Simultaneous correlation to many linear functionals : a new cryptanalytic technique which can almost halve the effective key size of certain stream ciphers

Matthew W. Dodd

Racal Comsec Limited, Salisbury, Wiltshire
matthew@mdodd.net

1 Introduction

Since Siegenthaler published his well-known papers dealing with correlation attacks on stream ciphers, much research has focused on his idea of modelling keystream produced by linearly clocking stream ciphers as noisy shift register sequences. That this model arises so often in practise is due to the fact that any function $GF(2^n) \rightarrow GF(2)$ is correlated to some *linear* function $GF(2^n) \rightarrow GF(2)$. The observation that there is generally simultaneous correlation to many such linear functionals leads one to consider launching attacks using correlation to functionals in vector spaces of various sizes.

Now a function $f : GF(2^n) \rightarrow GF(2)$ is completely characterised by its correlation to all functionals. It turns out that the correlations to the elements of subspaces of functionals characterise "reduced" (generally non-deterministic) versions of f, such as those obtained when random bits provide some of the inputs. As a consequence, attacks based on correlations to spaces of linear functionals turn out to be the same as attacks attempting to determine a most likely key, or component of a key, given that the keystream was generated on the (generally) "non-deterministic" keystream generator with "reduced" output function or "reduced" key space rather than the original one. A more precise formulation of what we mean by such maximum likelihood attacks will be the starting point of this paper.

2 Maximum likelihood attacks on stream ciphers

2.1 The keystream generator

Throughout this paper we shall be considering a keystream generator with state space V, state transition function $T : V \rightarrow V$, and output function $f : V \rightarrow \{0, 1\}$. Later on, we shall require that f is (approximately) balanced — that is, it takes the values 0 and 1 (approximately) equally often : this will be true of any good keystream generator used to protect privacy. We shall identify V with the vector space of n-bit vectors over $GF(2)$.

The generator is initially loaded with a key, so that the keystream bits $b_1, \ldots, b_N$ it produces satisfy

$$b_i = f(T^i x) \qquad (i = 1, \ldots, N) \tag{1}$$

when x assumes the value of the key.

2.2 Maximum likelihood attack

First of all, a definition : for any function $g : V \rightarrow R$ and subspace $S \leq V$, we define a "reduced" version, $\overline{g}_S$, of g, by taking, for each $x \in V$, $\overline{g}_S(x)$ to be the random variable which assumes the values 0 and 1 with the probability that g assumes that value on a uniformly selected element

of the coset x+$S^{\perp}$($S^{\perp}$is the space of vectors of V whose inner product with every element of S is 0).

Now we observe that for any subspace $U \leq V$, the equations (1) can be reduced to

$$b_i = \overline{f}_U(T^i x) \qquad (i = 1, \ldots, N), \tag{2}$$

corresponding to a reduction of the output function, or to

$$b_i = \overline{f \circ T^i}_U(x) \qquad (i = 1, \ldots, N), \tag{3}$$

corresponding to a reduction of the key space; we will also use a common symbolism

$$b_i = \overline{g}_i(x) \qquad (i = 1, \ldots, N). \tag{4}$$

For either of these reductions we may attempt to determine the most likely $x \in V$ given the probabilistic equations hold; that is, maximise

$$\begin{aligned} &\Pr(x|\overline{g}_i(x) = b_i \,\forall i \in \{1, \ldots, N\}) \\ &= \Pr(\overline{g}_i(x) = b_i \,\forall i|x) \Pr(x)/\Pr(b_1, \ldots, b_N). \end{aligned} \tag{5}$$

Thus for equiprobable initial states and given keystream, x equivalently maximises

$$\Pr(\overline{g}_i(x) = b_i \,\forall i) = \prod_{i=1}^{N} \Pr(\overline{g}_i(x) = b_i).$$

In the case of reduction (2), this probability can be computed, for each value of x, in time N. The reduction (3) is in general less tractable, but we shall see later that this is not so when state transition is linear.

