one LFSR serve as the input of a nonlinear function and this nonlinear filter function outputs the keystream sequences. There are various examples of stream ciphers using LFSRs as building blocks. These include ABC [1], SOSEMANUK [6], Sfinks [10], A5 [11], Yamb [24], Snow [27], Polar Bear [34], WG [50] and E_0 [60],.

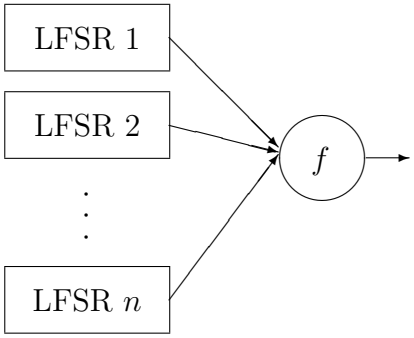

Figure 2.5: A Nonlinear Combination Generator

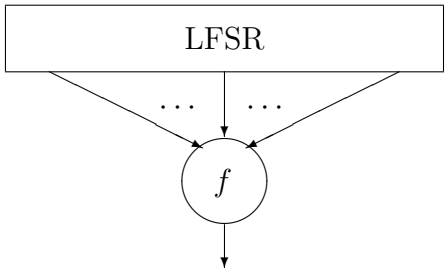

Figure 2.6: A Nonlinear Filter Generator

Because the state transition or update function for an LFSR is linear, keystream generators based on LFSRs still have weakness. For example, they are vulnerable to algebraic attacks [13, 14, 21, 22, 37]. The basic principle of algebraic attacks is to express the whole cipher as a large system of multivariate algebraic equations. This system of equations can be solved to recover the secret key or the initial state of the LFSRs. However, this fails for FCSRs. These devices were suggested to use for generation of binary sequences with large periods. They share many good properties with LFSR sequences. Their inherent non-linearity makes them promising building blocks in stream cipher design. There have been several proposals of stream ciphers

based on FCSRs [2, 4, 5]. One of them is the F-FCSR stream cipher in the eStream project [55]. The F-FCSR stream cipher uses an FCSR in Galois mode and takes a linear combination of the state bits to produce output. The stream cipher is extremely fast due to the very simple output function. LFSR based stream ciphers are also vulnerable to correlation attack. This attack works by exploiting known correlations between inputs (or combinations of inputs) and outputs (or combinations of outputs). Such attack depends on finding statistical biases in the outputs and reduces the expected number of keys that must be tested in a search of the key space. We will recall more about this attack in Chapter 5.

2.5 Mathematic Tools

The important mathematical tools used in this paper are exponential sums and the discrete Fourier transform. Exponential sums are significant techniques in number theory and are useful in various applications of finite fields. Group homomorphisms called characters play a basic role in analyzing exponential sums in finite fields.

A character of an Abelian group $\mathbf{G}$ is a group homomorphism from $\mathbf{G}$ to the multiplicative group $\mathbb{C}^\times = \mathbb{C} \setminus 0$ of the complex numbers. That is, it is a function $\chi : \mathbf{G} \to \mathbb{C}$ such that $\chi(ab) = \chi(a)\chi(b)$. Let $\mathbf{R}$ be a ring. An additive character is a character on the additive group of $\mathbf{R}$. We can get a multiplicative character on the multiplicative group of units of $\mathbf{R}$. The Legendre symbol (see definition (2.5.1)) is an example of a multiplicative character.

An integer a is called a quadratic residue modulo n if $\gcd(a, n) = 1$ and

$$x^2 \equiv a \pmod{n} \tag{2.7}$$

has a solution (i.e., if a is a "perfect square modulo n"). If eq. (2.7) has no solution, we say a is a quadratic nonresidue modulo n.

Note here that if integer a does not satisfy the condition that $\gcd(a, n) = 1$, it can not be classified as a quadratic residue or a quadratic nonresidue. In particular, 0 is considered neither a quadratic residue nor a quadratic nonresidue. While the modulus n can be any arbitrary positive integer, we are more interested in the case when n is an odd prime.

Proposition 2.5.1 *Let p be an odd prime. Then any $a \in V = \{1, 2, \cdots, p-1\}$ is either a quadratic residue or a quadratic nonresidue modulo p.*

In other words, exactly half of $|V|$ (i.e., $(p-1)/2$) are quadratic residues modulo p and exactly half are quadratic nonresidue modulo p.

In number theory, the Legendre symbol is an important quadratic character $\chi(a) := \left(\frac{a}{p}\right)$ where $a \in \mathbb{F} = \mathbb{F}_p$ with p prime.

Definition 2.5.1 *Let p be an odd prime. Let a be an integer. The Legendre symbol of a modulo p is defined as $\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}$ and*

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } a \text{ is a quadratic residue modulo } p, \\ -1 & \text{if } a \text{ is a quadratic nonresidue modulo } p, \\ 0 & \text{if } a \equiv 0 \pmod{p}. \end{cases}$$

The properties of the Legendre symbol are useful in the analysis of distribution properties of pseudorandom sequences in next chapter.

Lemma 2.5.1 *Let p be an odd prime. Let $\gcd(a, p) = 1$ and $\gcd(b, p) = 1$ where $a, b \in \mathbb{Z}$. We have the following properties related to the Legendre symbol.*

- *if $a \equiv b \pmod{p}$, then $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$;*
- $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$;

- $\left(\frac{a^2}{p}\right) = 1$;
- $\left(\frac{-1}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1 \pmod 4, \\ -1 & \text{if } p \equiv 3 \pmod 4. \end{cases}$
- $\left(\frac{2}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1, 7 \pmod 8, \\ -1 & \text{if } p \equiv 3, 5 \pmod 8. \end{cases}$

We consider the sum of characters over all elements of a field. Notice that for any $a \in \mathbb{F}_p$ with p prime, $\chi(a)$ is a complex root of unity since if $a^k = 1$ we have $(\chi(a))^k = \chi(a^k) = \chi(1) = 1$. Let ξ be a complex primitive qth root of unity. In other words, q is the smallest number such that $\xi^q = 1$. The exponential sums enter into our problem by means of the following well known basic identity [62]

$$\sum_{b=0}^{q-1} \xi^{bc} = \begin{cases} q & \text{if } c \equiv 0 \pmod q, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Our analysis is based on Weil's theorem, which is the one of the most beautiful results in 20th century mathematics.

Lemma 2.5.2 *(Weil's Theorem [44, p. 223]) Let q be a prime greater than 2 and ξ be a complex primitive qth root of unity. For a polynomial $g(x) \in (\mathbb{Z}/(q))[x]$ with* $\deg(g) \geq 1$, *we have*

$$\left| \sum_{c \in \mathbb{Z}_q} \xi^{g(c)} \right| \leq (\deg(g) - 1) q^{1/2}.$$

In particular,

$$\left| \sum_{z \in Q} \xi^{bz} \right| = \frac{1}{2} \left| \sum_{c=1}^{q-1} \xi^{bc^2} \right| = \frac{1}{2} \left| \sum_{c=0}^{q-1} \xi^{bc^2} - 1 \right| \leq \frac{1}{2} \left(q^{1/2} + 1 \right),$$

where $b \not\equiv 0 \pmod q$ and Q is the set of quadratic residues modulo q.

The Fourier transform of a complex valued function $f : \mathbb{Z}/(q) \to \mathbb{C}$ is given by

$$\hat{f}(b) = \frac{1}{q} \sum_{c=0}^{q-1} f(c)\xi^{-bc}.$$

By the Fourier inversion formula we have

$$f(c) = \sum_{b=0}^{q-1} \hat{f}(b)\xi^{bc}.$$

We also need the following two lemmas.

Lemma 2.5.3 *Let ξ be a complex primitive qth root of unity. For positive integers c_1, c_2, N and z, we have*

$$\left| \sum_{j=c_1}^{c_2} \xi^{jNz} \right| = \left| \frac{\sin(\pi Nz(c_2 - c_1 + 1)/q)}{\sin(\pi Nz/q)} \right|.$$

Proof. We have

$$\begin{aligned} \left| \sum_{j=c_1}^{c_2} \xi^{jNz} \right| &= \left| \sum_{j=c_1}^{c_2} \xi^{jNz} \right| = \left| \frac{\xi^{Nz(c_2-c_1+1)} - 1}{\xi^{Nz} - 1} \right| \\ &= \left| \frac{\sin(\pi Nz(c_2 - c_1 + 1)/q)}{\sin(\pi Nz/q)} \right|. \end{aligned}$$

□

Lemma 2.5.4 *[18] For positive integers q and b with $q > 1$, we have*

$$\sum_{b=1}^{q-1} \frac{|\sin(\pi bn/q)|}{|\sin(\pi b/q)|} < \frac{4}{\pi^2} q \log q + 0.38q + 0.608 + 0.116\frac{d^2}{q}$$

where $d = \gcd(b, q)$.

Chapter 3 Distribution Properties of Half-ℓ-sequence

3.1 Introduction

FCSRs are important building blocks of pseudorandom sequence generators in stream ciphers. Since speed and efficiency are significant for pseudorandom sequence generators, it is vital to have efficient hardware or software implementations for FCSRs. Lee and Park [43] proposed software implementations for word-based FCSRs in 2011. They improved the efficiency of software implementation of FCSRs by extending the size of register cells from 1 bit to k bits where k is the size of words in a given CPU (e.g. $k = 32$). Sequences of k bits are produced at every clocking. Their two implementations using full-size words (32 bits) and half-size words (16 bits) require the connection integer of corresponding FCSRs to be congruent to -1 modulo N where $N = 2^{32}$ or 2^{16}. Let q be the connection integer. When $q \equiv -1 \pmod{N}$ with $N = 2^k$ and $k \geq 3$, 2 is a quadratic residue (QR) modulo q by the law of quadratic reciprocity [49]. Hence also $N = 2^k$ is a QR modulo q. So every power of N is a QR modulo q. Then the multiplicative order of N modulo q is at most $\phi(q)/2$. Since N is a quadratic residue modulo q, it follows from eq. (2.6), the exponential representation of an FCSR sequence, that for every h, either all $N^{-i}h \pmod{q}$ are quadratic residues or all are non-quadratic residues. It follows that the period of the corresponding sequence is at most $(q-1)/2$ when q is a prime. We give a definition of these sequences and call them half-ℓ-sequences [33]. The purpose of this chapter is to estimate the distribution of the number of occurrences of one symbol, the number of pairs of symbols with a fixed distance between them, and the number of triples of consecutive symbols within one period of a half-ℓ-sequence.

Let $N = 2^k$ with $k \geq 3$ and $q = q_0 + mN$ for $q_0, m \in \mathbb{Z}$ with $0 \leq q_0 < N$ and $\gcd(q_0, N) = 1$. We consider an N-adic sequence $\mathbf{s}$. I.e., $\mathbf{s}$ is generated by an FCSR

Γ whose connection integer is q.

Definition 3.1.1 *A sequence* $\mathbf{s}$ *with prime connection integer* q *is called a half-*ℓ*-sequence if the period of* $\mathbf{s}$ *is* $(q-1)/2$.

Let $\mathbf{s} = \{s_i\}_{i=0}^{\infty}$ be a half-ℓ-sequence with period T and connection integer q. We consider the distribution of s_n, the distribution of $(s_n, s_{n+\tau})$, and the distribution of (s_n, s_{n+1}, s_{n+2}) for $\mathbf{s}$ with $0 < \tau < T$. For $j = 0, 1, \cdots$, let

$$\frac{u_j}{q} = \sum_{i=0}^{\infty} s_{i+j} N^i$$

be the N-adic expression of sequence $\mathbf{s}$ starting from s_j. Then u_j and u_{j+1} are related by the equation $u_j = qs_j + Nu_{j+1}$. Thus

$$u_{j+1} \equiv N^{-1} u_j \pmod{q} \tag{3.1}$$

and

$$s_j \equiv q^{-1} u_j \pmod{N}. \tag{3.2}$$

From eq. (3.1) it follows that either all u_j are quadratic residues modulo q, or all are non-quadratic residues modulo q. In the sequel, we always suppose they are quadratic residues modulo q without loss of generality. On the other hand, from eq. (3.2) it follows that for any $v \in \{0, 1, \cdots, N-1\}$, the number of occurrences of v in one period of $\mathbf{s}$ equals the number of quadratic residues u with $u \equiv qv \pmod{N}$.

For $0 \le \tau < T$ let

$$G_v^{(\tau)} = \{x \in \mathbb{Z} : 0 \le x < q, N^\tau x (\mathrm{mod} q) \equiv v (\mathrm{mod} N)\},\ 0 \le v < N.$$

Let m_v be the largest integer j such that $v + jN < q$ for $0 \leq v < N$. We have

$$G_v^{(0)} = \{v + jN : 0 \leq j \leq m_v\}.$$

Let $Z_k = \{x \in \mathbb{Z} : \lceil kq/N^\tau \rceil \leq x \leq \lfloor (k+1)q/N^\tau \rfloor\}$ for $0 < \tau < T$, $0 \leq k < N^\tau - 1$ and $Z_{N^\tau - 1} = \{x \in \mathbb{Z} : \lceil (N^\tau - 1)q/N^\tau \rceil \leq x < q\}$. We find that $G_v^{(\tau)}$ is the disjoint union of $N^{\tau-1}$ many Z_k with $-kq \equiv v \pmod{N}$, i.e.,

$$G_v^{(\tau)} = \bigcup_{j=0}^{N^{\tau-1}-1} Z_{t_v + jN}, \tag{3.3}$$

where $0 \leq t_v < N$ satisfies $t_v \equiv -vq^{-1} \pmod{N}$. Also, $|Z_{t_v+jN}| \leq \lceil q/N^\tau \rceil < q/N^\tau + 1$.

Define complex valued functions $g_v^{(\tau)} : \mathbb{Z}/(q) \to \mathbb{C}$ as

$$g_v^{(\tau)}(x) = \begin{cases} 1, & \text{if } x \in G_v^{(\tau)}, \\ 0, & \text{otherwise.} \end{cases}$$

The Fourier transform of $g_v^{(\tau)}(x)$ is given by

$$\hat{g}_v^{(\tau)}(x) = \frac{1}{q} \sum_{c \in G_v^{(\tau)}} \xi^{-xc}.$$

3.2 Distribution of s_n

We consider the distribution of elements in one period of a half-ℓ-sequence $\mathbf{s}$. For $0 \leq v_1 < N$, let $\mu(v_1)$ be the number of integers n with $s_n = v_1$ for $0 \leq n < T$. From eq. (3.2), we know that $\mu(v_1)$ equals the number of quadratic residues that are congruent to

$$(qv_1 \pmod{N}).$$

Let Q be the set of quadratic residues. From eq.(3.3), we have

$$\mu(v_1) = |Q \cap G_{v_1}^{(0)}|.$$

Theorem 3.2.1 *For an N-ary half-ℓ-sequence* $\mathbf{s}$ *with prime connection integer q, the number $\mu(v_1)$ of occurrences of s_n with $s_n = v_1$ for $0 \le n < T$ satisfies*

$$\left|\mu(v_1) - \frac{q-1}{2N}\right| \;<\; \frac{q-1}{2q} + \frac{(q^{1/2}+1)}{2} \cdot \left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right),$$

where $0 \le v_1 < N$.

Proof. Using Fourier transforms, we have

$$\begin{aligned}\mu(v_1) &= \sum_{x\in Q} g_{v_1}^{(0)}(x) = \sum_{x\in Q}\sum_{a=0}^{q-1} \hat{g}_{v_1}^{(0)}(a)\xi^{ax} \\ &= \sum_{a=0}^{q-1} \hat{g}_{v_1}^{(0)}(a) \sum_{x\in Q} \xi^{ax}.\end{aligned}$$

We first consider the case when $a = 0$.

$$\hat{g}_{v_1}^{(0)}(0) = \frac{1}{q}\sum_{c\in G_{v_1}^{(0)}} \xi^0 = \frac{1}{q}\left|\{0 \le v_1 + jN < q : 0 \le j \le m_{v_1}\}\right|. \tag{3.4}$$

We want to determine the number of j in eq. (3.4). We have

$$|\{0 \le v_1 + jN < q : 0 \le j \le m_{v_1}\}| = \left\lfloor \frac{q - v_1}{N} \right\rfloor + 1.$$

Using the formula

$$\frac{n_1}{n_2} - 1 < \left\lfloor \frac{n_1}{n_2} \right\rfloor \le \frac{n_1}{n_2},$$

we can get the lower and upper bounds

$$\frac{q-v_1}{N} < \left\lfloor \frac{q-v_1}{N} \right\rfloor + 1 \leq \frac{q-v_1}{N} + 1.$$

Hence we have

$$\frac{1}{N} - \frac{v_1}{Nq} < \hat{g}_{v_1}^{(0)}(0) \leq \frac{1}{N} + \frac{N-v_1}{Nq}.$$

Since when $a = 0$ we have

$$\sum_{x \in Q} \xi^{ax} = \frac{q-1}{2}$$

then

$$\left(\frac{1}{N} - \frac{v_1}{Nq}\right) \frac{q-1}{2} < \hat{g}_{v_1}^{(0)}(0) \sum_{x \in Q} \xi^{ax} \leq \left(\frac{1}{N} + \frac{N-v_1}{Nq}\right) \frac{q-1}{2}.$$

Then we consider the case when $a \neq 0$. We have

$$\begin{aligned}
\left| \sum_{a=1}^{q-1} \hat{g}_{v_1}^{(0)}(a) \sum_{x \in Q} \xi^{ax} \right| &\leq \left| \sum_{a=1}^{q-1} \hat{g}_{v_1}^{(0)}(a) \right| \left| \sum_{x \in Q} \xi^{ax} \right| \\
&= \left| \sum_{a=1}^{q-1} \frac{1}{q} \sum_{c \in G_v^{(0)}} \xi^{-ac} \right| \left| \sum_{x \in Q} \xi^{ax} \right| \\
&= \frac{1}{q} \left| \sum_{a=1}^{q-1} \sum_{j=0}^{m_{v_1}} \xi^{a(v_1+jN)} \right| \left| \sum_{x \in Q} \xi^{ax} \right|.
\end{aligned}$$

Putting everything together, we get

$$\begin{aligned}
\mu(v_1) &= \sum_{a=0}^{q-1} \hat{g}_{v_1}^{(0)}(a) \sum_{x \in Q} \xi^{ax} \\
&= \hat{g}_{v_1}^{(0)}(0) \sum_{x \in Q} \xi^0 + \sum_{a=1}^{q-1} \hat{g}_{v_1}^{(0)}(a) \sum_{x \in Q} \xi^{ax}.
\end{aligned}$$

We have

$$\begin{aligned} & \left|\mu(v_1) - \frac{q-1}{2N}\right| \\ = & \left|\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 + \sum_{a=1}^{q-1}\hat{g}^{(0)}_{v_1}(a)\sum_{x\in Q}\xi^{ax} - \frac{q-1}{2N}\right| \\ \leq & \left|\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N}\right| + \left|\sum_{a=1}^{q-1}\hat{g}^{(0)}_{v_1}(a)\sum_{x\in Q}\xi^{ax}\right|. \end{aligned}$$

Then we apply the bounds on $\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^{ax}$ to get

$$\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N} > \left(\frac{1}{N} - \frac{v_1}{Nq}\right)\frac{q-1}{2} - \frac{q-1}{2N}$$

and

$$\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N} \leq \left(\frac{1}{N} + \frac{N-v_1}{Nq}\right)\frac{q-1}{2} - \frac{q-1}{2N}.$$

That is,

$$\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N} > \left(-\frac{v_1}{Nq}\right)\frac{q-1}{2}$$

and

$$\hat{g}^{(0)}_{v_1}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N} \leq \left(\frac{N-v_1}{Nq}\right)\frac{q-1}{2}.$$

Then we have

$$\left|\hat{g}_{v_1}^{(0)}(0)\sum_{x\in Q}\xi^0 - \frac{q-1}{2N}\right| < \max\left\{\left|\left(-\frac{v_1}{Nq}\right)\frac{q-1}{2}\right|, \left|\left(\frac{N-v_1}{Nq}\right)\frac{q-1}{2}\right|\right\},$$

and hence

$$\begin{aligned}
&\left|\mu(v_1) - \frac{q-1}{2N}\right| \\
< \;& \frac{q-1}{2q} + \left|\sum_{a=1}^{q-1}\hat{g}_{v_1}^{(0)}(a)\sum_{x\in Q}\xi^{ax}\right| \\
\leq \;& \frac{q-1}{2q} + \frac{1}{q}\left|\sum_{a=1}^{q-1}\sum_{j=0}^{m_{v_1}}\xi^{a(v_1+jN)}\right|\left|\sum_{x\in Q}\xi^{ax}\right| \\
\leq \;& \frac{q-1}{2q} + \frac{(q^{1/2}+1)}{2q}\cdot\left(\frac{4}{\pi^2}q\log q + 0.38q + 0.608 + \frac{0.116}{q}\right),
\end{aligned}$$

which completes the proof. The last inequality follows from Lemma 2.5.3 and Lemma 2.5.4. □

In [33], we first gave the following bound

$$\left|\mu(v_1) - \frac{q-1}{2N}\right| \leq \frac{1}{2}\left(1 + \ln\left(\frac{q-1}{2}\right)\right)\left(q^{1/2}+1\right)$$

for the half-ℓ-sequence in the one symbol case. Wang and Tan presented a new bound in the one symbol case by using a more tighter bound on $1/\sin x$ [63]. This new bound is

$$\left|\mu(v) - \frac{q-1}{2N}\right| \leq \left(\frac{1}{\pi}\ln\left(\frac{q-1}{2}\right) + 0.3\right)\left(q^{1/2}+1\right) + \frac{q-1}{2q}.$$

In this section, we give a tighter bound than that in [33, 63]. In Table 3.1, we presents the comparison results of our bound with their bounds using the same series

of connection integers as were used by Wang and Tan [63].

Table 3.1: Comparison of the bounds in one symbol case

Connection integer q	bound in [33]	bound in [63]	our new bound
47	16	10	8
211	43	28	20
401	66	42	30
977	116	73	51
2003	180	114	79
4001	276	175	120
8191	426	270	184
20011	727	460	313
40009	1095	694	470
99991	1874	1188	800
131071	2194	1390	936
398287	4172	2645	1772
662551	5586	3542	2368
956261	6890	4369	2918
1299743	8206	5204	3472
3029711	13263	8412	5598
9999991	25978	16480	10933
2147483647	504997	320643	210591

3.3 Distribution of $(s_n, s_{n+\tau})$

We consider the distribution of $(s_n, s_{n+\tau})$ for $\mathbf{s}$ with $0 < \tau < T$. For $0 \leq v_1, v_2 < N$, let $\mu(\tau; v_1, v_2)$ be the number of integers n with $s_n = v_1$ and $s_{n+\tau} = v_2$ for $0 \leq n < T$. From eq. (3.2), we know that $\mu(\tau; v_1, v_2)$ equals the number of quadratic residues of blocks of length $\tau + 1$ that are congruent to

$$(qv_1 \pmod N, \mathbf{b}, qv_2 \pmod N),$$

where $\mathbf{b}$ is a block of elements of $\mathbb{Z}/(N)$. From eq. (3.3), we have

$$\mu(\tau; v_1, v_2) = |Q \cap G_{v_2}^{(0)} \cap G_{v_1}^{(\tau)}|.$$

Theorem 3.3.1 *For an N-ary half-ℓ-sequence* $\mathbf{s}$ *with prime connection integer q and $0 \le \tau < T$, the number $\mu(\tau; v_1, v_2)$ of occurrences of $(s_n, s_{n+\tau})$ with $s_n = v_1$ and $s_{n+\tau} = v_2$ for $0 \le n < T$ satisfies*

$$\begin{aligned}\left|\mu(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| &< \frac{(q-1)(2N^{\tau-2} + N^{\tau-1})}{2q} \\ &\quad + \frac{(q^{1/2}+1)N^{\tau-1}}{2}\left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right),\end{aligned}$$

where $0 \le v_1, v_2 < N$.

Proof. Using Fourier transforms, we have

$$\begin{aligned}\mu(\tau; v_1, v_2) &= \sum_{x\in Q} g_{v_2}^{(0)}(x) g_{v_1}^{(\tau)}(x) = \sum_{x\in Q}\sum_{a=0}^{q-1} \hat{g}_{v_2}^{(0)}(a)\xi^{ax} \sum_{b=0}^{q-1} \hat{g}_{v_1}^{(\tau)}(b)\xi^{bx} \\ &= \sum_{a=0}^{q-1} \hat{g}_{v_2}^{(0)}(a) \sum_{z=0}^{q-1} \hat{g}_{v_1}^{(\tau)}(z-a) \sum_{x\in Q} \xi^{zx} \quad (\text{we use } z = a+b) \\ &= \sum_{a=0}^{q-1} \frac{1}{q} \sum_{c\in G_{v_2}^{(0)}} \xi^{-ac} \sum_{z=0}^{q-1} \frac{1}{q} \sum_{d\in G_{v_1}^{(\tau)}} \xi^{-d(z-a)} \sum_{x\in Q} \xi^{zx} \\ &= \frac{1}{q^2} \sum_{z=0}^{q-1}\sum_{x\in Q} \xi^{zx} \sum_{a=0}^{q-1}\sum_{j=0}^{m_{v_2}} \xi^{a(v_2+jN)} \sum_{d\in G_{v_1}^{(\tau)}} \xi^{-d(z-a)} \\ &= \frac{1}{q^2} \sum_{z=0}^{q-1}\sum_{x\in Q} \xi^{zx} \sum_{j=0}^{m_{v_2}} \sum_{d\in G_{v_1}^{(\tau)}} \xi^{-dz} \sum_{a=0}^{q-1} \xi^{a(d+v_2+jN)} \\ &\triangleq \frac{1}{q^2} \sum_{z=0}^{q-1}\sum_{x\in Q} \xi^{zx} \mu_z(\tau; v_1, v_2),\end{aligned}$$

where

$$\mu_z(\tau; v_1, v_2) = \sum_{j=0}^{m_{v_2}} \sum_{d\in G_{v_1}^{(\tau)}} \xi^{-dz} \sum_{a=0}^{q-1} \xi^{a(d+v_2+jN)}.$$

We only need to consider the case

$$d + v_2 + jN \equiv 0 \pmod q \ \text{ for } \ 0 \le j \le m_{v_2} \ \text{ and } \ d \in G_{v_1}^{(\tau)}, \tag{3.5}$$

since otherwise $\sum_{a=0}^{q-1} \xi^{a(d+v_2+jN)} = 0$ by (2.8). We want to determine the number of pairs (j, d) satisfying eq. (3.5). For $0 \le i < N^{\tau-1}$ we define

$$D_i = \{0 \le j \le m_{v_2} : -(v_2 + jN) \pmod q \in Z_{t_{v_1}+iN}\}.$$

All D_i's are disjoint with each other. We can show each $j \in D_i$ if and only if $d + v_2 + jN \equiv 0 \pmod q$ for some $d \in Z_{t_{v_1}+iN} \subset G_{v_1}^{(\tau)}$ and

$$\left\lceil \frac{q - v_2 - \lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor}{N} \right\rceil \le j \le \left\lfloor \frac{q - v_2 - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} \right\rfloor. \qquad (3.6)$$

In fact, $D = \bigcup_{i=0}^{N^{\tau-1}-1} D_i$ exactly contains all j satisfying eq. (3.5). We estimate the number of j, i.e. $|D|$. Since

$$|D_i| = \left\lfloor \frac{q - v_2 - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} \right\rfloor - \left\lceil \frac{q - v_2 - \lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor}{N} \right\rceil + 1,$$

for $0 \le i < N^{\tau-1}$, using the formulas

$$\frac{n_1}{n_2} - 1 < \left\lfloor \frac{n_1}{n_2} \right\rfloor \le \frac{n_1}{n_2} \quad \text{and} \quad \frac{n_1}{n_2} \le \left\lceil \frac{n_1}{n_2} \right\rceil < \frac{n_1}{n_2} + 1,$$

we can get the lower and upper bounds

$$\begin{aligned}
|D_i| &> \frac{q - v_2 - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} - 1 \\
&\quad - \left(\frac{q - v_2 - \lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor}{N} + 1 \right) + 1 \\
&= \frac{\lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} - 1 \\
&> \frac{1}{N} \left(\frac{(t_{v_1} + 1 + iN)q}{N^\tau} - 1 - \frac{(t_{v_1} + iN)q}{N^\tau} - 1 \right) - 1 \\
&= \frac{q}{N^{\tau+1}} - \frac{2}{N} - 1
\end{aligned}$$

and

$$\begin{aligned}
|D_i| &\leq \frac{q - v_2 - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} - \frac{q - v_2 - \lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor}{N} + 1 \\
&= \frac{\lfloor (t_{v_1} + 1 + iN)q/N^\tau \rfloor - \lceil (t_{v_1} + iN)q/N^\tau \rceil}{N} + 1 \\
&\leq \frac{(t_{v_1} + 1 + iN)q/N^\tau - (t_{v_1} + iN)q/N^\tau}{N} + 1 \\
&= \frac{q}{N^{\tau+1}} + 1.
\end{aligned}$$

Hence $|D| = \sum_{i=0}^{N^{\tau-1}-1} |D_i|$ is bounded by the following inequalities

$$\frac{q}{N^2} - 2N^{\tau-2} - N^{\tau-1} < |D| \leq \frac{q}{N^2} + N^{\tau-1}.$$

We first get a bound for $\mu_0(\tau; v_1, v_2)$. We have $\mu_0(\tau; v_1, v_2) = q|D|$, so

$$\frac{q^2}{N^2} - 2qN^{\tau-2} - qN^{\tau-1} < \mu_0(\tau; v_1, v_2) \leq \frac{q^2}{N^2} + qN^{\tau-1}.$$

Then we consider $\mu_z(\tau; v_1, v_2)$ for $z \neq 0$.

$$\begin{aligned}
|\mu_z(\tau; v_1, v_2)| &= \left| \sum_{j=0}^{m_{v_2}} \sum_{d \in G_{v_1}^{(\tau)}} \xi^{-dz} \sum_{a=0}^{q-1} \xi^{a(d+v_2+jN)} \right| \\
&= q \left| \sum_{j \in D} \xi^{(v_2+jN)z} \right| = q \left| \sum_{i=0}^{N^{\tau-1}-1} \sum_{j \in D_i} \xi^{(v_2+jN)z} \right| \\
&\leq q \sum_{i=0}^{N^{\tau-1}-1} \left| \sum_{j \in D_i} \xi^{jNz} \right| \quad \text{(we use (3.6))}.
\end{aligned}$$

Putting everything together, we get

$$\begin{aligned}\mu(\tau; v_1, v_2) &= \frac{1}{q^2}\sum_{z=0}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(\tau; v_1, v_2) \\ &= \frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) + \frac{1}{q^2}\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(\tau; v_1, v_2).\end{aligned}$$

We have

$$\begin{aligned}&\left|\mu(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| \\ = &\left|\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) + \frac{1}{q^2}\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| \\ \leq &\left|\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| + \left|\frac{1}{q^2}\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(\tau; v_1, v_2)\right|.\end{aligned}$$

Then we apply the bounds on $\mu_0(\tau; v_1, v_2)$ to get

$$\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2} > \frac{q-1}{2q^2}\left(\frac{q^2}{N^2} - 2qN^{\tau-2} - qN^{\tau-1}\right) - \frac{q-1}{2N^2}$$

and

$$\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2} \leq \frac{q-1}{2q^2}\left(\frac{q^2}{N^2} + qN^{\tau-1}\right) - \frac{q-1}{2N^2}.$$

That is,

$$\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2} > -\frac{(q-1)(2N^{\tau-2} + N^{\tau-1})}{2q}$$

and

$$\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2} \leq \frac{(q-1)N^{\tau-1}}{2q}.$$

Then we have

$$\begin{aligned}
&\left|\frac{q-1}{2q^2}\mu_0(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| < \\
&\max\left\{\left|-\frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q}\right|, \left|\frac{(q-1)N^{\tau-1}}{2q}\right|\right\},
\end{aligned}$$

and hence

$$\begin{aligned}
&\left|\mu(\tau; v_1, v_2) - \frac{q-1}{2N^2}\right| \\
< \;& \frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q} + \frac{1}{q^2}\left|\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(\tau; v_1, v_2)\right| \\
\leq \;& \frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q} + \frac{1}{q^2}\sum_{z=1}^{q-1}|\mu_z(\tau; v_1, v_2)| \cdot \left|\sum_{x\in Q}\xi^{zx}\right| \\
\leq \;& \frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q} + \frac{q^{1/2}+1}{2q^2}\sum_{z=1}^{q-1}|\mu_z(\tau; v_1, v_2)| \\
\leq \;& \frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q} + \frac{(q^{1/2}+1)N^{\tau-1}}{2q}\sum_{z=1}^{q-1}\left|\sum_{j\in D_i}\xi^{jNz}\right| \\
\leq \;& \frac{(q-1)(2N^{\tau-2}+N^{\tau-1})}{2q} \\
&+\frac{(q^{1/2}+1)N^{\tau-1}}{2q} \cdot \left(\frac{4}{\pi^2}q\log q + 0.38q + 0.608 + \frac{0.116}{q}\right),
\end{aligned}$$

which completes the proof. The last inequality follows from Lemma 2.5.3 and Lemma 2.5.4. □

Note here N^τ should be sufficiently smaller than q. Otherwise, the bound in the above theorem is trivial.

3.4 Distribution of (s_n, s_{n+1}, s_{n+2})

The idea of previous sections can help us to investigate the number of occurrences of three consecutive symbols in the half-ℓ-sequence $\mathbf{s}$. According to the definition of $G_v^{(\omega)}$, the number of occurrences of three consecutive symbols (s_n, s_{n+1}, s_{n+2}) with $s_n = v_1$, $s_{n+1} = v_2$, and $s_{n+2} = v_3$, denoted by $\mu(1, 2; v_1, v_2, v_3)$, can be obtained by

$$\mu(1, 2; v_1, v_2, v_3) = |Q \cap G_{v_3}^{(0)} \cap G_{v_2}^{(1)} \cap G_{v_1}^{(2)}|.$$

Theorem 3.4.1 *For an N-ary half-ℓ-sequence $\mathbf{s}$ with prime connection integer q, the number $\mu(1, 2; v_1, v_2, v_3)$ of occurrences of (s_n, s_{n+1}, s_{n+2}) with $s_n = v_1$, $s_{n+1} = v_2$, and $s_{n+2} = v_3$ for $0 \le n < T$ satisfies*

$$\begin{aligned}\left|\mu(1, 2; v_1, v_2, v_3) - \frac{q-1}{2N^3}\right| \;<\; & \frac{(q-1)}{2q}\left(\frac{4}{N} + 3\right) \\ & + \left(q^{1/2} + 1\right)\left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right),\end{aligned}$$

where $0 \le v_1, v_2, v_3 < N$.

Proof. Using Fourier transforms, we have

$$
\begin{aligned}
& \mu(1,2;v_1,v_2,v_3) \\
= & \sum_{x\in Q} g_{v_3}^{(0)}(x) g_{v_2}^{(1)}(x) g_{v_1}^{(2)}(x) \\
= & \sum_{x\in Q} \sum_{a=0}^{q-1} \hat{g}_{v_3}^{(0)}(a)\xi^{ax} \sum_{b=0}^{q-1} \hat{g}_{v_2}^{(\tau)}(b)\xi^{bx} \sum_{c=0}^{q-1} \hat{g}_{v_1}^{(2)}(c)\xi^{cx} \\
= & \sum_{a=0}^{q-1} \hat{g}_{v_3}^{(0)}(a) \sum_{b=0}^{q-1} \hat{g}_{v_2}^{(1)}(b) \sum_{z=0}^{q-1} \hat{g}_{v_1}^{(2)}(z-a-b) \sum_{x\in Q} \xi^{zx} \quad (\text{we use } z=a+b+c) \\
= & \sum_{a=0}^{q-1} \frac{1}{q} \sum_{d\in G_{v_3}^{(0)}} \xi^{-ad} \sum_{b=0}^{q-1} \frac{1}{q} \sum_{e\in G_{v_2}^{(1)}} \xi^{-be} \sum_{z=0}^{q-1} \frac{1}{q} \sum_{f\in G_{v_1}^{(2)}} \xi^{-f(z-a-b)} \sum_{x\in Q} \xi^{zx} \\
= & \frac{1}{q^3} \sum_{z=0}^{q-1} \sum_{x\in Q} \xi^{zx} \sum_{a=0}^{q-1} \sum_{j=0}^{m_{v_3}} \xi^{a(v_3+jN)} \sum_{b=0}^{q-1} \sum_{e\in G_{v_2}^{(1)}} \xi^{-be} \sum_{f\in G_{v_1}^{(2)}} \xi^{-f(z-a-b)} \\
= & \frac{1}{q^3} \sum_{z=0}^{q-1} \sum_{x\in Q} \xi^{zx} \sum_{j=0}^{m_{v_3}} \sum_{e\in G_{v_2}^{(1)}} \sum_{f\in G_{v_1}^{(2)}} \xi^{-fz} \sum_{b=0}^{q-1} \xi^{b(f-e)} \sum_{a=0}^{q-1} \xi^{a(v_3+jN+f)} \\
\triangleq & \frac{1}{q^3} \sum_{z=0}^{q-1} \sum_{x\in Q} \xi^{zx} \mu_z(1,2;v_1,v_2,v_3),
\end{aligned}
$$

where

$$
\mu_z(1,2;v_1,v_2,v_3) = \sum_{j=0}^{m_{v_3}} \sum_{e\in G_{v_2}^{(1)}} \sum_{f\in G_{v_1}^{(2)}} \xi^{-fz} \sum_{b=0}^{q-1} \xi^{b(f-e)} \sum_{a=0}^{q-1} \xi^{a(v_3+jN+f)}.
$$

We only need to consider the case

$$
v_3 + jN + f \equiv 0 \pmod{q} \text{ and } f - e \equiv 0 \pmod{q}, \tag{3.7}
$$

for $0 \le j \le m_{v_3}$, $e \in G^{(1)}_{v_2}$, and $f \in G^{(2)}_{v_1}$, since otherwise by eq. (2.8) we have

$$\sum_{b=0}^{q-1} \xi^{b(f-e)} \sum_{a=0}^{q-1} \xi^{a(v_3+jN+f)} = 0.$$

We want to determine the number of pairs (j, f) satisfying eq. (3.7). From eq. (3.3) with $\omega = 1$, since $-q < -(v_3 + jN) \le 0$ such j must satisfy both

$$\left\lceil \frac{t_{v_2} q}{N} \right\rceil \le q - (v_3 + jN) \le \left\lfloor \frac{(t_{v_2} + 1)q}{N} \right\rfloor$$

and

$$\left\lceil \frac{(t_{v_1} + iN)q}{N^2} \right\rceil \le q - (v_3 + jN) \le \left\lfloor \frac{(t_{v_1} + 1 + iN)q}{N^2} \right\rfloor, \quad 0 \le i < N.$$

We derive that j is an element of

$$T = \left\{ j : \left\lceil \frac{q - v_3 - \lfloor (t_{v_2} + 1)q/N \rfloor}{N} \right\rceil \le j \le \left\lfloor \frac{q - v_3 - \lceil t_{v_2} q/N \rceil}{N} \right\rfloor \right\}$$

and

$$E_i = \left\{ j : \left\lceil \frac{q - v_3 - \lfloor (t_{v_1} + 1 + iN)q/N^2 \rfloor}{N} \right\rceil \le j \le \left\lfloor \frac{q - v_3 - \lceil (t_{v_1} + iN)q/N^2 \rceil}{N} \right\rfloor \right\},$$

for $0 \le i < N$. We find that $q/N^2 - 2/N - 1 < |T| \le q/N^2 + 1$ and $q/N^3 - 2/N - 1 < |E_i| \le q/N^3 + 1$. On the other hand, we get that $|T|$ is smaller than the right bound of E_i minus the left bound of E_{i+1}, but larger than the right bound of E_i minus the right bound of E_{i+1}. So we conclude

$$D = [1, m_{v_3}] \bigcap T \bigcap (\bigcup_{i=0}^{N-1} E_i)$$

contains at most two disjoint intervals, say A and B, of integers. In fact, $D = A \cup B$

contains all j satisfying eq. (3.7). We get

$$\frac{q}{N^3} - \frac{4}{N} - 3 < |A| + |B| \leq \frac{q}{N^3} + 2.$$

Hence we can get

$$\frac{q}{N^3} - \frac{4}{N} - 3 < |D| \leq \frac{q}{N^3} + 2.$$

We first get a bound for $\mu_0(1, 2; v_1, v_2, v_3)$. We have

$$\mu_0(1, 2; v_1, v_2, v_3) \;\; = \;\; q^2|D|$$

and

$$q^2\left(\frac{q}{N^3} - \frac{4}{N} - 3\right) < \mu_0(1, 2; v_1, v_2, v_3) \leq q^2\left(\frac{q}{N^3} + 2\right).$$

Then we consider $\mu_z(1, 2; v_1, v_2, v_3)$ for $z \neq 0$.

$$\begin{aligned} |\mu_z(\tau; v_1, v_2)| &= \left| \sum_{j=0}^{m_{v_3}} \sum_{e \in G_{v_2}^{(1)}} \sum_{f \in G_{v_1}^{(2)}} \xi^{-fz} \sum_{b=0}^{q-1} \xi^{b(f-e)} \sum_{a=0}^{q-1} \xi^{a(v_3+jN+f)} \right| \\ &= q^2 \left| \sum_{j \in D} \xi^{(v_3+jN)z} \right| = q^2 \left| \sum_{j \in A \cup B} \xi^{jNz} \right|. \end{aligned}$$

Putting everything together, we get

$$\begin{aligned} & \mu(1, 2; v_1, v_2, v_3) \\ = \;& \frac{1}{q^3} \sum_{z=0}^{q-1} \sum_{x \in Q} \xi^{zx} \mu_z(1, 2; v_1, v_2, v_3) \\ = \;& \frac{q-1}{2q^3} \mu_0(1, 2; v_1, v_2, v_3) + \frac{1}{q^3} \sum_{z=1}^{q-1} \sum_{x \in Q} \xi^{zx} \mu_z(1, 2; v_1, v_2, v_3). \end{aligned}$$

We have

$$\begin{aligned} & \left|\mu(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\right| \\ = & \left|\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)+\frac{1}{q^3}\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\right| \\ \leq & \left|\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\right|+\left|\frac{1}{q^3}\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(1,2;v_1,v_2,v_3)\right|. \end{aligned}$$