In section 6 we shall use a reformulation of the maximum likelihood condition, which makes use of the limit $\ln(1 + x) = x + O(x^2)$ as $x \to 0$ in the case when all $\Pr(\overline{g}_i(x) = 0) \approx \frac{1}{2}$: in this case, maximising (5) is equivalent to maximising

$$\begin{aligned} &\ln\left((\tfrac{1}{2})^{-N} \prod_{i=1}^{N} \Pr(\overline{g}_i(x) = b_i)\right) \\ &= \sum_{i=1}^{N} \ln(\Pr(\overline{g}_i(x) = b_i)/\tfrac{1}{2}) \\ &\approx \sum_{i=1}^{N} (2 \Pr(\overline{g}_i(x) = b_i) - 1) \\ &= \sum_{i=1}^{N} E((-1)^{\overline{g}_i(x) \oplus b_i}). \end{aligned} \tag{6}$$

2.3 Uniqueness of maximum likelihood solutions

In this section we observe that the reduced equations (2) or (3) may not yield a unique most likely solution x. This is apparent in the case of (3), where x can only possibly be determined up to a coset $x + U^{\perp}$.

To see that ambiguity is possible also for (2), observe that we can have a subspace W of V containing U such that T is well-defined on cosets $x + W^{\perp}$. (Such a W will usually arise from a decomposition of the keystream generator states into those of two sub-generators : that is, up to a reordering of positions in a vector of V, we have a cartesian product $V = W \times W^{\perp}$ (identifying W with the subspace $\{(w, 0) : w \in W\}$, and similarly for $W^{\perp}$) for W (and $W^{\perp}$) closed under T.) Now $\overline{f}_U$ the same distribution on all elements in a coset $x + U^{\perp} \supseteq x + W^{\perp}$, and hence is well-defined on cosets $x + W^{\perp}$. These properties of T and $\overline{f}_U$ together imply that (2) can only determine x up to a coset $x + W^{\perp}$.

2.4 How large N should be

We now address the question : how large need N be in order that the most likely coset of solutions to equations (4) is the correct one, i.e. is the coset of the initial state of the generator? We answer this by way of a corollary to the following theorem.

A theorem of Brynielsson [1]

Theorem 1. *Let $\underline{a}$ and $\underline{b}$ be two distributions on K objects, taking values with probabilities a_j and b_j $(j = 1, \ldots, K)$ respectively; X be a uniform random variable on $\{1, \ldots, M\}$; Y_i $(i = 1, \ldots, M)$ be independent random variables having the multinomial distribution $M(N, \underline{a})$ for $i = X$, but the multinomial distribution $M(N, \underline{b})$ for $i \neq X$; and, lastly, y_i be an observation of Y_i $(i = 1, \ldots, M)$. Denote by $d(\underline{a}, \underline{b})$ the directed divergence $\sum_{j=1}^{K} a_j \log(a_j/b_j)$, and by p_i $(i = 1, \ldots, M)$ the probability*

$$\Pr(X = i | Y_1 = y_1, \ldots, Y_M = y_M).$$

Then the ordering on the i induced by the p_i is the same as that induced by the likelihood ratio

$$\frac{\Pr(y_i \text{ is an observation of } M(N, \underline{a}))}{\Pr(y_i \text{ is an observation of } M(N, \underline{b}))},$$

and the probability that X is amongst the k greatest values of i under this ordering (for any k such that $K \ll \log(M/k)/d(\underline{a}, \underline{b})$) is approximately

$$\begin{cases} 0 \text{ if } N < \log(M/k)/d(\underline{a}, \underline{b}) \\ 1 \text{ if } N > \log(M/k)/d(\underline{a}, \underline{b}) \end{cases}$$

(for large M).

A corollary

Corollary 1. *Suppose that we have an array $\overline{g}_i(x)$ $(i = 1, \ldots, N; x \in V)$ of Bernoulli random variables, taking value 0 with probabilities $p_{i,x}$, for which (for any i) $\overline{g}_i(x)$ is the same distribution as $\overline{g}_i(y)$ when x and y are in the same coset, but otherwise are independent distributions; the $\overline{g}_i(x)$ (for fixed x) are independent; the parameters $p_{i,x}$ are themselves (for pairs of i and pairs of x in different cosets) known independent realisations of some* a priori *distribution D with mean $\frac{1}{2}$; and $(b_i)_{i=1}^N$ is an observation of $(\overline{g}_i(x))_{i=1}^N$ for (any) x in some particular coset — that of x_0, say. Then, for large number 2^s of cosets, the maximum likelihood method determines the coset of x_0 after $N \approx s/\overline{I}(\overline{g}_i(x); x)$ observations of $\overline{g}_i(x_0)$, where $\overline{I}(\overline{g}_i(x); x)$ denotes the average mutual information (averaged over D) between $\overline{g}_i(x)$ and x.*

Proof. If necessary, replace D by an approximation taking

$$K \ll s/\overline{I}(\overline{g}_i(x); x)$$

values, and make appropriate approximations in what follows.

Fix any x. For each i, there are $2K$ possibilities :

$$b_i = b \text{ and } \Pr(\overline{g}_i(x) = 0) = p,$$

for $b = 0, 1$ and each probability p associated with D, and these possibilities themselves have probabilities pairwise independent for distinct indices i. The sequence $y_x = (N_{x;b,p})_{b=0,1;p \in D}$, where

$$N_{x;b,p} = \ \# \text{ of times } b_i = b \text{ and } p_{i,x} = p,$$

is an observation of the multinomial distribution on N observations of independent events with these $2K$ probabilities.

For $x \in$ the coset of x_0, these $2K$ probabilities are

$$\begin{aligned}\Pr(b_i = 0, p_{i,x} = p) &= p \Pr(p_{i,x} = p)\\ \Pr(b_i = 1, p_{i,x} = p) &= (1-p)\Pr(p_{i,x} = p),\end{aligned}$$

defining a distribution $\underline{a}$, while for other x

$$\begin{aligned}&\Pr(b_i = b, p_{i,x} = p)\\ &= \Pr(b_i = b)\Pr(p_{i,x} = p) \qquad \text{by independence}\\ &= \tfrac{1}{2}\Pr(p_{i,x} = p) \qquad \text{by balance,}\end{aligned}$$

defining a distribution $\underline{b}$.

Writing, for convenience, $\Pr(p)$ for $\Pr(p_{i,x} = p)$, the directed divergence $d(\underline{a}, \underline{b})$ is

$$\begin{aligned}&\left(\sum_p p\Pr(p)\log\frac{p\Pr(p)}{\frac{1}{2}\Pr(p)}\right) + \left(\sum_p (1-p)\Pr(p)\log\frac{(1-p)\Pr(p)}{\frac{1}{2}\Pr(p)}\right)\\ &= \sum_p \Pr(p)(p\log(p/\tfrac{1}{2}) + (1-p)\log((1-p)/\tfrac{1}{2}))\\ &= \overline{I}(\overline{g}_i(x); x).\end{aligned}$$

Thus by theorem 1,

$$\frac{\Pr(y_x \text{ is an observation of } M(N, \underline{a}))}{\Pr(y_x \text{ is an observation of } M(N, \underline{b}))}$$

will be maximised for x in the correct coset (i.e. that of x_0) for

$$N \approx (\log 2^s)/\overline{I}(\overline{g}_i(x); x) = s/\overline{I}(\overline{g}_i(x); x).$$

But

$$\begin{aligned}&\frac{\Pr(y_x \text{ is an observation of } M(N, \underline{a}))}{\Pr(y_x \text{ is an observation of } M(N, \underline{b}))}\\ &= \frac{\binom{N}{(N_{x;b,p})_{b,p}}\prod_p (p\Pr(p)))^{N_{x;0,p}}\prod_p((1-p)\Pr(p)))^{N_{x;1,p}}}{\binom{N}{(N_{x;b,p})_{b,p}}\prod_p(\frac{1}{2}\Pr(p)))^{N_{x;0,p}}\prod_p(\frac{1}{2}\Pr(p)))^{N_{x;1,p}}}\\ &= 2^N\prod_p p^{N_{x;0,p}}(1-p)^{N_{x;1,p}}\\ &= 2^N\Pr((b_i)_{i=1}^N \text{ comes from } (\overline{g}_i(x))_{i=1}^N),\end{aligned}$$

whose maximisation is equivalent to the maximum likelihood method described in section 2.2. □