Then we apply the bounds on $\mu_0(\tau;v_1,v_2)$ to get

$$\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}>\frac{q-1}{2q^3}q^2\left(\frac{q}{N^3}-\frac{4}{N}-3\right)-\frac{q-1}{2N^3}$$

and

$$\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\leq\frac{q-1}{2q^3}q^2\left(\frac{q}{N^3}+2\right)-\frac{q-1}{2N^3}.$$

That is,

$$\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3} \;>\; -\frac{(q-1)}{2q}\left(\frac{4}{N}+3\right)$$

and

$$\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\leq\frac{q-1}{q}.$$

Then we have

$$\begin{aligned} & \left|\frac{q-1}{2q^3}\mu_0(1,2;v_1,v_2,v_3)-\frac{q-1}{2N^3}\right|< \\ & \max\left\{\left|-\frac{(q-1)}{2q}\left(\frac{4}{N}+3\right)\right|,\left|\frac{q-1}{q}\right|\right\}, \end{aligned}$$

and hence

$$\begin{aligned}
& \left|\mu(1,2;v_1,v_2,v_3) - \frac{q-1}{2N^3}\right| \\
< \; & \frac{(q-1)}{2q}\left(\frac{4}{N}+3\right) + \frac{1}{q^3}\left|\sum_{z=1}^{q-1}\sum_{x\in Q}\xi^{zx}\mu_z(1,2;v_1,v_2,v_3)\right| \\
\leq \; & \frac{(q-1)}{2q}\left(\frac{4}{N}+3\right) + \frac{1}{q^3}\sum_{z=1}^{q-1}|\mu_z(1,2;v_1,v_2,v_3)|\cdot\left|\sum_{x\in Q}\xi^{zx}\right| \\
\leq \; & \frac{(q-1)}{2q}\left(\frac{4}{N}+3\right) + \frac{q^{1/2}+1}{2q^3}\sum_{z=1}^{q-1}|\mu_z(1,2;v_1,v_2,v_3)| \\
\leq \; & \frac{(q-1)}{2q}\left(\frac{4}{N}+3\right) + \frac{(q^{1/2}+1)}{2q}\sum_{z=1}^{q-1}\left|\sum_{j\in A\cup B}\xi^{jNz}\right| \\
\leq \; & \frac{(q-1)}{2q}\left(\frac{4}{N}+3\right) \\
& + \frac{(q^{1/2}+1)}{2q}\cdot 2\left(\frac{4}{\pi^2}q\log q + 0.38q + 0.608 + \frac{0.116}{q}\right),
\end{aligned}$$

which completes the proof. The last inequality follows from Lemma 2.5.3 and Lemma 2.5.4. □

3.5 A Sharper Bound When $N = 2$

Theorem 3.5.1 *Let* $\mathbf{a} = a_0, a_1, a_2, \ldots$ *be a binary half-ℓ-sequence with* $q \equiv 1 \mod 8$ *and* q *an odd prime. Then* $\mathbf{a}$ *is balanced.*

Proof. Since $q \equiv 1 \mod 8$ and q is an odd prime, the order of 2 is $(q-1)/2$. Then we have $2^{(q-1)/2} \equiv 1 \mod q$. As a result, $2^{(q-1)/4} \equiv \pm 1 \mod q$. Because the order of 2 is $(q-1)/2$, $2^{(q-1)/4} \not\equiv 1 \mod q$, we have $2^{(q-1)/4} \equiv -1 \mod q$.

There is an integer h so that for all $j \geq 0$ we have

$$a_j \equiv 2^{-j}h \mod q \mod 2.$$

Consider a_j where $j \in [0, (q-1)/2)$. Then we have

$$a_j \equiv 2^{-j}h \mod q \mod 2.$$

and

$$\begin{aligned} a_{\frac{q-1}{4}+j} &\equiv 2^{-(\frac{q-1}{4}+j)}h \mod q \mod 2 \\ &\equiv \left(2^{-\frac{q-1}{4}}2^{-j}\right)h \mod q \mod 2 \\ &\equiv -2^{-j}h \mod q \mod 2 \\ &\equiv \left(q - 2^{-j}h\right) \mod q \mod 2, \end{aligned}$$

which is the complementary bit to a_j. So the first half of half-ℓ-sequence $\mathbf{a}$ is the bit-wise complement of its second half. Then the numbers of 1's and 0's in $\mathbf{a}$ are equal. So $\mathbf{a}$ is balanced.

□

3.6 Experimental Results

In this section we analyze the imbalance properties of half-ℓ-sequences by showing how tight the bounds in Theorem 3.2.1, 3.3.1 and 3.4.1 are using experiment results. It is impractical to investigate half-ℓ-sequences with big qs and $N = 2^{32}$ or 2^{16} as discussed in Lee and Park's paper by experiments. As a result, we choose half-ℓ-sequences with smaller Ns and qs for investigation.

3.6.1 Finding Satisfactory Connection Integers

To investigate the imbalance properties of half-ℓ-sequences, we need to find satisfactory connection integers for FCSRs to generate half-ℓ-sequences. The condition

on connection integers to generate an N-ary half-ℓ-sequences is similar to that of an N-ary ℓ-sequences, both of which relate to the order of N modulo q. Let's first look at the the condition on the connection integer to generate an N-ary ℓ-sequence. Let q be a prime number. The condition on the connection integer q for the generation of an N-ary ℓ-sequence is that the order of N modulo q is $q-1$. In other words, N is primitive modulo q.

Lemma 3.6.1 *Let p, q be prime numbers with $q = 2p+1$. N is primitive modulo q if and only if $N^p \not\equiv 1 \pmod{q}$ and $N^2 \not\equiv 1 \pmod{q}$.*

Indeed, since $ord_q(N) | \varphi(q) = q - 1 = 2p$.

Similarly, we need $ord_q(N) = (q-1)/2$ for an N-ary half-ℓ-sequence according to the definition of N-ary half-ℓ-sequence 3.1.1. That is, $N^{(q-1)/2} \equiv 1 \pmod{q}$ or the Legendre symbol $\left(\frac{N}{q}\right) = 1$. In the experiment, we generated some connection integers q of the form

- $q = 2p + 1$ with p and q prime; and
- $q \equiv -1 \pmod{N}$.

The sequences generated by an FCSR with those connection integers are half-ℓ-sequences or ℓ-sequences. Indeed, it is possible that either $N^{(q-1)/2} \equiv 1 \pmod{q}$ or $N^{q-1} \equiv 1 \pmod{q}$. Note that when $N \geq 2^3$, $q \equiv -1 \pmod{8}$. Then $\left(\frac{2}{q}\right) = 1$. Therefore,

$$\left(\frac{N}{q}\right) = \left(\frac{2^k}{q}\right) = \left(\left(\frac{2}{q}\right)\right)^k = 1,$$

or $N^{(q-1)/2} \equiv 1 \pmod{q}$. Since $p = (q-1)/2$ is a prime, $ord_q(N) = (q-1)/2$. Hence the sequences generated in this way are all half-ℓ-sequences. These may not be the only half-ℓ-sequences, but they are just the easiest to find. We use the sieve of Eratosthenes algorithm [9] to find all the primes numbers satisfying the conditions

above. The sieve of Eratosthenes algorithm finds all prime numbers up to a given limit. There are two ways to generate the half-ℓ-sequence based on a specific q. One uses eq. (2.2) with initial states and carry. The other one uses the algebraic expression in eq. (2.6). In our experiments, we used the algebraic expression to generate the half-ℓ-sequences.

3.6.2 One Symbol Case

In this section we analyze the imbalance properties of half-ℓ-sequences by showing how tight the bounds are in the one symbol case.

Let

$$\mu^{'}(v) = \left|\mu(v) - \frac{q-1}{2N}\right|,\ 0 \leq v < N,$$

$$\sigma_v(q) = \frac{q-1}{2q} + \frac{(q^{1/2}+1)}{2} \cdot \left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right)$$

and

$$\delta_v(q) = \frac{\max\{\mu^{'}(v) : 0 \leq v < N\}}{\sigma_v(q)}.$$

The quantity $\mu^{'}(v)$ is the difference between the number of occurrences of v in one period of a half-ℓ-sequence and the average number of occurrences. The smaller $\delta_v(q)$ is, the more balanced the sequence is. Ideally, for a pseudo-random sequence, we would like $\delta_v(q)$ to be close to zero. We generated the sequences for corresponding q and calculated $\delta_v(q)$ for these qs. We would like to see how $\delta_v(q)$ changes as the connection integers increase for a particular N. We have done experiments for $N = 8, 16$ and 32. For each value of N, we generated $\delta_v(q)$ with FCSR sizes 2, 3 and 4. Note that if the size of an FCSR is m, then the corresponding connection integer $q \in (N^m, N^{m+1})$. Fig. 3.1 shows that $\delta_v(q)$ for $N = 8$ with FCSR size 2, 3 and 4 is in the range $(0.05, 0.35)$. Fig. 3.2 shows that $\delta_v(q)$ for $N = 16$ with FCSR size 2, 3 and 4 is in the range $(0.04, 0.3)$. Fig. 3.3 shows that $\delta_v(q)$ for $N = 32$ with FCSR size

2, 3 and 4 is in the range $(0.03, 0.25)$. We don't see an increase or decrease pattern as q increases from the three figures. We don't know a way to find the best qs if the period is large enough to be useful. $\mu'(v)$ is the product of $\delta_v(q)$ and $\sigma_v(q)$. As the connection integer q increases, $\sigma_v(q)$ will increase accordingly. Thus $\mu'(v)$ will increase accordingly.

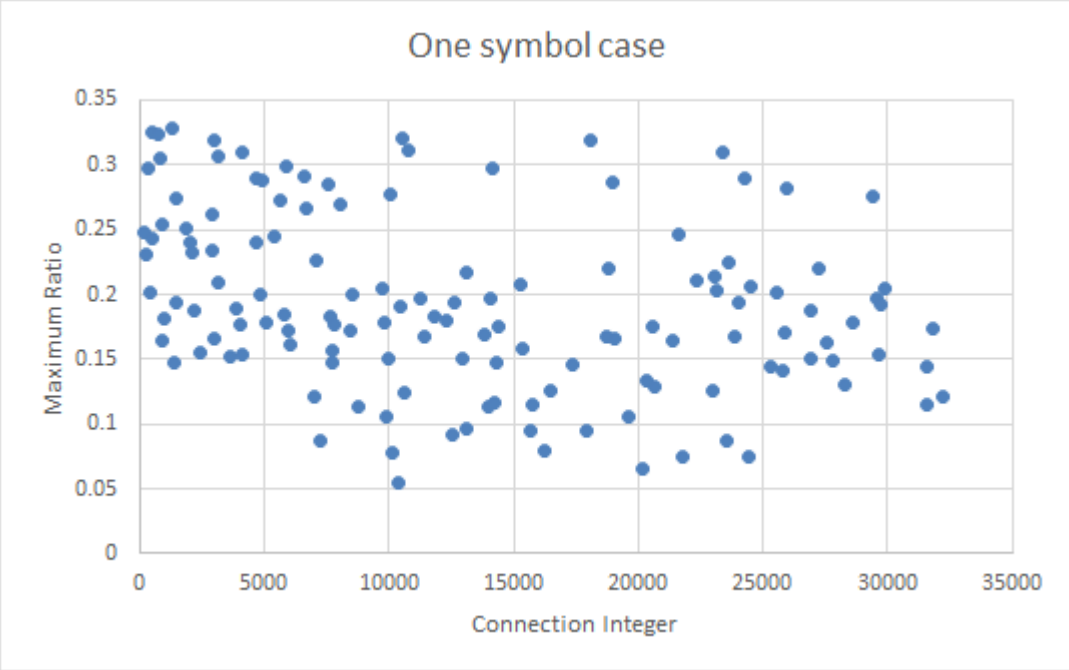

Figure 3.1: Maximum ratio when $N = 8$ in one symbol case

3.6.3 Two Consecutive Symbol Case

In this section we analyze the imbalance properties of half-ℓ-sequences by showing how tight the bounds are in the two consecutive symbol case. This is a special case when $\tau = 1$ in Theorem 3.3.1.

Let

$$\mu'(v_1, v_2) = \left| \mu(v_1, v_2) - \frac{q-1}{2N^2} \right|, \ 0 \le v_1, v_2 < N,$$

$$\sigma_{v_1,v_2}(q) = \frac{(q-1)}{2q}\left(\frac{2}{N}+1\right) + \frac{(q^{1/2}+1)}{2}\left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right)$$

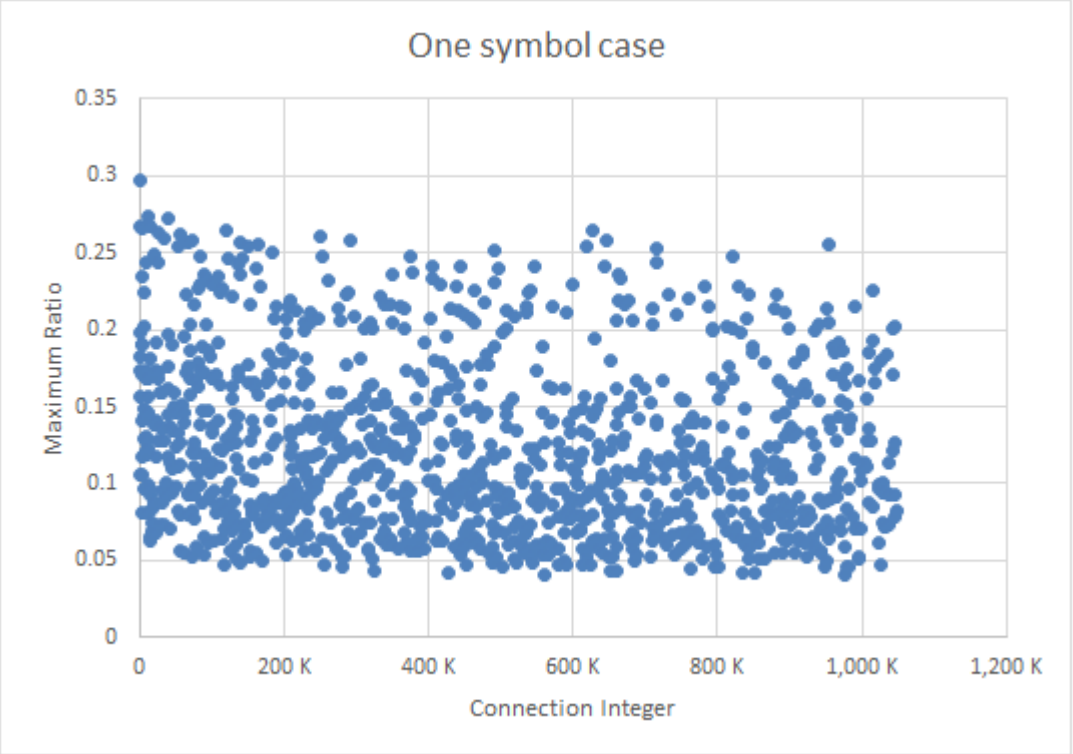

Figure 3.2: Maximum ratio when $N = 16$ in one symbol case

and

$$\delta_{v_1,v_2}(q) = \frac{\max\{\mu^{'}(v_1, v_2) : 0 \leq v_1, v_2 < N\}}{\sigma_{v_1,v_2}(q)}.$$

The quantity $\mu^{'}(v_1, v_2)$ is the difference between the number of occurrences of v_1, v_2 in one period of a half-ℓ-sequence and the average number of occurrences. The smaller $\delta_{v_1,v_2}(q)$ is, the more balanced the sequence is. Ideally, for a pseudorandom sequence, we would like $\delta_{v_1,v_2}(q)$ to be close to zero. We generated the sequences for corresponding q and calculated $\delta_{v_1,v_2}(q)$ for these qs. We would like to see how $\delta_{v_1,v_2}(q)$ changes as the connection integers increase for a particular N. We have done experiments for $N = 8, 16$ and 32. For each value of N, we generated $\delta_{v_1,v_2}(q)$ with FCSR sizes 2, 3 and 4. Note that if the size of an FCSR is m, then the corresponding connection integer $q \in (N^m, N^{m+1})$. Fig. 3.4 shows that $\delta_{v_1,v_2}(q)$ for $N = 8$ with FCSR size 2, 3 and 4 is in the range $(0.03, 0.17)$. Fig. 3.5 shows that $\delta_{v_1,v_2}(q)$ for $N = 16$ with FCSR size 2, 3 and 4 is in the range $(0.02, 0.12)$. Fig. 3.6 shows that

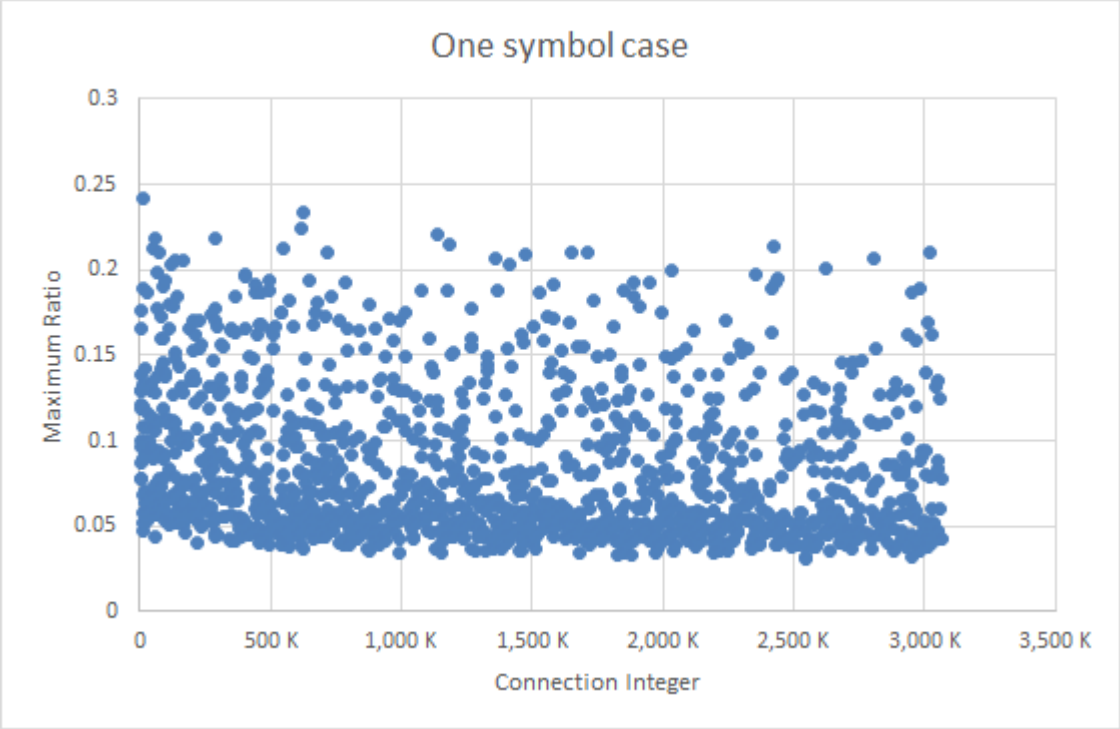

Figure 3.3: Maximum ratio when $N = 32$ in one symbol case

$\delta_{v_1,v_2}(q)$ for $N = 32$ with FCSR size 2, 3 and 4 is in the range $(0.01, 0.06)$. We don't see an increase or decrease pattern as q increases from the three figures. We don't know a way to find the best qs if the period is large enough to be useful. $\mu^{'}(v_1, v_2)$ is the product of $\delta_{v_1,v_2}(q)$ and $\sigma_{v_1,v_2}(q)$. As the connection integer q increases, $\sigma_{v_1,v_2}(q)$ will increase accordingly. Thus $\mu^{'}(v_1, v_2)$ will increase accordingly.

3.6.4 Three Consecutive Symbol Case

In this section we analyze the imbalance properties of half-ℓ-sequences by showing how tight the bounds are in the three consecutive symbol case.