***N* for reduced output function** In this case, we are considering the equations (2). We assume here, and subsequently when considering unicity distance for reduced output function, that the assumptions of the corollary apply for cosets of an m-dimensional subspace $W \leq V$ containing U.

$$I(\overline{g}_i(x); x) = I(\overline{f}_U(T^i x); x) = I(\overline{f}_U(x); x),$$

so the unicity distance is

$$N \approx m/I(\overline{f}_U(x); x). \tag{7}$$

***N* for reduced key space** Similarly with equations (3), if we make analogous assumptions, we can apply the corollary to learn that the unicity distance is

$$N \approx m'/\overline{I}(\overline{f \circ T^i}_U(x); x),$$

where m' denotes the dimension of $U^\perp$.

We shall develop this result further in section 5.

2.5 An observation concerning "correlation immunity"

Following Siegenthaler [5], we define a function $f(x_1, \ldots, x_n)$ to be *correlation-immune* to a set $\{i_1, \ldots, i_m\}$ of input positions if

$$I(f(x_1, \ldots, x_n); x_{i_1}, \ldots, x_{i_m}) = 0.$$

This mutual information is equal to

$$I(\overline{f}_U(x); x)$$

where U is the subspace of V generated by the i_1th, $\ldots$, i_mth standard basis vectors. Thus, by equation (7), if f is correlation-immune, the maximum likelihood method will fail no matter how much keystream is considered.

3 Correlation attacks on linearly clocking stream ciphers

Henceforth we focus our attention on the case where state transition is an invertible linear transformation on V; we will represent it as A, rather than T, and consider A to be an $n \times n$ matrix. (As a point of interest, we note that A is similar to a matrix in rational canonical form, so that, with no loss of generality, the generator can be taken to comprise a number of separate Galois-clocking LFSRs.)

This section will review some "traditional" theory of correlation attacks; in section 6 we will reformulate these ideas in the language of section 2, which will cast new light on existing attacks and produce some new ones.

3.1 Linear correlations to *f*

Linear correlation attacks exploit correlation between f and linear functions on V i.e. functionals $\in V^*$. For any function $g : V \to \{0, 1\}$ we define the *correlation* $c_{g,v}$ of g to the functional "$\cdot v$" by

$$c_{g,v} := \Pr(g(x) = x \cdot v) - \Pr(g(x) \neq x \cdot v).$$

For convenience, we will write c_v for a correlation $c_{f,v}$ of our keystream generator's output function f.

Observing that the set of all ± 1-valued functionals $\{x \mapsto (-1)^{x.v}\}_{v \in V}$ is a complete orthogonal subset of the real vector space of real-valued functions $V \to R$ with inner product

$$\langle g_1, g_2 \rangle = \sum_{v \in V} g_1(v) g_2(v),$$

we can apply Parseval to $(-1)^g$ to learn that

$$\| (-1)^g \|^2 = \sum_{v \in V} (\langle g, 2^{-n/2}(-1)^{\cdot v} \rangle)^2,$$

whence

$$\sum_{v \in V} c_{g,v}^2 = 1.$$

In particular, $c_{g,v}$ is non-zero for at least one $v \in V$.

3.2 Linear correlation attack

Fix a v for which $c_v \neq 0$, and let W be any A-invariant subspace of V containing v; as usual, put $m = \dim W$.

The sequences of bits $(f(A^i x))_i$ generated by the KG can be written $(((A^i x) \cdot v) \oplus e_{i,x})_i$, where the $e_{i,x}$ are modelled as independent Bernoulli random variables which take the value 0 with probability $(1 + c_v)/2 \neq \frac{1}{2}$. Thus the KG generates "noisy LFSR sequences". Considerable research effort has been directed towards the efficient implementation of minimum distance decoding of such sequences (when the error probability $< \frac{1}{2}$) to the underlying linear sequence. No published algorithm known to the author solves this problem in full generality with time complexity $< 2^m$ (but see Dodd [2]). Siegenthaler [6] presents the straightforward method : to maximise (or minimise) the number of agreements ("correlation") between each of the possible underlying linear sequences and the keystream. This method has time complexity $N2^m$. Later, in section 6, we shall characterise Siegenthaler's attack in terms of the language of section 2.

As previously observed by Mund *et al.* [4], a more efficient method, borrowed from the theory of decoding first-order Reed-Muller codes (see, *e.g.*, MacWilliams and Sloane [3]) makes use of the Walsh-Hadamard transform. Before describing this method, we introduce some notation and results concerning the Walsh-Hadamard transform.

3.3 The Walsh-Hadamard transform

For any subspace $S \leq V$ and real-valued function g on S, the Walsh-Hadamard transform $T_S(g)$ of g on S is also a real-valued function on S, defined by

$$T_S(g)(v) = \sum_{s \in S} (-1)^{s \cdot v} g(s) \qquad (v \in S),$$

where $\cdot$ denotes the usual inner product on V.

Denoting the dimension of S by d, we have the following results :

1. The array $T_S(g)(v)$ $(v \in S)$ for a function $g : S \to R$ can be computed in time $d2^d$ and space 2^d real storage locations.
2. $T_S(T_S(g)) = 2^d g$ $(g : S \to R)$.
3. If g is a $\{0,1\}$-valued function on S, $T_S((-1)^g)(v) = 2^d c_{g,v}$ (we defined $c_{g,x}$ in section 3.1).
4. $\sum_{s \in S} (T_S(g)(s))^2 = 2^d \sum_{s \in S} (g(s))^2$ $(g : S \to R)$.

3.4 Reed-Muller decoding algorithm

We now return to the situation introduced in section 3.2, but assume also that $W^\perp$ is A-invariant. (In section 5.3 we shall prove that this implies W is invariant under the transpose A^* of A.)

First of all, observe that for any $x, v \in W$

$$\begin{aligned}
(A^i x) \cdot v &= (A^i x)^T v \\
&= x^T (A^*)^i v, \qquad \text{where } A^* \text{ denotes the transpose of } A \\
&= x \cdot ((A^*)^i v) \\
&= x \cdot v_i, \qquad \text{where } v_i = (A^*)^i v \in W.
\end{aligned}$$

Now we can compute

$$\begin{aligned}
&\#\{\text{agreements between } x \cdot v_i \text{ and } b_i\} \text{ - } \#\{\text{disagreements between } x \cdot v_i \text{ and } b_i\} \\
&= \sum_{i=1}^{N} (-1)^{(x \cdot v_i) \oplus b_i}
\end{aligned}$$

$$= \sum_{w \in W} (-1)^{x.w} h(w), \qquad \text{putting } h(w) = \textstyle\sum_{i:v_i=w} (-1)^{b_i}$$
$$= T_W(h)(x),$$

the Walsh-Hadamard transform of h on W.

3.5 Complexity of the Reed-Muller attack

Using the results cited in section 3.3, we can summarise the complexity of the Reed-Muller attack (the corresponding parameters for Siegenthaler's attack are shown in brackets) :

- time complexity = $m2^m + N$ ($N2^m$);
- space complexity = 2^m (0).

3.6 Two significant observations

These two attacks, relying on correlation to the single functional "$\cdot v$", appear, heuristically, to waste information as compared to the maximum likelihood attack. Generally, many c_v will be non-zero, corresponding to *simultaneous correlation* of the keystream to many linear sequences. Moreover we can see from the results of section 3.3 that any function $f : V \mapsto \{0, 1\}$ is characterised by $T_V((-1)^f)$ and so also by its correlations $(c_v)_{v \in V}$ to linear functionals.

These observations will be explored subsequently in this paper.

4 Characterising a function by correlations to linear functionals

Given a real-valued function g on V, what function is characterised by the correlations of g to the functionals in a subspace S^* of V^* i.e. by the correlations $(c_{g,s})_{s \in S}$ for a subspace S of V? We now demonstrate that these correlations in fact characterise $\overline{g}_S$.