Let

$$\mu^{'}(v_1, v_2, v_3) = \left|\mu(v_1, v_2, v_3) - \frac{q-1}{2N^3}\right|, \ 0 \le v_1, v_2, v_3 < N,$$

$$\sigma_{v_1,v_2,v_3}(q) = \frac{(q-1)}{2q}\left(\frac{4}{N} + 3\right) + \left(q^{1/2} + 1\right)\left(\frac{4}{\pi^2}\log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2}\right)$$

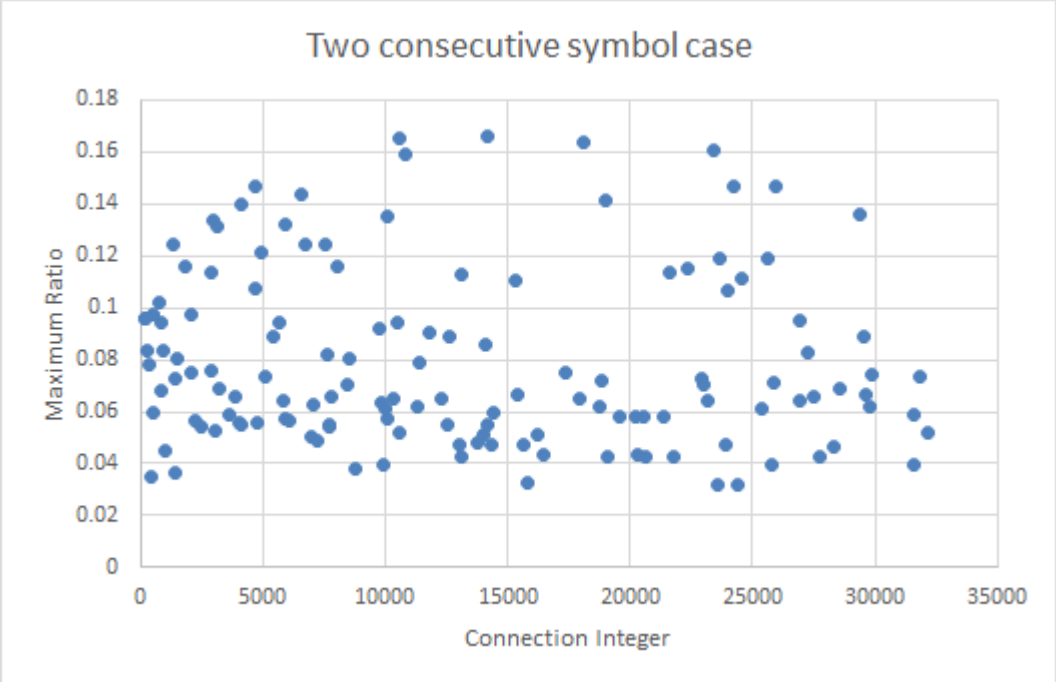

Figure 3.4: Maximum ratio when $N = 8$ in two consecutive symbol case

and

$$\delta_{v_1,v_2,v_3}(q) = \frac{\max\{\mu^{'}(v_1, v_2, v_3) : 0 \leq v_1, v_2, v_3 < N\}}{\sigma_{v_1,v_2,v_3}(q)}.$$

The quantity $\mu^{'}(v_1, v_2, v_3)$ is the difference between the number of occurrences of v in one period of a half-ℓ-sequence and the average number of occurrences. The smaller $\delta_{v_1,v_2,v_3}(q)$ is, the more balanced the sequence is. Ideally, for a pseudo-random sequence, we would like $\delta_{v_1,v_2,v_3}(q)$ to be close to zero. We generated the sequences for corresponding q and calculated $\delta_{v_1,v_2,v_3}(q)$ for these qs. We would like to see how $\delta_{v_1,v_2,v_3}(q)$ changes as the connection integers increase for a particular N. We have done experiments for $N = 8, 16$ and 32. For each value of N, we generated $\delta_{v_1,v_2,v_3}(q)$ with FCSR sizes 2, 3 and 4. Note that if the size of an FCSR is m, then the corresponding connection integer $q \in (N^m, N^{m+1})$. Fig. 3.7 shows that $\delta_{v_1,v_2,v_3}(q)$ for $N = 8$ with FCSR size 2, 3 and 4 is in the range $(0.008, 0.05)$. Fig. 3.8 shows that $\delta_{v_1,v_2,v_3}(q)$ for $N = 16$ with FCSR size 2, 3 and 4 is in the range $(0.002, 0.17)$. Fig. 3.9 shows that $\delta_{v_1,v_2,v_3}(q)$ for $N = 32$ with FCSR size 2, 3 and 4 is in the range

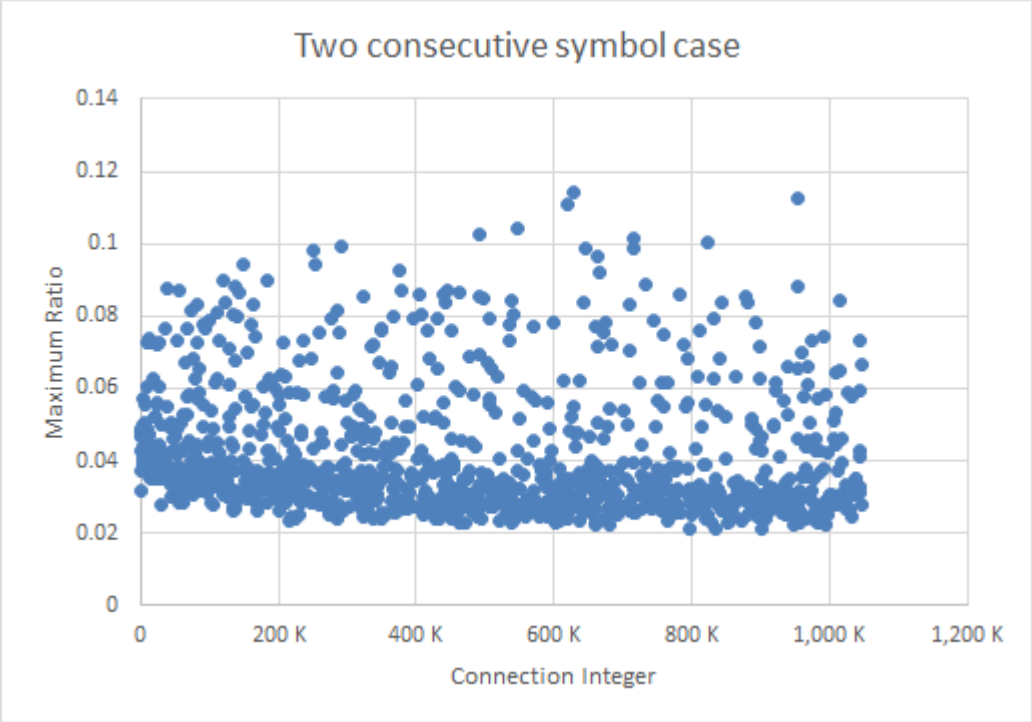

Figure 3.5: Maximum ratio when $N = 16$ in two consecutive symbol case

$(0.0006, 0.0075)$. We don't see an increase or decrease pattern as q increases from the three figures. We don't know a way to find the best qs if the period is large enough to be useful. $\mu'(v_1, v_2, v_3)$ is the product of $\delta_{v_1,v_2,v_3}(q)$ and $\sigma_{v_1,v_2,v_3}(q)$. As the connection integer q increases, $\sigma_{v_1,v_2,v_3}(q)$ will increase accordingly. Thus $\mu'(v_1, v_2, v_3)$ will increase accordingly.

3.7 Concluding Remarks

In this chapter, we show some nice features of the distribution of s_n, $(s_n, s_{n+\tau})$ and (s_n, s_{n+1}, s_{n+2}) in one period of a half-ℓ-sequence. We discuss a special binary case half-ℓ-sequence which is balanced. Our methods for investigating the distribution properties of half-ℓ-sequences can be extended to investigation of distribution properties of sequences whose period is $\phi(q)/2^k$ with $k \geq 2$, e.g., quarter-ℓ-sequence when $k = 2$. We can get similar bounds but with larger constants due to the polynomial degree increase in Weil's theorem. The implementation efficiency for these sequences

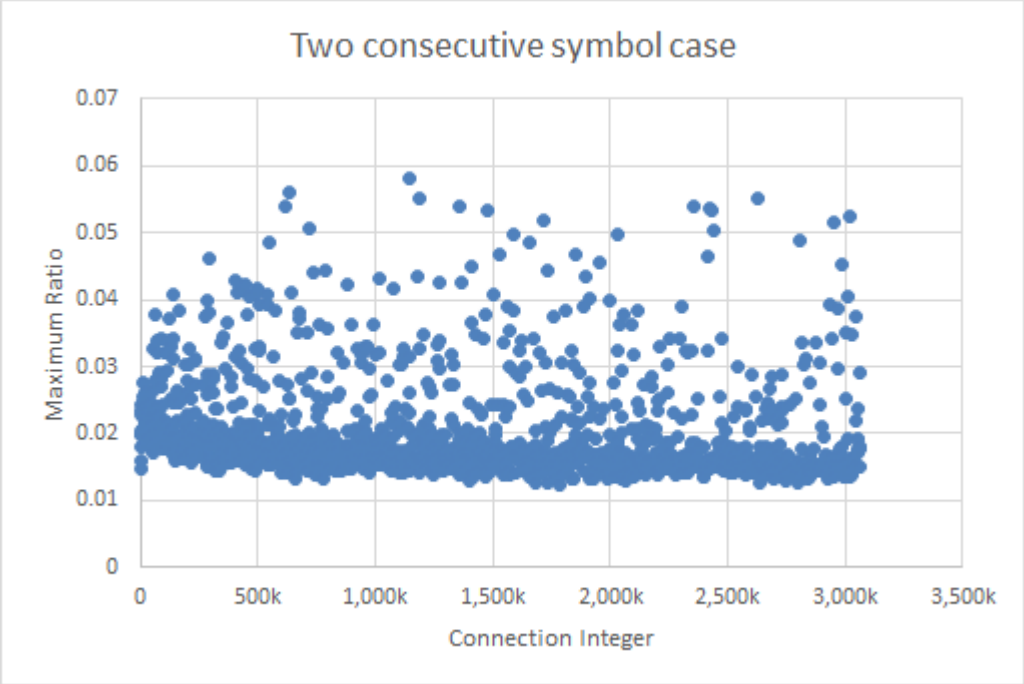

Figure 3.6: Maximum ratio when $N = 32$ in two consecutive symbol case

might also decrease, since the size of the FCSR must increase to achieve the same period.

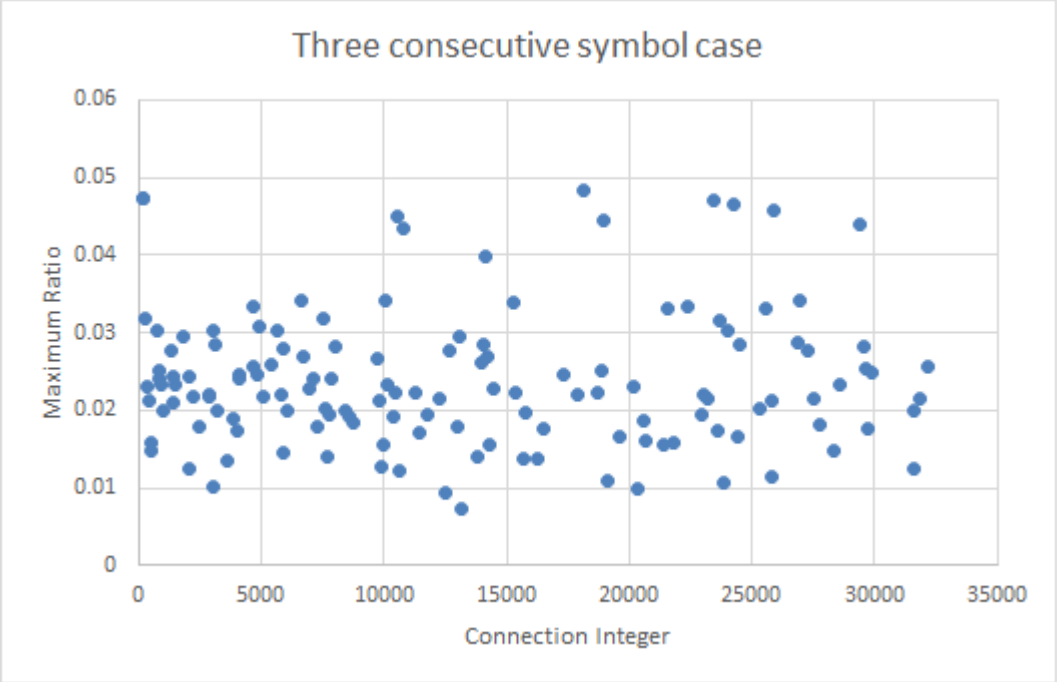

Figure 3.7: Maximum ration when $N = 8$ in three consecutive symbol case

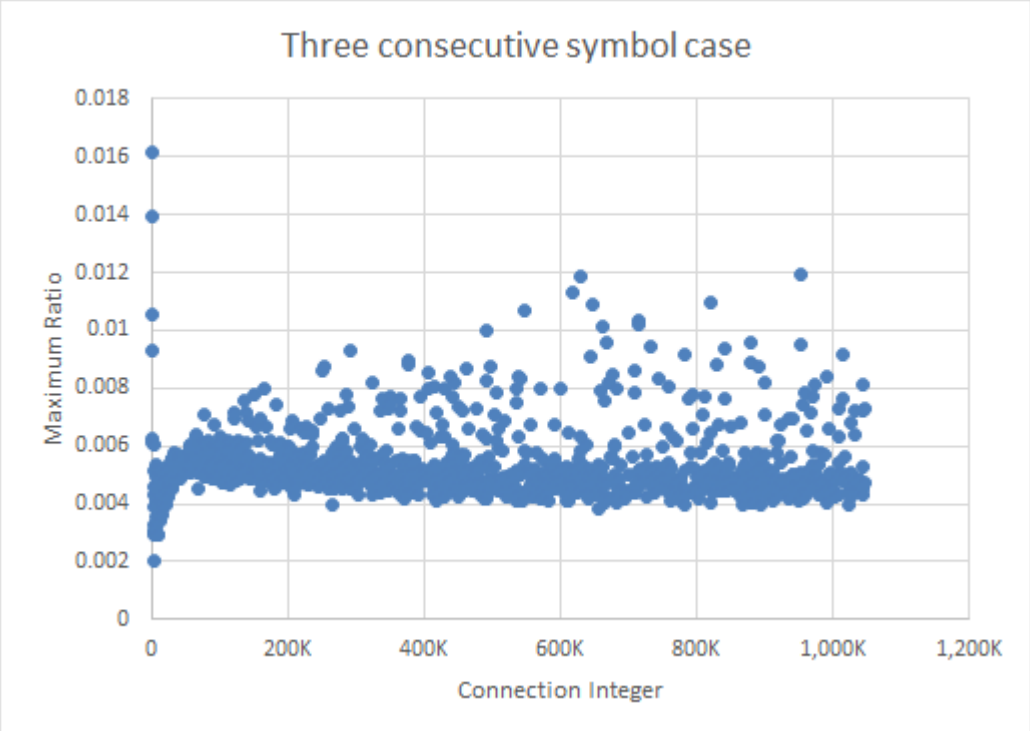

Figure 3.8: Maximum ration when $N = 16$ in three consecutive symbol case

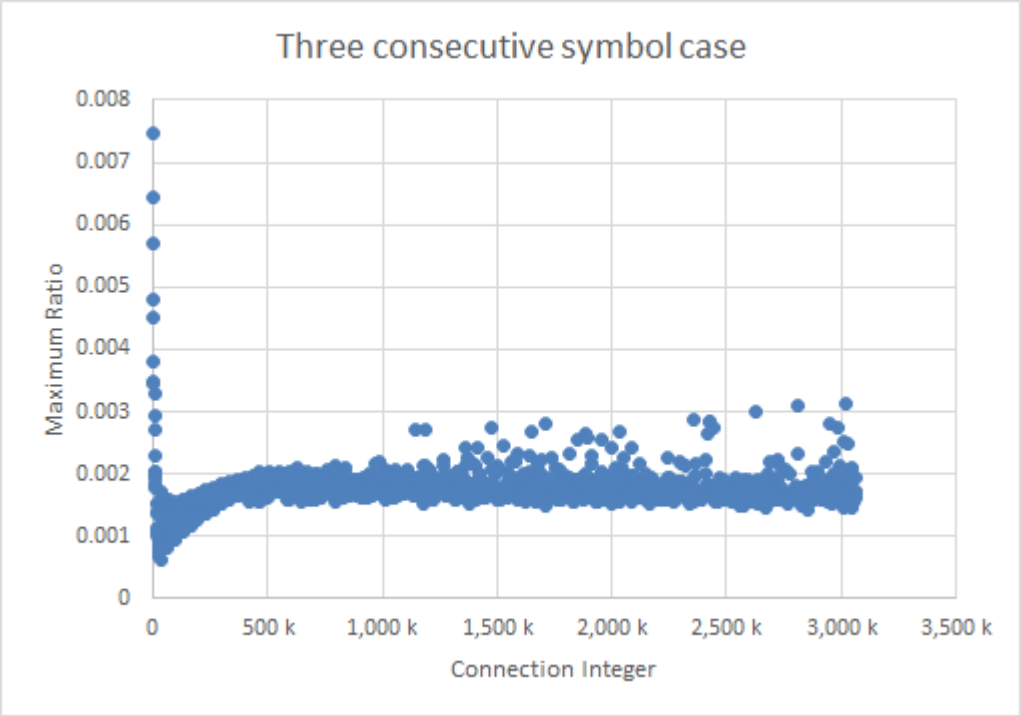

Figure 3.9: Maximum ratio when $N = 32$ in three consecutive symbol case

Chapter 4 Statistical Properties of Pseudorandom Sequences

4.1 Introduction

In this chapter, we investigate various statistical properties of sequences related to FCSRs. First, we introduce the distribution properties of N-ary sequences that combine two half-ℓ-sequences using addition modulo N. We give bounds on the number of occurrences of two symbols with a fixed distance between them in an ℓ-sequence and obtain conditions on the connection integer that guarantee the distribution is highly uniform. Furthermore, we discuss the autocorrelations of half-ℓ-sequences.

4.2 Distribution Properties of Combined Half-ℓ-sequences

Summation combiners [56] are stream ciphers that combine two or more binary m-sequences using addition-with-carry operations. They have been studied during the 1980's due to their speed and simple construction in hardware implementation. The period of the resulting combined sequence reaches approximately the product of the periods of the constituent sequences. The linear span of the resulting sequence was known to approach its period. However, the 2-adic complexity of the resulting sequence is no more than the sum of the 2-adic complexities of the constituent sequences. In 2006, Goresky and Klapper considered the case of combining two binary FCSR sequences using binary addition [29]. In particular, they considered combining two distinct ℓ-sequences using addition modulo two. Indeed, even though ℓ-sequences have good distribution properties, linear complexities and correlation properties, they themselves should never be used as keystreams because of the existence of the 2-adic rational approximation algorithm [30]. In this section we consider a similar problem by combining two distinct half-ℓ-sequences. We investigate the period and shifted

properties of combined half-ℓ-sequences. Distribution properties of combined half-ℓ-sequences are presented. At the end of this section we show some experimental results about combined half-ℓ-sequences.

4.2.1 Period and Shift Properties

Goresky and Klapper investigated the period of sequences that combine two binary FCSR sequences using binary addition [29]. Let $\mathbb{Z}/(N)$ be the residue ring modulo N. For all $a, b, c \in \mathbb{Z}/(N)$, if $a + b \pmod N = a + c \pmod N$, then $b \pmod N = c \pmod N$. It is straightforward to derive the following lemma.

Lemma 4.2.1 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *be a periodic FCSR sequence of (minimal) period* T_1 *with each* $a_i \in \mathbb{Z}/(N)$*, and let* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be a periodic FCSR sequence of (minimal) period* T_2 *with each* $b_i \in \mathbb{Z}/(N)$*. Let* $\mathbf{c} = \{c_i\}_{i=0}^{\infty}$ *be a sequence with* $c_i = a_i + b_i \pmod N$ *for each* i*. Suppose that for every prime* p*, the largest power of* p *that divides* T_1 *is not equal to the largest power of* p *that divides* T_2*. Then* $\mathbf{c}$ *is periodic and the period of* $\mathbf{c}$ *is* $lcm(T_1, T_2)$*, the least common multiple of* T_1 *and* T_2*.*

We consider combining two distinct N-ary half-ℓ-sequences $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ and $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ using addition modulo N to obtain a sequence $\mathbf{c} = \{c_i\}_{i=0}^{\infty}$ where $a_i, b_i, c_i \in \mathbf{Z}/(N)$. Suppose $\mathbf{a}$ is a half-ℓ-sequence that is generated by an FCSR with connection integer q_1 and that $\mathbf{b}$ is a half-ℓ-sequence that is generated by an FCSR with connection integer q_2. Then the period T_1 of sequence $\mathbf{a}$ is $(q_1 - 1)/2$ and the period T_2 of sequence $\mathbf{b}$ is $(q_2 - 1)/2$. We can easily gain the period of the sequence that combines two N-ary half-ℓ-sequences using modular addition.

Theorem 4.2.1 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be* N*-ary half-*ℓ*-sequences with connection integers* q_1 *and* q_2 *respectively. Suppose that for every prime* p*, the largest power of* p *that divides* $(q_1 - 1)/2$ *is not equal to the largest power of* p *that divides*

$(q_2-1)/2$. *Then the sequence* $\mathbf{c} = \{c_i\}_{i=0}^{\infty}$ *obtained by* $c_i = a_i + b_i \pmod N$ *has period*

$$\frac{(q_1-1)(q_2-1)}{2\gcd(q_1-1, q_2-1)}.$$

Proof. According to Lemma 4.2.1, the period of $\mathbf{c}$ is the least common multiple of $(q_1-1)/2$ and $(q_2-1)/2$, which is

$$\frac{(q_1-1)(q_2-1)}{4\gcd((q_1-1)/2, (q_2-1)/2)} = \frac{(q_1-1)(q_2-1)}{2\gcd(q_1-1, q_2-1)}.$$

□

Lemma 4.2.2 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be binary half-*ℓ*-sequences with connection integers* q_1 *and* q_2 *respectively. Suppose that one of the connection integers is congruent to* 1 (mod 8). *Without loss of generality, let* $q_1 \equiv 1 \pmod 8$. *Let* $\mathbf{c} = \{c_i\}_{i=0}^{\infty}$ *be a sequence with* $c_i = a_i + b_i \pmod N$. *Suppose that* $\frac{q_2-1}{\gcd(q_1-1,q_2-1)}$ *is odd and* $\frac{q_1-1}{\gcd(q_1-1,q_2-1)}$ *is even. Then the second half of* $\mathbf{c}$ *is the complement of the first half.*

Proof. The second half of a period of the sequence $\mathbf{a}$ is the complement of the first half and the same is true for the sequence $\mathbf{b}$ according to Theorem 3.5.1. Let T be the period of $\mathbf{c}$. Then

$$\frac{T}{2} = \frac{q_1-1}{4}\frac{q_2-1}{\gcd(q_1-1, q_2-1)} = \frac{q_2-1}{4}\frac{q_1-1}{\gcd(q_1-1, q_2-1)}.$$

By symmetry therefore we have $a_{i+T/2} = \bar{a}$ and $b_{i+T/2} = b_i$ whenever $0 \le i < T/2$. Here $\bar{a}$ denotes the complement of $a_i \in \mathbb{Z}/(2)$. Hence,

$$c_{i+T/2} = \bar{a_i} + b_i \pmod 2 = \bar{c_i}$$

which completes the proof.