Lemma 1. *For any $S \leq V$,*

$$T_V(E((-1)^{\overline{f}_S))(v) = \begin{cases} T_V((-1)^f)(v) & \text{if } v \in S \\ 0 & \text{otherwise,} \end{cases}$$

where E denotes expected value.

Proof.

$$\begin{aligned} T_V(E((-1)^{\overline{f}_S))(v) &= \sum_{v' \in V} (-1)^{v' \cdot v} E((-1)^{\overline{f}_{S^{(v')}}}) \\ &= \sum_{v' \in V} (-1)^{v' \cdot v} \frac{1}{|S^\perp|} \sum_{s \in S^\perp} (-1)^{f(v'+s)} \\ &= \frac{1}{|S^\perp|} \sum_{v'' \in V, s \in S^\perp} (-1)^{(v''+s) \cdot v} (-1)^{f(v'')} \\ &= \frac{1}{|S^\perp|} \sum_{s \in S^\perp} (-1)^{s \cdot v} T_V((-1)^f)(v) \end{aligned}$$

Now $S = (S^\perp)^\perp$ ($S \subseteq (S^\perp)^\perp$ and they have the same dimension[1]), so that if $v \notin S$, $\exists s' \in S^\perp$ such that $v \cdot s' = 1$; then

$$\begin{aligned}\sum_{s\in S^\perp} (-1)^{s\cdot v} &= \sum_{s\in S^\perp} (-1)^{(s+s')\cdot v} \\ &= -\sum_{s\in S^\perp} (-1)^{s\cdot v},\end{aligned}$$

so this sum is 0. If $v \in S$, $\sum_{s\in S^\perp}(-1)^{s\cdot v} = |S^\perp|$. Hence the result. □

Corollary 2. *$\overline{g}_S$ is characterised by the $(c_{g,s})_{s\in S}$.*

5 The unicity distance N in terms of correlations

In this section we demonstrate that the expressions for the unicity distances obtained in section 2.4 can be couched in terms of the correlations c_v of the output function f.

5.1 Information in terms of correlations

Proposition 1. *If $S \leq V$ is any subspace and $g : V \to R$ any balanced function for which all $\Pr(\overline{g}_S(x) = 0) \approx \frac{1}{2}$, then we can approximate*

$$I(x; \overline{g}_S(x)) \approx \frac{1}{2\ln 2} \sum_{s\in S} c_s^2.$$

Proof. As a preliminary, we use the Taylor expansion

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 \cdots$$

to establish that :-

$$\begin{aligned} i(p) &:= p\ln(p/\tfrac{1}{2}) + q\ln(q/\tfrac{1}{2}) \qquad (\text{where } q = 1-p) \\ &= p[(2p-1) - \frac{1}{2}(2p-1)^2 + \frac{1}{3}(2p-1)^3 \cdots] + \\ &\quad q[(2q-1) - \frac{1}{2}(2q-1)^2 + \frac{1}{3}(2q-1)^3 \cdots] \\ &= p[(p-q) - \frac{1}{2}(p-q)^2 + \frac{1}{3}(p-q)^3 \cdots] + \\ &\quad q[-(p-q) - \frac{1}{2}(p-q)^2 - \frac{1}{3}(2q-1)^3 \cdots] \\ &= \frac{1}{2}(p-q)^2 + \frac{1}{12}(p-q)^4 + \cdots \\ &\approx \frac{1}{2}(p-q)^2 \qquad \text{for } p \approx \tfrac{1}{2}. \end{aligned}$$

[1] Let X be any subspace of V. If M is a matrix whose columns are a basis of X, im $M \simeq V/\ker M$, so dim im M = dim V - dim ker M; but im $M = X$ and ker $M = X^\perp$, so dim X = dim V- dim $X^\perp$. Applying this twice, dim $(S^\perp)^\perp$ = dim V - dim $S^\perp$= dim V - (dim V - dim S) = dim S.

Now we compute

$$\begin{aligned}
I(x;\overline{g}_S(x)) &= H(\overline{g}_S(x)) - H(\overline{g}_S(x)|x) \\
&= 1 - \sum_{x\in V} \Pr(x)\left(1 - \frac{1}{\ln 2} i(\Pr(\overline{g}_S(x)=0))\right) \\
&= \frac{1}{\ln 2} 2^{-n} \sum_{x\in V} i(\Pr(\overline{g}_S(x)=0)) \\
&\approx \frac{1}{2\ln 2} 2^{-n} \sum_{x\in V} (\Pr(\overline{g}_S(x)=0) - \Pr(\overline{g}_S(x)=1))^2 \\
&= \frac{1}{2\ln 2} 2^{-n} \sum_{x\in V} E((-1)^{\overline{g}_S(x)})^2 \\
&= \frac{1}{2\ln 2} 2^{-n} 2^{-n} \sum_{v\in V} T_V(E((-1)^{\overline{g}_S}))(v)^2 \quad \text{(by 3.3, point 4)} \\
&= \frac{1}{2\ln 2} 2^{-2n} \sum_{s\in S} T_V((-1)^g)(s)^2 \qquad \text{(by section 4)} \\
&= \frac{1}{2\ln 2} \sum_{s\in S} c_{g,s}^2 \qquad \text{(by 3.3, point 3).}
\end{aligned}$$

□

(Notice that, with our assumption that g is balanced, $c_{g,0} = 0$.)

5.2 Unicity distance when reducing the output function

Now that state transition is linear, A is well-defined on the cosets of a subspace of V if and only if that subspace is A-invariant; and since the intersection of any two A-invariant subspaces containing $U \leq V$ is also an A-invariant subspace containing U, there there is a smallest such subspace, W (which will also be the smallest subspace of V containing U for which A is well-defined on cosets of $W^\perp$). As before, let $m = \dim W$.

As we stated in equation (7), the unicity distance when reducing the output function f to $\overline{f}_U$ is

$$N \approx m / I(x;\overline{f}_U(x)).$$

Applying the result of proposition 1, we see that

$$N \approx \frac{2m\ln 2}{\sum_{u\in U} c_u^2},$$

or

$$N \approx \frac{m}{\sum_{u\in U} c_u^2}.$$

5.3 Unicity distance when reducing the key space

Similarly in this case, section 2.4 gives the unicity distance in terms of $\overline{I}(x;\overline{(f\circ A^i)}_U)$. To compute this information using proposition 1, we first compute

$$\begin{aligned}
T_V((-1)^{f\circ A^i})(v) &= \sum_{v'\in V} (-1)^{v\cdot v'} (-1)^{f(A^i v')} \\
&= \sum_{v'\in V} (-1)^{v^T v'} (-1)^{f(A^i v')}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{v'' \in V} (-1)^{v^T A^{-i} v''} (-1)^{f(v'')} \\
&= \sum_{v'' \in V} (-1)^{((A^*)^{-i} v)^T v''} (-1)^{f(v'')} \\
&= T_V((-1)^f)((A^*)^{-i} v) \\
&= 2^n c_{((A^*)^{-i} v)} \qquad \text{by section 3.3, point 3.}
\end{aligned} \tag{8}
$$

Therefore, by proposition 1,

$$
I(x; \overline{f \circ A^i}_U) \approx \frac{1}{2 \ln 2} \sum_{v \in U} c^2_{((A^*)^{-i} v)}. \tag{9}
$$

Now we wish to average this over values i, but in order to do this, we need some preliminary results.

Lemma 2. *For any subspace $S \leq V$, $A^{-1} S^\perp = (A^* S)^\perp$.*

Proof. For any $v, v' \in V$,

$$
(Av) \cdot v' = (Av)^T v' = v^T (A^* v') = v \cdot (A^* v');
$$

Therefore

$$
\begin{aligned}
v \in (A^* S)^\perp &\Leftrightarrow v \cdot (A^* s) = 0 \qquad \text{for all } s \in S \\
&\Leftrightarrow Av \cdot s = 0 \qquad \text{for all } s \in S \\
&\Leftrightarrow Av \in S^\perp \\
&\Leftrightarrow v \in A^{-1} S^\perp
\end{aligned}
$$

□

Corollary 3. *A subspace $S \leq V$ is A^*-invariant $\Leftrightarrow$ $S^\perp$ is A-invariant.*