□

Lemma 4.2.3 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be binary half-ℓ-sequences with connection integers* $q_1 \equiv 1 \pmod 8$ *and* $q_2 \equiv 1 \pmod{16}$ *respectively. Let* $\mathbf{b}^{(\tau)} = (b_\tau, b_{\tau+1}, \cdots)$ *be the shift of the sequence* $\mathbf{b}$ *by* τ. *If* $\gcd((q_1-1)/2, (q_2-1)/2) = 4$ *and* $\tau = 4k$ *for some* k, *then the sequence* $\mathbf{d} = \mathbf{a} + \mathbf{b}^{(\tau)} \pmod 2$ *is a shift of the sequence* $\mathbf{c} = \mathbf{a} + \mathbf{b} \pmod 2$.

Proof. Since $\tau = 4k$ for some k and $\gcd((q_1-1)/2, (q_2-1)/2) = 4$, we have that $(q_2-1)/4 - \tau$ is divisible by $\gcd((q_1-1)/2, (q_2-1)/2)$. As a result, there exist integers m and n such that

$$\frac{q_2-1}{4} - \tau = m\frac{q_1-1}{2} - n\frac{q_2-1}{2}.$$

That is

$$m\frac{q_1-1}{2} = \frac{q_2-1}{4} - \tau + n\frac{q_2-1}{2}.$$

Therefore, for all j,

$$\begin{aligned}
d_{j+m(q_1-1)/2} &= a_{j+m(q_1-1)/2} + b^{(\tau)}_{j+(q_2-1)/4-\tau+n(q_2-1)/2} \pmod 2 \\
&= a_{j+m(q_1-1)/2} + b_{j+(q_2-1)/4+n(q_2-1)/2} \pmod 2 \\
&= a_j + b_{j+(q_2-1)/4} \pmod 2 \\
&= a_j + \bar{b}_j \pmod 2 \\
&= \bar{c}_j.
\end{aligned}$$

By Lemma 4.2.2 the sequence $\mathbf{c}$ is a shift of its complement, so $\mathbf{d}$ is also a shift of $\mathbf{c}$.

□

Now we return to the general case of N-ary half-ℓ-sequences.

Lemma 4.2.4 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be* N*-ary half-*ℓ*-sequences with connection integers* q_1 *and* q_2 *respectively. Let* $\mathbf{b}^{(1)} = (b_1, b_2, \cdots)$ *be a shift of the sequence* $\mathbf{b}$ *by* 1*. If* $\gcd((q_1-1)/2, (q_2-1)/2) = 1$*, then the sequence* $\mathbf{d} = \mathbf{a} + \mathbf{b}^{(1)} \pmod N$ *is a shift of the sequence* $\mathbf{c} = \mathbf{a} + \mathbf{b} \pmod N$*.*

Proof. Since $\gcd((q_1-1)/2, (q_2-1)/2) = 1$, we have that $(q_2-1)/2 - 1$ is divisible by $\gcd((q_1-1)/2, (q_2-1)/2)$. As a result, there exist integers m and n such that

$$\frac{q_2-1}{2} - 1 = m\frac{q_1-1}{2} - n\frac{q_2-1}{2}.$$

That is

$$m\frac{q_1-1}{2} = (n+1)\frac{q_2-1}{2} - 1.$$

Therefore, for all j,

$$\begin{aligned} d_{j+m(q_1-1)/2} &= a_{j+m(q_1-1)/2} + b^{(1)}_{j+(n+1)(q_2-1)/2-1} \pmod N \\ &= a_j + b_j \pmod N \\ &= c_j. \end{aligned}$$

□

4.2.2 Distribution of Combined Half-ℓ-sequences

Theorem 4.2.2 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be* N*-ary half-*ℓ*-sequences with connection integers* q_1 *and* q_2 *respectively. Let* $\mathbf{c} = \mathbf{a} + \mathbf{b} \pmod N$ *with period* T*. For* $1 \le k < T$*, let* $M_{\mathbf{a}}(\bar{x})$ *be the number of occurrences of* k *consecutive symbols* $\bar{x} = (x_1, x_2, \cdots, x_k)$ *in* $\mathbf{a}$ *and* $M_{\mathbf{b}}(\bar{y})$ *be the number of occurrences of* k *consecutive symbols* $\bar{y} = (y_1, y_2, \cdots, y_k)$ *in* $\mathbf{b}$ *where* $x_i, y_i \in \mathbb{Z}/(N)$ *with* $1 \le i \le k$*. Let* $M_{\mathbf{c}}(\bar{v})$ *be the number of occurrences of* $\bar{v} = (v_1, v_2, \cdots, v_k)$ *in* $\mathbf{c}$ *where* $v_i \in \mathbb{Z}/(N)$*. If*

$\gcd((q_1-1)/2,(q_2-1)/2)=1$, *then*

$$M_{\mathbf{c}}(\bar{v}) = \sum_{\bar{x}\in(\mathbb{Z}/(N))^k} M_{\mathbf{a}}(\bar{x})\cdot M_{\mathbf{b}}(\bar{v}-\bar{x} \pmod N).$$

Proof. From Lemma 4.2.4, we know that $\mathbf{a}+\mathbf{b}^{(1)} \pmod N$ is a shift of $\mathbf{c}$. By induction, we can see that $\mathbf{d}=\mathbf{a}+\mathbf{b}^{(\tau)} \pmod N$ is a shift of $\mathbf{c}$ for $1\le\tau<T$.

If we count the occurrences of the length k block $\bar{v}$ in both $\mathbf{c}$ and $\mathbf{d}$, then we will have twice the number of occurrences of the length k block in $\mathbf{c}$. We have $\mathbf{c}=\mathbf{a}+\mathbf{b} \pmod N$ and each symbol in $\mathbf{a}$ is matched with each symbol in $\mathbf{b}$. Thus, to count the occurrences of $\bar{v}$ in $\mathbf{c}$, we want to count the number of pairs (an occurrence of $\bar{x}$ in $\mathbf{a}$, an occurrence of $\bar{v}-\bar{x} \pmod N$ in $\mathbf{b}$). Thus we sum over all satisfactory blocks $\bar{x}$ and $\bar{v}-\bar{x} \pmod N$ the number of occurrences of $\bar{x}$ in $\mathbf{a}$ times the number of occurrences of $\bar{v}-\bar{x} \pmod N$ in $\mathbf{b}$, that is,

$$M_{\mathbf{c}}(\bar{v}) = \sum_{\bar{x}\in(\mathbb{Z}/(N))^k} M_{\mathbf{a}}(\bar{x})\cdot M_{\mathbf{b}}(\bar{v}-\bar{x} \pmod N).$$

□

Corollary 4.2.1 *Let* $\mathbf{a}=\{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b}=\{b_i\}_{i=0}^{\infty}$ *be binary half-ℓ-sequences with connection integers* q_1 *and* q_2 *respectively. Let* $\mathbf{c}=\mathbf{a}+\mathbf{b} \pmod N$. *If the connection integer of one of the two half-ℓ-sequences is congruent to* $1 \pmod 8$, *i.e.* $q_1\equiv 1 \pmod 8$ *and* $\gcd((q_1-1)/2,(q_2-1)/2)=1$, *then* $\mathbf{c}$ *is balanced.*

Proof. Let A_1 and A_0 be the number of occurrences 1s and 0s in $\mathbf{a}$ respectively. Then $A_1=A_0$, since $q_1\equiv 1 \pmod 8$ according to Theorem 3.5.1. Let B_1 and B_0 be the number of occurrences 1s and 0s in $\mathbf{b}$ respectively. If $\gcd((q_1-1)/2,(q_2-1)/2)=1$, then

- the number of occurrences 1s in $\mathbf{c}=A_1*B_0+A_0*B_1=A_1*(B_0+B_1)$;

- the number of occurrences 0s in $\mathbf{c}= A_1 * B_1 + A_0 * B_0 = A_1 * (B_0 + B_1)$;

which completes the proof. □

Theorem 4.2.3 *Let* $\mathbf{a} = \{a_i\}_{i=0}^{\infty}$ *and* $\mathbf{b} = \{b_i\}_{i=0}^{\infty}$ *be* N*-ary half-*ℓ*-sequences with connection integers* $q_1 \equiv \pm 1 \pmod 8$ *and* $q_2 \equiv \pm 1 \pmod 8$ *respectively. Let* $\mu(v)$ *be the number of occurrences of* v *in* $\mathbf{c} = \mathbf{a} + \mathbf{b} \pmod N$ *where* $0 \le v < N$*. If* $\gcd((q_1 - 1)/2, (q_2 - 1)/2) = 1$*, then*

$$\left| \mu(v) - \frac{(q_1 - 1)(q_2 - 1)}{4N} \right| \le N \cdot \sigma(q_1) \cdot \sigma(q_2),$$

where $\sigma(q)$ *is the maximum deviation of the number of occurrences of one symbol from the average number of occurrences within one period of a half-*ℓ*-sequence with connection integer* q*.*

Proof. Let $t_i = (q_i - 1)/(2N)$, the average number of occurrences of a symbol in $\mathbf{a}$ and $\mathbf{b}$, and for $0 \le v < N$ let $c_v = A_v - t_1$ and $d_v = B_v - t_2$. Then $\sum_v c_v = \sum_v d_v = 0$. Therefore for $0 \le w < N$,

$$\begin{aligned} \sum_{u+v=w} A_u B_v &= \sum_{u+v=w} (t_1 + c_u)(t_2 + d_v) \\ &= N t_1 t_2 + (\sum_{0 \le u < N} c_u) t_2 + (\sum_{0 \le v < N} d_v) t_1 + \sum_{u+v=w} c_u d_v \\ &= \frac{(q_1 - 1)(q_2 - 1)}{4N} + \sum_{u+v=w} c_u d_v. \end{aligned}$$

Thus

$$|\mu(w) - \frac{(q_1 - 1)(q_2 - 1)}{4N}| \le \sum_{u+v=w} |c_u d_v| \le N \cdot \sigma(q_1) \cdot \sigma(q_2).$$

In particular, $\sigma(q_i)$ is bound by Theorem 3.2.1. □

4.2.3 Experimental Results

In this section, we show experimental results on the combined half-ℓ-sequences based on different pairs of connection integers q_1 and q_2.

Let $\theta = \gcd((q_1 - 1)/2, (q_2 - 1)/2)$ and

$$\mu^{'}(v) = \left| \mu(v) - \frac{(q_1 - 1)(q_2 - 1)}{4N\theta} \right|, \ 0 \leq v < N.$$

The quantity $\mu^{'}(v)$ is the difference between the number of occurrences of v in one period of a combined half-ℓ-sequence and the average number of occurrences. Let $\gamma = \max\{\mu^{'}(v) : 0 \leq v < N\}$.

Recall that our bound for $\sigma(q)$ when q is the connection integer of a half-ℓ-sequence is

$$\sigma(q) = \frac{q-1}{2q} + \frac{(q^{1/2}+1)}{2} \cdot \left(\frac{4}{\pi^2} \log q + 0.38 + \frac{0.608}{q} + \frac{0.116}{q^2} \right).$$

From Theorem 4.2.3, we get the bound for a combined half-ℓ-sequence based q_1 and q_2 when $\theta = 1$. We denote it as

$$\zeta = N\sigma(q_1)\sigma(q_2).$$

We investigate combined half-ℓ-sequences when $N = 2, 4$ and 8. Table 4.1 shows the distribution of combined half-ℓ-sequences when $N = 2$. Notice that when one of the connection integers of the two binary half-ℓ-sequences is congruent to 1 (mod 8), we have $\gamma = 0$, which means the combined sequence is balanced. For example, in Table 4.1 when $q_1 = 41 \equiv 1 \pmod 8$ we have $\gamma = 0$. This result is consistent with the result shown in Corollary 4.2.1. However, when $N = 8$, γ can never be 0 since the connection integers must be congruent to $-1 \pmod N$. From Table 4.1, 4.2 and 4.3, we can see that $\gamma < \zeta$ when $\theta = 1$. This result is consistent with what we find in Theorem 4.2.3. The fact that $\gamma < \zeta$ is not restricted when $\theta = 1$. Actually when

$\theta \neq 1$, e.g., $\theta = 11$ with $q_1 = 23$ and $q_2 = 199$ we have $\gamma < \zeta$.

4.3 Distribution of $(s_n, s_{n+\tau})$ in an ℓ-sequence

Using similar techniques as for half-ℓ-sequences in Chapter 3, we have the following theorem for the distribution of $(s_n, s_{n+\tau})$ in one period of an ℓ-sequence $\mathbf{s}$.

Theorem 4.3.1 *For an N-ary ℓ-sequence $\mathbf{s}$ with prime connection integer q and $0 \leq \tau < T$, the number $\mu(\tau; v_1, v_2)$ of occurrences of $(s_n, s_{n+\tau})$ with $s_n = v_1$ and $s_{n+\tau} = v_2$ for $0 \leq n < T$ satisfies*

$$\left| \mu(\tau; v_1, v_2) - \frac{q-1}{N^2} \right| \leq N^{\tau-1} \left(1 + \ln \left(\frac{q-1}{2} \right) \right).$$

where $0 \leq v_1, v_2 < N$.

However, we can get a sharper bound using another method according to Lemma 4.3.1 below.

Lemma 4.3.1 *[30] Let $\mathbf{s}$ be an N-ary ℓ-sequence based on connection integer q with q an odd prime. Then the number $M_{\mathbf{s}}(\mathbf{b})$ of occurrences of any block $\mathbf{b}$ of size τ within a single period of $\mathbf{s}$ is*

$$\left\lfloor \frac{q}{N^\tau} \right\rfloor \leq M_{\mathbf{s}}(\mathbf{b}) \leq \left\lfloor \frac{q}{N^\tau} \right\rfloor + 1.$$

Theorem 4.3.2 *For an N-ary ℓ-sequence $\mathbf{s}$ with prime connection integer q and $0 \leq \tau < T$, the number $\mu(\tau; v_1, v_2)$ of integers n with $s_n = v_1$ and $s_{n+\tau} = v_2$ for $0 \leq n < T$ satisfies*

$$\left| \mu(\tau; v_1, v_2) - \frac{q-1}{N^2} \right| < N^{\tau-1} + N^2.$$

Table 4.1: Distribution of combined half-ℓ-sequences when $N = 2$

q_1	q_2	θ	γ	ζ
23	41	1	0	80
23	47	1	8	87
23	71	1	11	111
23	79	1	8	119
23	97	1	0	135
23	103	1	8	140
23	137	1	0	166
23	167	1	17	188
23	191	1	20	204
23	193	1	0	205
23	199	11	6	209
23	239	1	23	233
23	263	1	20	248
23	271	1	17	252
23	311	1	29	274
23	313	1	0	275
23	359	1	29	299
23	367	1	14	303
23	383	1	26	312
23	401	1	0	320
23	409	1	0	324
41	47	1	0	122
41	71	5	0	157
41	79	1	0	167
41	97	4	0	190
41	103	1	0	197
41	137	4	0	234
41	167	1	0	264
41	191	5	0	287
41	193	4	0	289
41	199	1	0	294
41	239	1	0	329
41	263	1	0	349
41	271	5	0	355
41	311	5	0	387
41	313	4	30	388
41	359	1	0	422
41	367	1	0	428
41	383	1	0	439
41	401	20	0	451
41	409	4	30	457
47	71	1	18	170

Table 4.2: Distribution of combined half-ℓ-sequences when $N = 4$

q_1	q_2	θ	γ	ζ
23	47	1	7	174
23	71	1	10	223
23	79	1	7	238
23	103	1	7	280
23	167	1	16	376
23	191	1	19	408
23	199	11	8	418
23	239	1	22	467
23	263	1	19	496
23	271	1	16	505
23	311	1	28	549
23	359	1	28	599
23	367	1	13	607
23	383	1	25	624
47	71	1	17	341
47	79	1	12	364
47	103	1	12	428
47	167	1	27	574
47	191	1	32	623
47	199	1	22	639
47	239	1	37	715
47	263	1	32	758
47	271	1	27	772
47	311	1	47	840
47	359	1	47	916
47	367	1	22	929
47	383	1	42	953
71	79	1	17	467
71	103	1	17	549
71	167	1	38	737
71	191	5	31	800
71	199	1	31	820
71	239	7	22	917
71	263	1	45	972
71	271	5	7	990
71	311	5	18	1077
71	359	1	66	1175
71	367	1	31	1191
71	383	1	59	1223
79	103	3	7	585
79	167	1	27	786
79	191	1	32	853

Table 4.3: Distribution of combined half-ℓ-sequences when $N = 8$

q_1	q_2	θ	γ	ζ
167	263	1	53	3272
167	359	1	77	3956
167	383	1	70	4115
167	479	1	101	4716
167	503	1	86	4859
167	719	1	125	6038
167	839	1	134	6632
167	863	1	86	6746
167	887	1	118	6859
263	359	1	93	5219
263	383	1	81	5429
263	479	1	122	6222
263	503	1	102	6410
263	719	1	152	7967
263	839	1	162	8750
263	863	1	101	8901
263	887	1	142	9050
359	383	1	114	6564
359	479	1	185	7523
359	503	1	150	7751
359	719	1	233	9632
359	839	1	246	10579
359	863	1	146	10762
359	887	1	214	10942
383	479	1	151	7826
383	503	1	132	8063
383	719	1	184	10020
383	839	1	198	11005
383	863	1	136	11195
383	887	1	176	11383
479	503	1	197	9240
479	719	1	307	11483
479	839	1	323	12612
479	863	1	193	12830
479	887	1	281	13045
503	719	1	245	11831
503	839	1	261	12993
503	863	1	164	13217
503	887	1	229	13439
719	839	1	407	16148
719	863	1	238	16426
719	887	1	353	16702

Proof. Let $\mathbf{b}$ denote any block of $\tau-1$ consecutive symbols. Let $M_{\mathbf{s}}(v_1, \mathbf{b}, v_2)$ be the occurrences of $v_1, \mathbf{b}, v_2$ within one period of s. According to Lemma 4.3.1, we have

$$\left| M_{\mathbf{s}}(v_1, \mathbf{b}, v_2) - \frac{q-1}{N^{\tau+1}} \right| \leq 1 + \frac{1}{N^{\tau+1}},$$

and

$$\mu(\tau; v_1, v_2) = \sum_{\mathbf{b}} M_{\mathbf{s}}(v_1, \mathbf{b}, v_2),$$

where the sum is over all possible choices of $\mathbf{b}$ and there are $N^{\tau-1}$ of them. Then we have

$$\begin{aligned} \left| \mu(\tau; v_1, v_2) - \frac{q-1}{N^2} \right| &= \left| N^{\tau-1} M_{\mathbf{s}}(v_1, \mathbf{b}, v_2) - \frac{q-1}{N^2} \right| \\ &= N^{\tau-1} \left| M_{\mathbf{s}}(v_1, \mathbf{b}, v_2) - \frac{q-1}{N^{\tau+1}} \right| \\ &\leq \left(1 + \frac{1}{N^{\tau+1}} \right) N^{\tau-1} = N^{\tau-1} + \frac{1}{N^2}. \end{aligned}$$

which completes the proof. □

Furthermore, we get a sharper bound when $N^{\tau+1} < q$ with constraints on q. To do this, we investigate the bound for $|\mu(\tau; v_1, v_2) - \mu(\tau; u_1, u_2)|$ instead where v_1, v_2 and u_1, u_2 are two pairs of values that vary from 0 to $N-1$. Here $\mu(\tau; v_1, v_2)$ is the number of integers n with $s_n = v_1$ and $s_{n+\tau} = v_2$ in an ℓ-sequence $\mathbf{s}$ and similarly $\mu(\tau; u_1, u_2)$ is the number of integers n with $s_n = u_1$ and $s_{n+\tau} = u_2$ in $\mathbf{s}$.

Theorem 4.3.3 *Let* $\mathbf{s}$ *be an* N*-ary* ℓ*-sequence with prime connection integer* $q = \sum_{i=0}^{r} q_i N^i$ *where* $0 \leq q_i < N$ *for some positive integer* r *and* $0 \leq \tau < T$. *If* $q_1 = q_2 = \cdots = q_\tau = 0$ *or* $q_1 = q_2 = \cdots = q_\tau = N-1$, *then the numbers* $\mu(\tau; v_1, v_2)$ *and*

$\mu(\tau; u_1, u_2)$ of occurrences of $(s_n, s_{n+\tau})$ with $s_n = v_1, s_{n+\tau} = v_2$ and $s_n = u_1, s_{n+\tau} = u_2$ for $0 \le n < T$ satisfy

$$|\mu(\tau; v_1, v_2) - \mu(\tau; u_1, u_2)| \le 1,$$

where $0 \le v_1, v_2, u_1, u_2 < N$.

Proof. Let $\mathbf{b}$ be a block of consecutive symbols with length $\tau + 1$ and let

$$b = \sum_{i=0}^{\tau} b_i N^i \quad \text{and} \quad q' = \sum_{i=0}^{\tau} q_i N^i$$

where $b_0 = v_1$, $b_\tau = v_2$. The proof of Lemma 4.3.1 in [30] shows that if $b < q'$, then $n(\mathbf{b}) = n_1 + 1$ and if $b \ge q'$, then $n(\mathbf{b}) = n_1$ where $n_1 = \lfloor q/N^{\tau+1} \rfloor$.

We first count the occurrences of $(s_n, s_{n+\tau})$ with $s_n = v_1, s_{n+\tau} = v_2$. If $b_\tau = v_2 > q_\tau$, then $b > q'$ and there are $N^{\tau-1}$ such bs. If $b_\tau = v_2 < q_\tau$, then $b < q'$ and there are $N^{\tau-1}$ such bs. If $v_2 = q_\tau$, then let $q'' = q' - q_\tau N^\tau = \sum_{i=0}^{\tau-1} q_i N^i$ and $b'' = b - v_2 N^\tau$. We see that $b < q'$ if and only if $b'' < q''$.

Let $\rho_1 = |\{b : 0 \le b'' < q''\}|$. Then

$$\rho_1 = \sum_{i=1}^{\tau-1} q_i N^{i-1} + \rho_1^{'}, \tag{4.1}$$

where the value of $\rho_1^{'}$ depends on the relation between v_1 and q_0. Actually,

$$\rho_1^{'} = \begin{cases} 1 & \text{if } v_1 < q_0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Let $\rho_2 = |\{b : b'' \ge q''\}|$. Since the total number of b''s is $N^{\tau-1}$, we have

$$\rho_2 = N^{\tau-1} - \rho_1.$$

From the analysis above we can get all the possible value for $\mu(\tau; v_1, v_2)$, which

are

(I) $(n_1+1)N^{\tau-1}$ if $v_2 < q_\tau$;

(II) $n_1 N^{\tau-1}$ if $v_2 > q_\tau$;

(III) $\sum_{i=1}^{\tau-1} q_i N^{i-1} + 1 + N^{\tau-1} n_1$ if $v_2 = q_\tau$ and $v_1 < q_0$;

(IV) $\sum_{i=1}^{\tau-1} q_i N^{i-1} + N^{\tau-1} n_1$ if $v_2 = q_\tau$ and $v_1 \geq q_0$.

The above four cases also apply to $\mu(\tau; u_1, u_2)$.

When $q_1 = q_2 = \cdots = q_\tau = 0$, by eqs. (4.1) and (4.2) we have $\rho_1 = 1$ or 0. Since v_1, v_2, u_1 and u_2 vary from 0 to $N-1$, $\mu(\tau; v_1, v_2)$ and $\mu(\tau; u_1, u_2)$ can only reach the possible value in (II), (III) or (IV). Thus, $\mu(\tau; v_1, v_2)$ can be either $n_1 N^{\tau-1}$ or $n_1 N^{\tau-1} + 1$. Similarly $\mu(\tau; u_1, u_2)$ can be either $n_1 N^{\tau-1}$ or $n_1 N^{\tau-1} + 1$. As a result,

$$|\mu(\tau; v_1, v_2) - \mu(\tau; u_1, u_2)| \leq 1.$$

When $q_1 = q_2 = \cdots = q_\tau = N-1$, by eqs. (4.1) and (4.2) we have

$$\rho_1 = N^{\tau-1} \quad \text{or} \quad N^{\tau-1} - 1.$$

In this case, $\mu(\tau; v_1, v_2)$ and $\mu(\tau; u_1, u_2)$ can only reach the possible value in (I), (III) or (IV). Thus

$$|\mu(\tau; v_1, v_2) - \mu(\tau; u_1, u_2)| \leq 1.$$

□

4.4 Autocorrelation of Binary Half-ℓ-sequences

Correlation properties of pseudorandom sequences are important measures of randomness. They have practical applications in spread spectrum communication systems, radar systems, cryptanalysis, and so on [30]. Recall that the autocorrelation function of a binary periodic sequence $\mathbf{s} = \{s_i\}_0^\infty$ with period T is defined as

$$C_{\mathbf{s}}(\tau) = \sum_{i=0}^{T-1} (-1)^{s_i + s_{i+\tau}}$$

for $0 \leq \tau < T$. The autocorrelation function measures the similarity of the sequence and its shifted versions. We have $C_{\mathbf{s}}(0) = T$. Much research has been done on criteria for optimal autocorrelation sequences [12, 67]. A sequence $\mathbf{s}$ is said to have optimal autocorrelation if for any $\tau \neq 0$, we have

(1) $C_{\mathbf{s}}(\tau) = -1$ and $T \equiv -1 \pmod 4$; or

(2) $C_{\mathbf{s}}(\tau) \in \{1, -3\}$ and $T \equiv 1 \pmod 4$; or

(3) $C_{\mathbf{s}}(\tau) \in \{2, -2\}$ and $T \equiv 2 \pmod 4$; or

(4) $C_{\mathbf{s}}(\tau) \in \{0, -4\}$ and $T \equiv 0 \pmod 4$.

Sequences satisfying criteria (1) include Legendre sequences, Hall's sextic residue sequences, twin-prime sequences, m-sequences, GMW sequences, and Maschiettie's hyperoval sequences. These sequences are also said to have *ideal 2-level autocorrelation*. One can find more detailed definitions of these sequences in [12, 67].

In this section, we investigate the autocorrelation properties of half-ℓ-sequences using eq. (2.6) with $N = 2$. For $0 \leq \tau < T$, we see that

$$s_i + s_{i+\tau} = \left(2^{-i}h \pmod q \pmod 2\right) + \left(2^{-\tau}2^{-i}h \pmod q \pmod 2\right).$$

As before, h is a quadratic residue modulo q and hence the autocorrelation function of $\mathbf{s}$ at shift τ can be written as

$$C_s(\tau) = \sum_{x \in Q} (-1)^{x + \left(2^{-\tau} \cdot x \pmod q\right)},$$

where Q is the set of quadratic residues modulo q. We need the following technical lemmas.

Lemma 4.4.1 *[61] Let $q > 3$ be a prime number. For $1 < u < q-1$, we have*

$$\left| \sum_{x=1}^{q-1} (-1)^{x + (u \cdot x \pmod q)} \right| \leq 2\left(\left\lceil \frac{q}{6} \right\rceil - 1 \right).$$

Moreover, for $k = 0, 1$,

$$\sum_{x=1}^{q-1} (-1)^{x + (u \cdot x \pmod q)} = (-1)^k \cdot 2\left(\left\lceil \frac{q}{6} \right\rceil - 1 \right)$$

if and only if

$$u \equiv (-1)^k \cdot 3 \text{ or } (-1)^k \cdot 3^{-1} \pmod q.$$

Let Q' denote the set of non-quadratic residues modulo q.

Lemma 4.4.2 *Let q be an odd prime and $q \equiv 7 \pmod 8$. We have*

$$\sum_{x \in Q} (-1)^{x + (u \cdot x \pmod q)} = \sum_{x \in Q'} (-1)^{x + (u \cdot x \pmod q)}. \tag{4.3}$$

Proof. When $q \equiv 7 \pmod 8$, by the law of quadratic reciprocity [49] we have $2 \in Q$

and $-1 \in Q^{'}$. Then one can check

$$\begin{aligned}
\sum_{x \in Q^{'}} (-1)^{x+(u \cdot x \pmod q)} &= \sum_{x \in Q} (-1)^{(-x) \pmod q + (u \cdot (-x) \pmod q)} \\
&= \sum_{x \in Q} (-1)^{(q-x)+(q-(u \cdot x \pmod q))} \\
&= \sum_{x \in Q} (-1)^{-(x+(u \cdot x \pmod q))} \\
&= \sum_{x \in Q} (-1)^{x+(u \cdot x \pmod q)}.
\end{aligned}$$

□

Remarks. Lemma 4.4.2 is not true if $q \equiv 1 \pmod 8$. For example, if $q = 41$ and $u \equiv 2 \pmod q$, then the left hand side of eq. (4.3) is 8 while the right hand side of eq. (4.3) is -8.

Theorem 4.4.1 *Let* $\mathbf{s}$ *be a binary half-*ℓ*-sequence with prime connection integer* $q > 3$ *and* $q \equiv 7 \pmod 8$. *For* $0 < \tau < (q-1)/2$, *the autocorrelation of* $\mathbf{s}$ *satisfies*

$$|C_{\mathbf{s}}(\tau)| \leq \left\lceil \frac{q}{6} \right\rceil - 1.$$

Moreover, if 3 *is a quadratic residue modulo* q, *then*

$$C_{\mathbf{s}}(\tau) = \left\lceil \frac{q}{6} \right\rceil - 1$$

if and only if

$$2^{-\tau} \equiv 3 \text{ or } 3^{-1} \pmod q.$$

If 3 *is a non-quadratic residue modulo* q, *then*

$$C_{\mathbf{s}}(\tau) = 1 - \left\lceil \frac{q}{6} \right\rceil$$

if and only if

$$2^{-\tau} \equiv -3 \text{ or } -3^{-1} \pmod{q}.$$

Proof. By Lemma 4.4.2, we get

$$C_{\mathbf{s}}(\tau) = \frac{1}{2}\sum_{x=1}^{q-1}(-1)^{x+\left(2^{-\tau}\cdot x \pmod{q}\right)}.$$

Then applying Lemma 4.4.1, we get the bound of $C_{\mathbf{s}}(\tau)$.

Since $2^{-\tau}$ is a quadratic residue modulo q, according to Lemma 4.4.1 again, we have

$$C_{\mathbf{s}}(\tau) = \left\lceil \frac{q}{6} \right\rceil - 1 \text{ if and only if } 2^{-\tau} \equiv 3 \text{ or } 3^{-1} \pmod{q},$$

if 3 is a quadratic residue modulo q, and otherwise

$$C_{\mathbf{s}}(\tau) = 1 - \left\lceil \frac{q}{6} \right\rceil \text{ if and only if } 2^{-\tau} \equiv -3 \text{ or } -3^{-1} \pmod{q}.$$

□

Since the autocorrelation value of a binary half-ℓ-sequence does not satisfy any of the four criteria for an optimal sequence, it is not optimal.

Remarks. Theorem 4.4.1 does not holds when $q \not\equiv 7 \pmod 8$. For example, when $q = 41 \equiv 1 \pmod 8$, $C_{\mathbf{s}}(\tau) = 8 > \lceil q/6 \rceil - 1 = 6$ when $2^{-\tau} = 2 \pmod q$.

4.5 Concluding Remarks

In this chapter, we introduce distribution properties of pseudorandom sequences by combining two half-ℓ-sequences using modular addition. A bound for the number of occurrences of combined half-ℓ-sequences in one symbol case is given. Bounds on the higher odrder distribution (e.g., the number of occurrences of two symbols) are not discussed here due to the increased number of cases and the resulting weaker

bound. We present bounds on the distribution of pairs $(s_n, s_{n+\tau})$ for ℓ-sequences. The autocorrelation of half-ℓ-sequences is also discussed.

Chapter 5 Correlation Immune Functions

5.1 Introduction

In this chapter, a new correlation attack on nonlinear combination generators is proposed. The success of this attack depends on the correlation between the output of a nonlinear function of several LFSRs and the output of the nonlinear combination function in the generator. To measure resistance to such attacks, we introduce the idea of q-correlation immune functions. We investigate the properties of these functions and their constructions.

Figure 5.1 shows a nonlinear combination generator with n LFSRs and a nonlinear combination function f. Let $X^t = \{x_1^t, x_2^t, \cdots, x_n^t\}$ denote the n output bits from

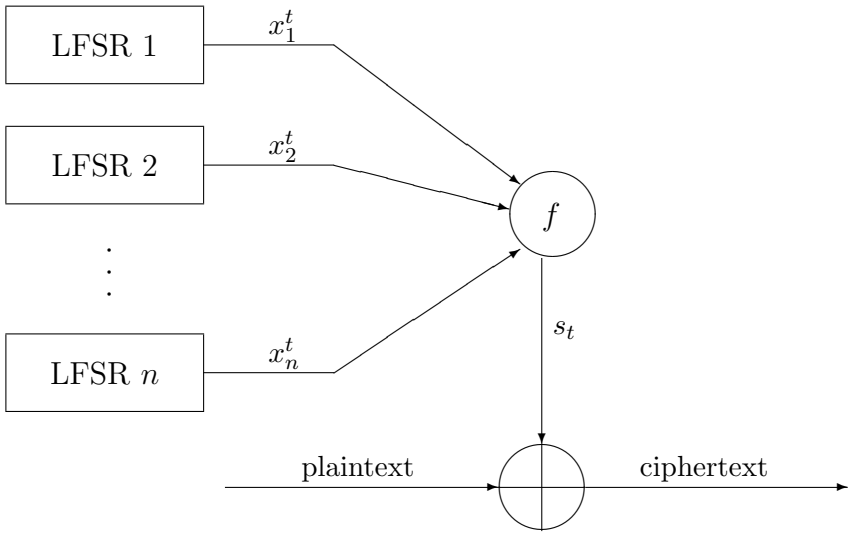

Figure 5.1: Stream Cipher with a Nonlinear Combination Generator

the n LFSRs at time t and r_i be the length of the ith LFSR for $1 \leq i \leq n$. The combination function f is a Boolean function of n variables from $\mathbb{F}_2^n$ to $\mathbb{F}_2$: at time t

the generator output a single bit

$$s_t = f(X^t).$$

This single bit s_t will XOR with a plaintext bit to get a ciphertext bit. The combination function f should be balanced in order to generate keystream bits as close to the uniform distribution as possible. In other words, f should output 0 or 1 with probability 1/2.

5.1.1 Correlation Attacks

There are many types of correlation attacks on stream ciphers and block ciphers. The goal of these attacks is to recover the secret key. In a brute force attack of the scheme in Figure 5.1, we may need $O(2^{r_1+r_2+\cdots+r_n})$ operations to get the initial states of the LFSRs. However, this is infeasible if $r_1 + ... + r_n$ is large enough (128 will do). Siegenthaler designed a correlation attack based on the correlation between the output of LFSRs and keystream from the combination function in a divide-and-conquer manner [59]. One can guess and exploit the initial states of the LFSRs one by one. The complexity of this attack needs $O(\sum_{i=1}^{n} 2^{r_i})$ operations, which is much less than is needed by a brute force attack. The attack applies if and only if the output keystream is correlated to the output of one LFSR or a linear combination of the output of LFSRs. For simplicity, we only discuss the situation when the output keystream is correlated to the output of one LFSR. This equivalently means that there exists i such that the linear bias

$$\varepsilon = \left| \text{prob}(x_i^t = s_t) - \frac{1}{2} \right| \neq 0,$$

where x_i^t and s_t are the outputs at time t from the ith LFSR and keystream generator respectively. This correlation can be detected by computing the correlation between

N bits of the keystream and the corresponding output bits of the LFSR generated from a guessed initial state x_i^0:

$$C(s_t, x_i^t) = \sum_{t=0}^{N-1} (-1)^{s_t \oplus x_i^t}.$$

The expected value of the quantity is $2N\varepsilon$ when x_i^0 is the correct value of the initial state for the ith LFSR. This attack consists of an exhaustive search of all the r_i bits of the initial state x_i^0. For each guess value x_i^0, the correlation between the N keystream bits and output bits of LFSR is calculated. By comparing the $C(s_t, x_i^t)$ with a given threshold one can determine whether a guess is right or wrong. The output keystream bits are expected to be uncorrelated with the output of ith LFSR for a wrong guess.

To resist such attacks, Siegenthaler introduced the notion of correlation immune functions, which reflects a relation between $f(x)$ and the linear functions (or affine functions). We recall that $f(x)$ is correlation immune of order k if its values are statistically independent of any subset of k input variables, or $f(x)$ is statistically independent of any linear or affine functions. Xiao and Massey investigated correlation immunity of order k via the Walsh-Hadamard transform (see Section 2.3) as shown in the following theorem.

Theorem 5.1.1 *(Xiao-Massey Theorem)[65] A Boolean function $f(x)$ is correlation immune of order k if and only if $W(f)(\omega) = 0$ for all $\omega \in \mathbb{F}_2^n$ with $1 \leq wt(\omega) \leq k$.*

5.1.2 New Correlation Attacks

The correlation attacks introduced by Siegenthaler are based on the correlation between keystream sequence and a linear combination of the outputs of several LFSRs. It is natural to ask whether we can find a correlation between keystream sequence and a nonlinear combination of the outputs of LFSRs, and, if so, will it need fewer

keystream bits and have lower time complexity? Let $q(x) = q(x_1, x_2, \cdots, x_n) \in \mathbf{B}_n$, which combines the outputs from k LFSRs. More precisely, $q(x)$ depends on only k variables $x_{i_1}, x_{i_2}, \cdots, x_{i_k}$ where $1 \leq k \leq n$. Actually, the bigger k is the more keystream bits are needed by the attack [17].

We can build a statistical model as Siegenthaler does for the correlation attack by measuring the correlation between $f(x)$ and $q(x)$. Let X_i be the random variable over the output of the ith LFSR for $1 \leq i \leq n$. Each X_i satisfy the probability distribution

$$\text{prob}(X_i) = \begin{cases} \text{prob}(x_i = 1) = 1/2 \\ \\ \text{prob}(x_i = 0) = 1/2. \end{cases}$$

Note here we assume the output sequence of each LFSR is an m-sequence.

Let Y^t be the random variable over the output of function f. We have $P(Y^t = 1) = P(Y^t = 0)$. Let Z^t be the random variable over the output of function $q(x)$. We have

$$\text{prob}(Z^t) = \begin{cases} \text{prob}(z = 1) = \frac{wt(q)}{2^k} \\ \\ \text{prob}(z = 0) = 1 - \frac{wt(q)}{2^k}. \end{cases}$$

We also have

$$\text{prob}[(Y^t \oplus Z^t) = 1] = \frac{1}{2} - \varepsilon.$$

We use random variable

$$\alpha = \sum_{t=0}^{N-1} (1 - 2(Y^t \oplus Z^t)) \tag{5.1}$$

as a measure for the correlation between Y^t and Z^t. Since all the terms $(Y^t \oplus Z^t)$ in the sum of eq. (5.1) are independent and identically distributed random variables, $\beta = \sum_{t=0}^{N-1} (Y^t \oplus Z^t)$ satisfies a binomial distribution. It has mean value m_β and

variance σ_β^2 given by

$$m_\beta = N(\frac{1}{2} - \varepsilon)$$

and

$$\sigma_\beta^2 = N(\frac{1}{2} + \varepsilon)(\frac{1}{2} - \varepsilon).$$

The mean value m_α and variance σ_α^2 will be

$$m_\alpha = N - 2N(\frac{1}{2} - \varepsilon) = 2\varepsilon$$

and

$$\sigma_\alpha^2 = 2^2\sigma_\beta^2 = 4N(\frac{1}{2} + \varepsilon)(\frac{1}{2} - \varepsilon).$$

When $\varepsilon = 0$, we have

$$m_\alpha = 0$$

and

$$\sigma_\alpha^2 = N.$$

The random variable α can be assumed to be normally distributed with mean value m_α and variance σ_α^2 for large N due to the central limit theorem.

Our attack works as follows.

step 1: Observe N bits of keystream s_t where $s_t \in \mathbb{F}_2$. These keystream bits are not required to be consecutive, but if they are not, then we need to record their positions.

step 2: Guess the initial states of k different LFSRs. There are $2^{r_{i_1}+\cdots+r_{i_k}}$ possible initial states. For each of the guessed initial states, we compute the state value $(x_{i_1}^t, x_{i_2}^t, \cdots, x_{i_k}^t)$ at corresponding time t. Then we compute $q(x_{i_1}^t, x_{i_2}^t, \cdots, x_{i_k}^t)$ and evaluate α.

step 3: Input a constant threshold value c^* for α. If $\alpha > c^*$, the guessed initial

states are regarded as the right ones. Otherwise, run the above steps again with new guessed initial states.

The detail of the calculation is related to hypothesis testing [51].

5.1.3 q-transform

The Walsh-Hadamard transform measures the relations between $f(x)$ and affine functions. Sometimes we need to consider a relation between $f(x)$ and a function of small degree but larger than one. A typical application is the algebraic attack on stream ciphers. Klapper introduced the notion of q-transform [38], which is a generalization of the Walsh-Hadamard transform, to measure the proximity of two functions.

Let GL_n be the set of nonsingular n by n matrices with entries in $\mathbb{F}_2$. The cardinality of GL_n is

$$N = (2^n - 2^0)(2^n - 2^1)\cdots(2^n - 2^{n-1}).$$

For a Boolean function $q(x) \in \mathbf{B}_n$, the *q-transform* of $f(x)$ at $A \in GL_n$ is the real valued function on GL_n

$$\overline{W}_q(f)(A) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x)+q(xA)}.$$

In fact, the q-transform measures the Hamming distance between $f(x)$ and the functions from the following set

$$\mathcal{S}_q = \{q(xA) : A \in GL_n\},$$

which is the smallest set of functions obtained from $q(x)$ by change of basis. We remark that, if $q(x_1, x_2, \ldots, x_n) = x_i$ for some $1 \leq i \leq n$ (in fact $q(x)$ can be any linear function), $\overline{W}_q(f)(A)$ exactly runs through $W(f)(\omega)$, where $\omega \neq \overline{0}$, $N/(2^n - 1)$

many times when A ranges over GL_n. Additionally we set

$$\overline{W}_q(f)(\mathbf{0}) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x)},$$

where $\mathbf{0}$ is the zero matrix of n by n.

Define the set for $q(x)$

$$\mathcal{H}_q = \{H \in GL_n : q_H(x) = q(x)\},$$

which is called the stabilizer of $q(x)$. One can show that $\mathcal{H}_q$ is a subgroup of GL_n since $\mathcal{H}_q$ is closed under multiplication. The cosets

$$A\mathcal{H}_q = \{AH : H \in \mathcal{H}_q\}, \quad A \in GL_n$$

give a partition of GL_n. One can check

$$\overline{W}_q(f)(AH) = \overline{W}_q(f)(A) \quad \text{for} \quad H \in \mathcal{H}_q. \tag{5.2}$$

So we only need to take a representative of each coset of $\mathcal{H}_q$ into account.

We can generalize the new correlation attack in section 5.1.2 by using $q_A(x) = q(xA)$ where $A \in GL_n$. All these $q_A(x)$ where $A \in GL_n$ can be used as nonlinear functions that combine the outputs of LFSRs to launch the new correlation attacks. We call these attacks q-correlation attacks. To resist q-correlation attacks, the function $f(x)$ must be statistically independent of $q_A(x)$ for $A \in GL_n$ just as $f(x)$ needs to be correlation immune to resist correlation attacks. The q-transform is a tool for understanding resistance to a q-correlation attack. We define q-correlation immune functions in the next section.

5.2 Definitions of q-correlation Immune Functions

In order to introduce the notion of the q-correlation immune function, we define the weight of a matrix as follows.

Definition 5.2.1 *Let $A = (a_1|a_2|\dots|a_n)$ be an n by n matrix over $\mathbb{F}_2$, where a_i is the i-th column of A for $1 \le i \le n$. We define the* weight *of A, denoted by $col.wt(A)$, as the maximal value among the weights of the columns of A, i.e.,*

$$col.wt(A) = \max\{wt(a_i) : 1 \le i \le n\}.$$

Furthermore we define the weight of a matrix A with respect to $q(x)$ or more precisely with respect to $\mathcal{H}_q$.

Definition 5.2.2 *Let $A \in GL_n$. We define the* weight *of A with respect to $q(x)$, denoted by $wt_q(A)$, as the minimal value among the weights of AH for all $H \in \mathcal{H}_q$, i.e.,*

$$wt_q(A) = \min\{col.wt(AH) : H \in \mathcal{H}_q\}.$$

Then we can choose a representative matrix A of the coset $A\mathcal{H}_q$ with the weight $wt_q(A)$.

Let

$$S_1 = \{A \in GL_n : q(xA) \text{ depends on at most } k \text{ vairiables}\}$$

and

$$S_2 = \{A \in GL_n : 1 \le wt_q(A) \le k\}.$$

Lemma 5.2.1 *Suppose $q(x) = \prod_{i\in I} x_i$ where $I \subset \{1, 2, \cdots, n\}$. In other words, $q(x)$ is a monomial. Suppose that $q_A(x) = q(xA)$ depends on only k variables. Then all x_j appearing in xA_i also appear in $q_A(x)$, so $wt(A_i) \le k$ for $i \in I$.*

Proof. Suppose that for some $i \in I$ we have $wt(A_i) > k$. Let $wt(A_s) = t > k$ where $s \in I$ we have $xA_s = x_{s_1} + x_{s_2} + \cdots + x_{s_t}$. Then $q(xA) = \prod_{i\in I}(xA_i) = (x_{s_1} + x_{s_2} + \cdots + x_{s_t}) \prod_{i\in I/\{s\}}(xA_i)$. We claim that $x_{s_1}, x_{s_2}, \cdots, x_{s_t}$ appear in $q(xA)$. Otherwise, suppose x_{s_1} vanishes in $q(xA)$ and let $\prod_{i\in I/\{s\}}(xA_i) = x_{s_1}C + D$ where C and D do not contain x_{s_1}. Then we have

$$\begin{aligned} & x_{s_1}(x_{s_1}C + D) + x_{s_1}x_{s_2}C + \cdots + x_{s_1}x_{s_t}C \\ = \; & x_{s_1}(C + D + x_{s_2}C + \cdots + x_{s_t}C) = 0, \end{aligned}$$

which means $C + D + x_{s_2}C + \cdots + x_{s_t}C = 0$ in $\mathbb{F}_2$. Then we have

$$\begin{aligned} & q(xA) = (x_{s_1} + x_{s_2} + \cdots + x_{s_t})(x_{s_1}C + D) \\ = \; & (x_{s_2} + \cdots + x_{s_t})D \\ = \; & (x_{s_2} + \cdots + x_{s_t})(1 + x_{s_2} + \cdots + x_{s_t})C \\ = \; & 0. \end{aligned}$$

Since $q(x)$ and $q(xA)$ have identical distributions of values (up to a permutation), $q(xA) = 0$, so is the same for $q(x)$ which is a contradiction to $q(x)$ as a monomial. Thus all $x_{s_1}, x_{s_2}, \cdots, x_{s_t}$ appear in $q(xA)$, which is a contradiction to that $q(xA)$ depends on k variables. So $wt(A_i) \leq k$.

□

Lemma 5.2.1 may not hold when $q(x)$ is not a monomial. For example, when $q(x) = (x_1 + x_3)(x_2 + x_4)$, there is an n by n matrix A with $q(xA) = x_1x_2$. There

exists an i with $1 \leq i \leq 4$ such that $wt(A_i) = 3 > 2$, e.g.,

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Lemma 5.2.2 *If $q(x)$ is a monomial and $q_A(x) = q(xA)$ depends on only k variables, then there is $B \in H_q$ such that $1 \leq col.wt(AB) \leq k$.*

Proof. It is trivial that $col.wt(AB) \geq 1$. Suppose $q(x) = \prod_{i \in I} x_i$ where $I \subset \{1, 2, \cdots, n\}$. Then $q(xA) = \prod_{i \in I}(xA_i)$. Since $q(xA)$ depends on only k variables, from Lemma 5.2.1 we have $wt(A_i) \leq k$. Let C be an invertible matrix with $C_i = A_i$ where $i \in I$ and each of whose remaining $n - |I|$ columns has weight 1. Also, we have

$$q(xC) = \prod_{i \in I}(xA_i) = q(xA).$$

As a result, we have $C = AB$ where $B \in H_q$ and $col.wt(AB) = col.wt(C) \leq k$.

□

We then have the following two types of definitions of q-correlation immune of order k according to S_1 and S_2.

Definition 5.2.3 *A Boolean function $f(x) \in \mathbf{B}_n$ is type* I *q-correlation immune of order k if $\overline{W}_q(f)(A) = 0$ for all $A \in S_1$.*

Definition 5.2.4 *A Boolean function $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k if $\overline{W}_q(f)(A) = 0$ for all $A \in S_2$.*

Theorem 5.2.1 *Let f be a Boolean function in $\mathbf{B}_n$. If $q(x)$ is a monomial and f is type* II *q-correlation immune of order k, then f is also a type* I *q-correlation immune of order k.*

Proof. If $q(x)$ is a monomial and $q_A(x) = q(xA)$ depends on only k variables, from Lemma 5.2.2, we know that there is $B \in H_q$ such that $1 \leq col.wt(AB) \leq k$. Thus, $1 \leq wt_q(A) \leq k$. Hence we have $S_1 \subset S_2$. As a result, if f is type II q-correlation immune of order k, then f is also a type I q-correlation immune of order k.

□

When $q(x) = x_1$, according to the analysis above, $f(x)$ is both type I and type II q-correlation immune of order k if and only if it is correlation immune of order k. So we can view both definitions as generalizations of correlation immunity.

From Definition 5.2.4, we find that the notion depends heavily on $\mathcal{H}_q$, which is not easy to determine. Below we prove an equivalent definition, which however only depends on the weight $col.wt(A)$ of A.

Theorem 5.2.2 *(Equivalent definition). A Boolean function $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k iff $\overline{W}_q(f)(A) = 0$ for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$.*

Proof. For $A \in GL_n$, we have $wt_q(A) \leq col.wt(A)$, so $1 \leq col.wt(A) \leq k$ implies $1 \leq wt_q(A) \leq k$. Thus if $\overline{W}_q(f)(A) = 0$ for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$, then $f(x)$ is type II q-correlation immune of order k.

Conversely, for $A \in GL_n$ with $1 \leq wt_q(A) \leq k$, there is an $H \in \mathcal{H}_q$ such that $col.wt(AH) = wt_q(A) \leq k$. So by (5.2), we get $\overline{W}_q(f)(A) = 0$, i.e., $f(x)$ is type II q-correlation immune of order k.

□

5.3 Equivalent Characterizations for q-correlation Immune Functions

Theorem 5.3.1 *For any Boolean function $q(x) \in \mathbf{B}_n$, a function $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k if and only if $f(x) + q_A(x)$ is balanced for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$.*

The proof of Theorem 5.3.1 follows from the definition of q-transform directly. From Theorem 5.3.1 we can know that if $f(x)$ is q-correlation immune of order k, then it is resistant to q-correlation attacks involving any k LFSR outputs, since the probability $\text{prob}(f(x) = q_A(x)) = 1/2$ for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$.

We would also like to discuss other equivalent statements for q-correlation immune functions.

Definition 5.3.1 *The functions $f(x)$ and $q_A(x)$ are statistically independent if*

$$\text{prob}(f(x) = 1 | q_A(x) = 1) = \text{prob}(f(x) = 1 | q_A(x) = 0) = \text{prob}(f(x) = 1). \quad (5.3)$$

Here we give an example first. Let $q(x) = q(x_1, x_2, x_3) = x_1x_2x_3 + x_1 + x_2 + x_3$. We find that for those A with $col.wt(A) = 1$, we have

$$q_A(x) = q(xA) = q(x).$$

We choose $f(x) = 1$ if $x \in \{(000), (100), (101)\}$ and $f(x) = 0$ otherwise, we can prove $f(x) + q_A(x)$ is balanced, i.e., $f(x)$ is a q-correlation immune function of order 1 by Theorem 5.3.1. However, $f(x)$ and $q_A(x)$ are not statistically independent because eq. (5.3) does not hold. We note that in this example, $q(x)$ is not balanced. But for balanced functions $q(x)$ we have following results.

Theorem 5.3.2 *Let $q(x) \in \mathbf{B}_n$ be a balanced function. Then $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k if and only if $f(x)$ and $q_A(x)$ are statistically*

independent for all $A \in GL_n$ *with* $1 \leq col.wt(A) \leq k$.

To prove Theorem 5.3.2, we need the following lemma. For a Boolean function $g(x)$, the weight of $g(x)$ denoted by $wt(g)$ is $wt(g) = |\{x \in \mathbb{F}_2^n : g(x) = 1\}|$.

Lemma 5.3.1 *For a function* $q(x) \in \mathbf{B}_n$, *we have* $wt(q) = wt(q_A)$ *for all* $A \in GL_n$. *In particular if* $q(x)$ *is balanced, then so is* $q_A(x)$ *for all* $A \in GL_n$.

Proof. The result follows from the fact that xA ranges over $\mathbb{F}_2^n$ as x does. □

Proof of Theorem 5.3.2: First we suppose that $f(x)$ is q-correlation immune of order k. By Theorem 5.3.1 we see that $f(x)+q_A(x)$ is balanced for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$. For such A we let

$$M_{ij} = |\{x \in \mathbb{F}_2^n : f(x) = i, q_A(x) = j\}|, \quad \text{where} \ \ i, j \in \mathbb{F}_2.$$

Then we have $M_{10}+M_{01} = 2^{n-1}$ since $f(x)+q_A(x)$ is balanced and $M_{11}+M_{01} = 2^{n-1}$ since $q_A(x)$ is balanced by Lemma 5.3.1. So we get $M_{11} = M_{10}$ and

$$\text{prob}(f(x) = 1|q_A(x) = 1) = \frac{M_{11}}{2^{n-1}} = \frac{2M_{11}}{2^n} = \frac{M_{11}+M_{10}}{2^n} = \frac{wt(f)}{2^n} = \text{prob}(f(x) = 1).$$

Similarly we have

$$\text{prob}(f(x) = 1|q_A(x) = 0) = \frac{wt(f)}{2^n} = \text{prob}(f(x) = 1),$$

and we complete the proof of the first part.

Now we suppose that $f(x)$ and $q_A(x)$ are statistically independent for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$. For such A, we see that $\text{prob}(q_A(x) = 0) = \text{prob}(q_A(x) =$

1) = 1/2 since $q_A(x)$ is balanced. Then for $b \in \mathbb{F}_2$ we get

$$\begin{aligned}
&\text{prob}(f(x) + q_A(x) = b) \\
&= \sum_{i \in \mathbb{F}_2} \text{prob}(f(x) = i,\ q_A(x) = b - i) \\
&= \sum_{i \in \mathbb{F}_2} \text{prob}(f(x) = i | q_A(x) = b - i) \cdot \text{prob}(q_A(x) = b - i) \\
&= \sum_{i \in \mathbb{F}_2} \text{prob}(f(x) = i) \cdot \text{prob}(q_A(x) = b - i) \\
&= \tfrac{1}{2} \sum_{i \in \mathbb{F}_2} \text{prob}(f(x) = i) = \tfrac{1}{2},
\end{aligned}$$

which indicates that $f(x) + q_A(x)$ is balanced. Hence we complete the proof by Theorem 5.3.1.

□

By a similar proof, we can get the following result.

Theorem 5.3.3 *Let $q(x) \in \mathbf{B}_n$ be a balanced function. Then $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k if and only if* $\text{prob}(q_A(x) = 1 | f(x) = 1) = \text{prob}(q_A(x) = 0 | f(x) = 1) = 1/2$ *for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$.*

5.4 Certain Properties of q-correlation Immune Functions

Proposition 5.4.1 *If $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order $k \geq 1$, then* $\deg(f) = n$ *iff $wt(q)$ is odd.*

Proof. We see that $wt(q_A) = wt(q)$ for all $A \in GL_n$ by Lemma 5.3.1 and $f(x)+q_A(x)$ is balanced for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$ by Theorem 5.3.1. So $wt(q)$ is odd if and only if $wt(f)$ is odd and hence $\deg(f) = n$.

□

Proposition 5.4.2 *Let $C \in GL_n$ with $col.wt(C) = 1$. If $f(x) \in \mathbf{B}_n$ is type* II *q-correlation immune of order k, then so is $f(xC)$.*

Proof. Write $f_C(x) = f(xC)$. We have for $A \in GL_n$

$$\overline{W}_q(f_C)(A) = \overline{W}_q(f)(C^{-1}A).$$

On the other hand, the restriction on $col.wt(C) = 1$ implies that the weight of each column of C is one, so is the weight of each column C^{-1}, i.e., $col.wt(C^{-1}) = 1$. Hence we get

$$col.wt(C^{-1}A) = col.wt(A), \quad A \in GL_n.$$

Now for all $A \in GL_n$ with $1 \leq col.wt(A) \leq k$, we derive $\overline{W}_q(f)(C^{-1}A) = 0$ since $f(x)$ is q-correlation immune of order k. Then we can get the desired result according to the analysis above. □

5.5 Construction of q-correlation Immune Functions

In this section, we discuss the possible techniques for the construction of q-correlation immune functions.

5.5.1 A General Construction

Here we give a general construction of type II q-correlation immune functions of order k. Let $1 \leq r < n$. Suppose that $q(x) \in \mathbf{B}_n$ depends only on $x_1, \ldots, x_r$. For $A \in GL_n$ with $col.wt(A) \leq k$, we see that $q_A(x)$ depends on at most kr coordinates.

Let $f(x) = f_1(x) + f_2(x) \in \mathbf{B}_n$, where $f_1(x) \in \mathbf{B}_n$ is a linear function that depends on at least $kr + 1$ coordinates and $f_2(x) \in \mathbf{B}_n$ (possibly nonlinear) depends only on the complement of the support of $f_1(x)$. Here the *support* of $f_1(x)$ means the set of variables appearing in the algebraic normal form of $f_1(x)$.

It follows that $f(x) + q_A(x)$ has the form $g(x) + x_i$ for some $g \in \mathbf{B}_n$ and $1 \leq i \leq n$ such that $g(x)$ does not depend on x_i. Therefore $f(x) + q_A(x)$ is balanced, so

$\overline{W}_q(f)(A) = 0$ for all $A \in GL_n$ with $col.wt(A) \leq k$. This means f is q-correlation immune of order k by Theorem 5.2.2.

We find that from the construction above

$$\deg(f) \leq n - (kr + 1).$$

However, there do exist $f(x)$ and $q(x)$ such that $f(x)$ is q-correlation immune of order n and $\deg(f) = n$. For example, let $q(\overline{0}) = 0$ and $wt(q) = 2^{n-1} - 1$. We have $q_A(\overline{0}) = 0$ and $wt(q_A) = 2^{n-1} - 1$ for all $A \in GL_n$. Then both $f(x)$ and $1 + f(x)$ satisfying

$$f(x) = \begin{cases} 1, & \text{if } \ x = \overline{0}, \\ 0, & \text{otherwise}, \end{cases}$$

are q-correlation immune of order n.

5.5.2 Construction Based on Linear Codes

Let $\mathbb{F}$ be a finite field of size p. An $[n, k, d]$ linear code $\mathcal{C}$ is a linear subspace of $\mathbb{F}^n$ of dimension k and with minimum distance d, i.e., the minimum Hamming weight of its nonzero code words is d. For every two codewords $c_1, c_2 \in \mathcal{C}$, we have $a_1c_1 + a_2c_2 \in \mathcal{C}$ where $a_1, a_2 \in \mathbb{F}$. Every basis of a linear $[n, k, d]$ code consists of k codewords. Therefore the size of $\mathcal{C}$ is p^k. A generator matrix of an $[n, k, d]$ linear code over $\mathbb{F}$ is a $k \times n$ matrix, typically denote by G. The rows of G form a basis of the code and G is not unique for a given linear code. The rank of a generator matrix $\mathcal{C}$ equals the dimension of $\mathcal{C}$.

Wu and Dawson [64] investigated correlation immunity based on linear codes as shown in Lemma 5.5.1. We next generalize their construction to type I q-correlation immunity.

Lemma 5.5.1 *[64] If G is a generating matrix of an $[n, k, d]$ linear code, then for any $g(y) \in \mathbf{B}_k$, the correlation immunity of $f(x) = g(xG^T)$ is at least $d - 1$.*

Lemma 5.5.2 *Let $q(x) = \prod_{j=1}^{m} x_{i_j}$. Let $A = (A_1|A_2|\cdots|A_n)$ be an n by n matrix. If $q(xA)$ depends on at most k variables, then*

$$|supp(xA_{i_1}) \cup supp(xA_{i_2}) \cup \cdots \cup supp(xA_{i_m})| \leq k.$$

Proof. Let $wt(A_s) = t$ where $s \in \{i_1, i_2, \cdots, i_m\}$ and $xA_s = x_{s_1} + x_{s_2} + \cdots + x_{s_t}$. By Lemma 5.2.1 we know that $x_{s_1}, x_{s_2}, \cdots, x_{s_t}$ all appear in $q(xA)$. As a result, if $q(xA)$ depends on at most k variables, then

$$|supp(xA_{i_1}) \cup supp(xA_{i_2}) \cup \cdots \cup supp(xA_{i_m})| \leq k.$$

□

Theorem 5.5.1 *Let $q(x)$ be a monomial depending on m variables. In other words, $q(x) = x_{i_1}x_{i_2}\cdots x_{i_m} \in \mathbf{B}_n$ where $0 \leq m \leq n$. Let C be an $[n, k, d \geq 2]$ code with generator matrix G. Let $g \in \mathbf{B}_k$ be nondegenerate and let $f(x) = g(xG^{tr}) \in \mathbf{B}_n$. Then f is type* I *q-correlation immune of order $d - 1$.*

Proof. Let $A = (A_1|A_2|\cdots|A_n)$, we have

$$q(xA) = (xA_{i_1})(xA_{i_2})\cdots(xA_{i_m}).$$

Let

$$\omega(A_{i_1}, A_{i_2}, \cdots, A_{i_m}) = |supp(xA_{i_1}) \cup supp(xA_{i_2}) \cup \cdots \cup supp(xA_{i_m})|.$$

From Lemma 5.5.2 we know that f is type I q-correlation immune of order t if for all the linearly independent $A_{i_1}, A_{i_2}, \cdots, A_{i_m}$ such that

$$\omega(A_{i_1}, A_{i_2}, \cdots, A_{i_m}) \leq t,$$

we have

$$rank(G^{tr}|A_{i_1}^{tr}|A_{i_2}^{tr}|\cdots|A_{i_m}^{tr}) = k + m. \tag{5.4}$$

Let $H = \{\sum_{j=1}^{m} a_j A_{i_j} | a_j \in \mathbb{F}_2\}$. Eq. (5.4) is equivalent to saying that $H \cap C = \{0^n\}$.

Let e be the minimum value of $\omega(A_{i_1}, A_{i_2}, \cdots, A_{i_m})$ over all $A_{i_1}, A_{i_2}, \cdots, A_{i_m}$ that are linearly independent and satisfy $H \cap C \neq \{0^n\}$.

We know that

$$\omega(A_{i_1}, A_{i_2}, \cdots, A_{i_m}) \geq \max\{wt(h), h \in H\}.$$

If $H \cap C \neq \{0^n\}$, there exists a nonzero $h^{'} \in H \cap C$. Since $wt(h^{'}) \geq d$, then

$$\omega(A_{i_1}, A_{i_2}, \cdots, A_{i_m}) \geq d.$$

Thus $e \geq d$. Then f is q-correlation immune of order $d - 1$. □

When $q(x)$ is linear, $f(x)$ is both type I and type II q-correlation immune of order k if and only if it is correlation immune of order k. At the end of this section, we discuss some relations between correlation immune functions and q-correlation immune functions based on linear codes when $q(x)$ is nonlinear.

Lemma 5.5.3 *Let G be the generating matrix of an $[n, 2, d]$ linear code where $d \geq 4$ and let the transpose of generating matrix $G^T = [g_1, g_2]$. Let $g(y) = y_1 y_2$ and $f(x) = g(xG^T) = (xg_1)(xg_2) = (\bigoplus_{i=1}^{m} x_{s_i})(\bigoplus_{j=1}^{n} x_{t_j})$. If $q(x) = x_c x_d$ where $c, d \in$*

$\{1, 2, \cdots, n\} \setminus \{s_1, s_2, \cdots, s_m, t_1, t_2, \cdots, t_n\}$. *Then the correlation immunity of* $f(x)$ *is at least* 3 *and* $f(x) + q(x)$ *is unbalanced.*

Proof. The correlation immunity of $f(x) = g(xG^T)$ is at least $d-1 \geq 3$ from Lemma 5.5.1. We also have that xg_1^T, xg_2^T, x_c, x_d are linearly independent. As a result, the rank of $f(x) + q(x) = (xg_1^T)(xg_2^T) + x_c x_d$ is 4. According to the classification of quadratic forms in [30], $f(x) + q(x)$ is unbalanced.

□

From Lemma 5.5.3, we can get a class of Boolean function $f(x)$ which is correlation immune of order higher than 3 but is not type I q-correlation immune of order 2.

Let's consider the case when $f(x)$ is q-correlation immune of order k but is not correlation immune of order k. Experimental results show the existence of such $f(x)$. For example, $f(x) = x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4$ is q-correlation immune of order 2 but is not correlation immune of order 2 over 4 variables.

5.6 Concluding Remarks

In this chapter, we present a new correlation attack by exploring the correlation between the output of a nonlinear function of several LFSRs and the output of the nonlinear combination function in the generator. To resist this attack, we propose the idea of q-correlation immune functions. We give two definitions of q-correlation immune functions. Certain properties and possible constructions are discussed.

Chapter 6 Future Work

This dissertation presents research work on distribution properties of N-ary half-ℓ-sequences with odd prime connection integers. One can generalize the definition of half-ℓ-sequences with prime power connection integers. In the future, I plan to explore more on some statistical properties of generalized half-ℓ-sequences. This dissertation discusses a new type of correlation attack and q-correlation immune functions. In this chapter, we outline the research directions in these areas and the topics that we are interested to work on in the future.

6.1 Half-ℓ-sequences with Prime Power Connection Integers

As mentioned above, we would like to investigate some statistical properties of generalized half-ℓ-sequences with prime power connection integers. Let $q = p^m$ with p an odd prime and $m \geq 1$. We extend the definition of half-ℓ-sequences in the following way.

Definition 6.1.1 *A sequence* $\mathbf{s}$ *with prime power connection integer* q *is called a half-*ℓ*-sequence if the period of* $\mathbf{s}$ *is* $\phi(q)/2$.

Indeed, when $m = 1$, $\mathbf{s}$ is the half-ℓ-sequence we discussed in Chapter 3.

Let ξ be a complex primitive qth root of unity. In Chapter 3, we need a bound on

$$\sum_{z \in Q} \xi^{bz}$$

where $b \not\equiv 0 \pmod{q}$ when q is an odd prime. Now we extend the result to the case when q is a prime power by using the following lemma.

Lemma 6.1.1 *[19] Let f be a polynomial over $\mathbb{Z}$ of degree $d \geq 1$ and $d_p \equiv d \pmod{p}$. Then for any prime p with $d_p \geq 1$ and any $m \geq 1$ we have*

$$\left| \sum_{x=1}^{p^m} \xi^{f(x)} \right| \leq 3(d-1)p^{m(1-1/d)}.$$

In particular,

$$\begin{aligned}
\left| \sum_{z \in Q} \xi^{bz} \right| &= \frac{1}{2} \left| \sum_{c \in \mathbb{Z}_{p^m}^*} \xi^{bc^2} \right| = \frac{1}{2} \left| \sum_{c=1}^{p^m} \xi^{bc^2} - \sum_{d=1}^{p^{m-1}} \xi^{c(pd)^2} \right| \\
&\leq \frac{1}{2} \left(\left| \sum_{c=1}^{p^m} \xi^{bc^2} \right| + \left| \sum_{d=1}^{p^{m-1}} (\xi^{p^2})^{cd^2} \right| \right) \\
&= \frac{1}{2} \left(\left| \sum_{c=1}^{p^m} \xi^{bc^2} \right| + p \left| \sum_{d=1}^{p^{m-2}} (\xi^{p^2})^{cd^2} \right| \right) \\
&\leq \frac{1}{2} \left(3p^{m/2} + p \cdot 3p^{(m-2)/2} \right) = 3p^{m/2}.
\end{aligned}$$

where $b \not\equiv 0 \pmod{p^m}$ and Q is the set of quadratic residues modulo p^m.

The Fourier transform of a complex valued function $f : \mathbb{Z}_{p^m} \to \mathbb{C}$ is given by

$$\hat{f}(b) = \frac{1}{p^m} \sum_{c=0}^{p^m-1} f(c) \xi^{-bc}.$$

By the Fourier inversion formula we have

$$f(c) = \sum_{b=0}^{p^m-1} \hat{f}(b) \xi^{bc}.$$

In the future, we plan to investigate the following problems related to some statistical properties of half-ℓ-sequences with prime power connection integers.

- The number of occurrences of one symbol within one period of a half-ℓ-sequence;
- The number of pairs of symbols with a fixed distance between them within one

period of a half-ℓ-sequence;

- The number of triples of consecutive symbols within one period of a half-ℓ-sequence;
- The autocorrelation of a half-ℓ-sequence.

6.2 Problems Related to q-transform

In this dissertation, we present a new correlation attack by exploiting the correlation between the output of a nonlinear function of several LFSRs and the output of the nonlinear combination function in the generator. We build a statistical model for this attack and describe the steps of this attack. Much work is needed for the analysis of this attack. In the future, we plan to work in the following directions.

- The data complexity or the number of keystream bits needed for the the attack;
- The success rate of this attack;
- The linear bias plays an important role in the analysis of data complexity and success rate. We are interested in finding efficient methods to obtain a nonlinear function such that the linear bias reaches maximum value;
- Analysis of the new correlation attack on existing LFSR based stream ciphers.

To resist the new correlation attack, we propose the idea of q-correlation immune functions based on q-transform. Certain properties are discussed for q-correlation immune functions. We also discuss the construction of q-correlation immune functions when $q(x)$ is a monomial. We plan to pursue more results on the construction of q-correlation immune functions when $q(x)$ is a polynomial in the future. We are also interested in counting the number of such functions.

6.3 Design of Stream Ciphers based on FCSRs

While there are several stream ciphers based on LFSRs, F-FCSR stream ciphers in eSTREAM project [55] are the first popular FCSR based stream ciphers that arouse much attention. The F-FCSR stream cipher uses an FCSR in Galois mode and takes a linear combination of the state bits to produce output. This stream cipher is extremely fast due to the very simple output function. It was initially in the eSTREAM portfolio, but was subsequently broken by Hell and Johansson [35] due to the linearity of its filter function. Nonetheless, in his plenary talk at SETA 2012, Johansson said he believes FCSRs have an important role as building blocks for future stream ciphers. I plan to replace the linear output function with a nonlinear output function, which generate equivalent sequence for F-FCSR and work on a more complex FCSR based construction to design secure stream ciphers.

Bibliography

[1] V. Anashin, A. Bogdanov, I. Kizhvatov, and S. Kumar. ABC: A new fast flexible stream cipher. *eSTREAM, ECRYPT Stream Cipher Project Report*, Report 2005/001, 2005.

[2] F. Arnault and T. Berger. F-FCSR: Design of a new class of stream ciphers. In H. Gilbert and H. Handschuh, editors, *Fast Software Encryption*, volume 3557 of *LNCS*, pages 83–97. Springer Berlin Heidelberg, 2005.

[3] F. Arnault, T. Berger, and C. Lauradoux. F-FCSR stream ciphers. In *New Stream Cipher Designs*, pages 170–178. Springer, 2008.

[4] F. Arnault, T. Berger, and A. Necer. A new class of stream ciphers combining LFSR and FCSR architectures. In A. Menezes and P. Sarkar, editors, *Progress in Cryptology—INDOCRYPT 2002*, volume 2551 of *LNCS*, pages 22–33. Springer Berlin Heidelberg, 2002.

[5] F. Arnault, T. P. Berger, and A. Necer. Feedback with carry shift registers synthesis with the euclidean algorithm. *IEEE Transactions on Information Theory*, 50(5):910–917, 2004.

[6] C. Berbain, O. Billet, A. Canteaut, N. Courtois, H. Gilbert, L. Goubin, A. Gouget, L. Granboulan, C. Lauradoux, M. Minier, et al. Sosemanuk, a fast software-oriented stream cipher. In *New stream cipher designs*, pages 98–118. Springer, 2008.

[7] E. R. Berlekamp. *Algebraic Coding Theory: Revised Edition*. World Scientific, 2015.

[8] B. S. BluetoothTM. version 1.2, november 2003.

[9] S. Bokhari. Multiprocessing the Sieve of Eratosthenes. *IEEE Computer*, 20(4), April 1986.

[10] A. Braeken, J. Lano, N. Mentens, B. Preneel, and I. Verbauwhede. SFINKS: A synchronous stream cipher for restricted hardware environments. In *SKEW - Symmetric Key Encryption Workshop*, April 2005.

[11] M. Briceno, I. Goldberg, and D. Wagner. A pedagogical implementation of a5/1, 1999.

[12] Y. Cai and C. Ding. Binary sequences with optimal autocorrelation. *Theoretical Computer Science*, 410(24):2316–2322, 2009.

[13] A. Canteaut. Open problems related to algebraic attacks on stream ciphers. In Ø. Ytrehus, editor, *Coding and Cryptography*, volume 3969 of *LNCS*, pages 120–134. Springer Berlin Heidelberg, 2006.

[14] A. Canteaut and M. Trabbia. Improved fast correlation attacks using parity-check equations of weight 4 and 5. In B. Preneel, editor, *EUROCRYPT 2000*, volume 1807 of *LNCS*, pages 573–588, Berlin, Germany, 2000. Springer-Verlag.

[15] C. Carlet. A larger class of cryptographic Boolean functions via a study of the Maiorana-McFarland construction. In M. Yung, editor, *Advances in Cryptology-CRYPTO 2002*, volume 2442 of *LNCS*, pages 549–564, Berlin, Germany, 2002. Springer-Verlag.

[16] C. Carlet, D. K. Dalai, K. C. Gupta, and S. Maitra. Algebraic immunity for cryptographically significant Boolean functions: analysis and construction. *IEEE Transactions on Information Theory*, 52(7):3105–3121, July 2006.

[17] C. Carlet, P. Guillot, and S. Mesnager. On immunity profile of Boolean functions. In H.-Y. S. Guang Gong, Tor Helleseth and K. Yang, editors, *Sequences and Their Applications–SETA 2006*, volume 4086 of *LNCS*, pages 364–375. Springer, 2006.

[18] T. Cochrane. On a trigonometric inequality of Vinogradov. *Journal of Number Theory*, 27(1):9–16, 1987.

[19] T. Cochrane and Z. Zheng. Pure and mixed exponential sums. *Acta Arith*, 91(3):249–278, 1999.

[20] I. C. S. L. M. S. Committee et al. Wireless lan medium access control (mac) and physical layer (phy) specifications, 1997.

[21] N. T. Courtois. General principles of algebraic attacks and new design criteria for cipher components. In V. R. H. Dobbertin and A. Sowa, editors, *Advanced Encryption Standard–AES*, volume 3373, pages 67–83. Springer, 2004.

[22] N. T. Courtois and W. Meier. Algebraic attacks on stream ciphers with linear feedback. In *Proceedings of the 22nd International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT'03, pages 345–359, Berlin, Heidelberg, 2003. Springer-Verlag.

[23] R. Couture and P. L'Écuyer. Distribution properties of multiply-with-carry random number generators. *Mathematics of Computation of the American Mathematical Society*, 66(218):591–607, 1997.

[24] L. Crypto. Yamb specification and source code. Technical report, April 2005.

[25] T. W. Cusick and P. Stanica. *Cryptographic Boolean functions and applications*. Academic Press, 2009.

[26] C. Ding, G. Xiao, and W. Shan. *The Stability Theory of Stream Ciphers*, volume 561 of *LNCS*. Springer-Verlag, 1991.

[27] P. Ekdahl and T. Johansson. A new version of the stream cipher SNOW. *In SAC '02: Revised Papers from the 9th Annual International Workshop on Selected Areas in Cryptography*, pages 47–61, 2003.

[28] S. Golomb. *Shift Register Sequence*. Aegean Park Press, Laguna Hills,CA, revised edition edition, 1982.

[29] M. Goresky and A. Klapper. Periodicity and distribution properties of combined FCSR sequences. In G. Gong, T. Helleseth, H.-Y. Song, and K. Yang, editors, *Sequences and Their Applications – SETA 2006*, volume 4086 of *Lecture Notes in Computer Science*, pages 334–341. Springer Berlin Heidelberg, 2006.

[30] M. Goresky and A. Klapper. *Algebraic Shift Register Sequence*. Cambridge University Press, April 2012.

[31] M. Goresky, A. Klapper, et al. Arithmetic crosscorrelations of feedback with carry shift register sequences. *IEEE Transactions on Information Theory*, 43(4):1342–1345, 1997.

[32] M. Goresky and A. M. Klapper. Fibonacci and galois representations of feedback-with-carry shift registers. *IEEE Transactions on Information Theory*, 48(11):2826–2836, Nov 2002.

[33] T. Gu and A. Klapper. Distribution properties of half-ℓ-sequence. In K.-U. Schmidt and A. Winterhof, editors, *Sequences and Their Applications-SETA 2014*, pages 234–245. Springer, 2014.

[34] J. Hastad, J. Mattsson, and M. Naslund. The stream cipher Polar Bear, April 2005.

[35] M. Hell and T. Johansson. Breaking the F-FCSR-H stream cipher in real time. In J. Pieprzyk, editor, *Advances in Cryptology-ASIACRYPT 2008*, volume 5350 of *LNCS*, pages 557–569. Springer Berlin Heidelberg, 2008.

[36] H. Imai and A. Yamagishi. CRYPTREC Project Cryptographic Evaluation Project for the Japanese Electronic Government. In T. Okamoto, editor, *Advances in Cryptology — ASIACRYPT 2000*, volume 1976 of *LNCS*, pages 399–400. Springer Berlin Heidelberg, 1976.

[37] T. Johansson and F. Jonsson. Fast correlation attacks through reconstruction of linear polynomials. In M. Bellare, editor, *Advances in Cryptology-CRYPTO 2000*, volume 1880 of *LNCS*, pages 300–315, Berlin, Germany, 2000. Springer-Verlag.

[38] A. Klapper. A new transform related to distance from a boolean function. In K.-U. Schmidt and A. Winterhof, editors, *Sequences and Their Applications-SETA 2014*, pages 47–59. Springer, 2014.

[39] A. Klapper and M. Goresky. Feedback shift registers, combiners with memory, and arithmetic codes. Tech. Rep. No. 239-93., Univ. of Kentucky Dept. of Comp. Sci., 1993.

[40] A. Klapper and M. Goresky. 2-adic shift registers. In R. Anderson, editor, *Fast Software Encryption*, volume 809 of *LNCS*, pages 174–178. Springer Berlin Heidelberg, 1994.

[41] A. Klapper and M. Goresky. Large period nearly deBruijn FCSR sequences. In L. C. Guillou and J.-J. Quisquater, editors, *Advances in Cryptology—EUROCRYPT'95*, volume 921 of *LNCS*, pages 263–273. Springer Berlin Heidelberg, 1995.

[42] A. Klapper and M. Goresky. Feedback shift registers, 2-adic span, and combiners with memory. *Journal of Cryptology*, 10:111–147, 1997.

[43] D. H. Lee and S. Park. Word-based FCSRs with fast software implementations. *Journal of Communications and Networks*, 13(1):1–5, 2011.

[44] R. Lidl and H. Niederreiter. *Finite Fields*, volume 20. Cambridge university press, 1997.

[45] G. Marsaglia. The mathematics of random number generator. In S. Burr, editor, *The unreasonable effectiveness of number theory*, pages 73–90. American Mathematical Society, 1992.

[46] G. Marsaglia. Yet another rng. *Posted to the electronic billboard sci. stat. math, August*, 1, 1994.

[47] G. Marsaglia and A. Zaman. A new class of random number generators. *Annals of Applied Probability*, 1(3):462–480, 1991.

[48] A. Menezes, P. van Oorschot, and S. Vanstone. *Handbook of Applied Cryptography*. CRC Press, `www.cacr.math.uwaterloo.ca/hac`, 1996.

[49] T. Nagell. Introduction to number theory. *New York*, 1951.

[50] Y. Nawaz and G. Gong. The WG stream cipher. Technical report, University of Waterloo, 2005.

[51] J. Neyman and E. S. Pearson. *On the problem of the most efficient tests of statistical hypotheses*. Springer, 1992.

[52] B. Preneel, A. Biryukov, E. Oswald, B. Rompay, L. Granboulan, E. Dottax, S. Murphy, A. Dent, J. White, M. Dichtl, et al. Nessie security report. *Deliverable D20, NESSIE Consortium. Feb*, 2003.

[53] W. Qi and H. Xu. Partial period distribution of FCSR sequences. *IEEE Transactions on Information Theory*, 49(3):761–765, 2003.

[54] R. L. Rivest. The RC4 encryption algorithm. *RSA Data Security Inc*, 1992.

[55] M. Robshaw. The eSTREAM project. In *New Stream Cipher Designs*, pages 1–6. Springer, 2008.

[56] R. Rueppel. *Analysis and design of stream ciphers*. Springer-Verlag, Berlin, 1986.

[57] P. Sarkar and S. Maitra. Nonlinearity bounds and construction of resilient Boolean functions. In M. Bellare, editor, *Advances in Cryptology-CRYPTO 2000*, volume 1880 of *LNCS*, pages 515–532, Berlin, Germany, 2000. Springer-Verlag.

[58] B. Schneier. *Applied cryptography: protocols, algorithms, and source code in C*. John Wiley & Sons, 2007.

[59] T. Siegenthaler. Decrypting a class of stream ciphers using ciphertext only. *IEEE Transactions on Computers*, C-34(1):81–85, 1985.

[60] SIG Bluetooth. *Bluetooth specification*, May 1, 2007, 2003.

[61] T. Tian and W.-F. Qi. Autocorrelation and distinctness of decimations of ℓ-sequences. *SIAM Journal on Discrete Mathematics*, 23(2):805–821, 2009.

[62] I. M. Vinogradov. *Elements of number theory*. Courier Corporation, 2003.

[63] Q. Wang and C. H. Tan. New bounds on the imbalance of a half-ℓ-sequence. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2688–2691. IEEE, 2015.

[64] C.-K. Wu and E. Dawson. Construction of correlation immune boolean functions. *Australasian Journal of Combinatorics*, 21:141–166, 2000.

[65] G. Xiao and J. Massey. A spectral characterization of correlation-immune combining functions. *IEEE Transactions on Information Theory*, 34(3):569–571, 1988.

[66] H. Xu and W.-F. Qi. Autocorrelations of maximum period FCSR sequences. *SIAM Journal on Discrete Mathematics*, 20(3):568–577, 2006.

[67] N. Y. Yu and G. Gong. Crosscorrelation properties of binary sequences with ideal two-level autocorrelation. In G. Gong, T. Helleseth, H.-Y. Song, and K. Yang, editors, *Sequences and Their Applications – SETA 2006*, volume 4086 of *LNCS*, pages 104–118. Springer Berlin Heidelberg, 2006.

Vita

Author's Name: Ting Gu

Place of Birth: Anlu, Hubei, China

Education:

Central China Normal University, Wuhan, China

B.S. awarded June 2007

M.E. awarded June 2010

Professional Position:

Research assistant, Teaching assistant, University of Kentucky

Selective Publications:

Gu, Ting, and Andrew Klapper. Statistical Properties of Half-ℓ-Sequences, to appear in Journal of Cryptography and Communications - Discrete Structures, Boolean Functions and Sequences.

Gu, Ting, and Andrew Klapper. Distribution Properties of Half-ℓ-Sequence, Sequences and Their Applications - SETA 2014. Springer International Publishing, 2014. 234-245.

Academic Service:

Reviewer for IEEE Transactions on Information Theory, 2015

Reviewer for IEEE Transactions on Information Forensics & Security, 2015, 2016

Reviewer for Signal Processing: Image Communication, 2015