Proof.

$$
\begin{aligned}
S = A^* S &\Leftrightarrow S^\perp = (A^* S)^\perp \\
&\Leftrightarrow S^\perp = (A^{-1}) S^\perp \\
&\Leftrightarrow A S^\perp = S^\perp
\end{aligned}
$$

□

Now we see that W is also the smallest A^*-invariant subspace of V containing U, and that the subspaces $(A^*)^{-1} U$ $(i = 1, \ldots, N)$ will include the non-zero elements of W with approximately equal probabilities $2^{m'-m}$ (recall that $m' = \dim U$). And since c_0=0, the average value of the right hand side of equation (9) is

$$
\frac{1}{2 \ln 2} 2^{m'-m} \sum_{w \in W} c^2_w.
$$

Hence

$$
N \approx 2^{m-m'} \frac{m'}{\sum_{w \in W} c^2_w}, \tag{10}
$$

In particular, if $W = V$, or if each $c^2_v \approx 2^{-n}$,

$$
N \approx m' 2^{n-m'}. \tag{11}
$$

6 Maximum likelihood attacks in terms of correlations

6.1 Reformulation of a maximum likelihood condition

In section 2, we saw that elements x of the most likely coset of initial keystream generator states given observed keystream $(b_i)_{i=1}^N$ maximise

$$\sum_{i=1}^{N} E((-1)^{\overline{g}_i(x)\oplus b_i}) \tag{12}$$

in the case where all $\Pr(\overline{g}_i(x) = 0) \approx \frac{1}{2}$.

In the following sections, we shall reformulate this condition — in the case of reduced output function or reduced key space — in terms of the correlations c_v of f, and explore cryptanalytic methods they suggest.

6.2 Reformulation for reduced output function

Considering (12) in the case where $\overline{g}_i(x) = \overline{f}_U(A^i x)$, we first note that

$$\begin{aligned} T_V(E((-1)^{\overline{f}_U}))(v) &= \begin{cases} 2^n c_v \text{ if } v \in U \\ 0 \quad \text{otherwise} \end{cases} \\ &= T_V(\sum_{u\in U} c_u(-1)^{\cdot u})(v). \end{aligned}$$

Therefore

$$E((-1)^{\overline{f}_U(x)})) = \sum_{u\in U} c_u(-1)^{x\cdot u},$$

and we can rewrite (12) as follows :

$$\begin{aligned} \sum_{i=1}^{N} E((-1)^{\overline{g}_i(x)\oplus b_i}) &= \sum_{i=1}^{N} (-1)^{b_i} E((-1)^{\overline{f}_U(A^i x)}) \\ &= \sum_{i=1}^{N} \sum_{u\in U} c_u(-1)^{(A^i x)\cdot u \oplus b_i}. \end{aligned} \tag{13}$$

6.3 Cryptanalytic applications

In this section we present two observations concerning the use of expression (13) in maximum likelihood attacks. Throughout the section we suppose $V = W \oplus W^\perp$, so that each coset of $W^\perp$ has a (unique) representative in W, and consequently we need only evaluate (13) for $x \in W$.

Siegenthaler's method If we put $U = \langle v \rangle$, maximising (13) over x amounts to maximising

$$c_v \sum_{i=1}^{N} (-1)^{(A^i x)\cdot v \oplus b_i},$$

which is just Siegenthaler's "closest fit" method. A corollary of section 5 is that this method succeeds with about m/c_v^2 bits of keystream.

Generalised "Reed-Muller method" A speedup akin to that of section 3.4 can be obtained for the task of finding maximum likelihood solutions for any output function reduction, by writing (13) as

$$\begin{aligned}\sum_{i=1}^{N}\sum_{u\in U} c_u(-1)^{(A^i x)\cdot u\oplus b_i} &= \sum_{i=1}^{N}\sum_{v\in U} c_v(-1)^{x\cdot(A^{*i}v)\oplus b_i}\\ &= T_W(h)(x),\end{aligned}$$

where $h(x) = \sum_{v,i:A^{*i}v=x} c_v(-1)^{b_i}$ $(x \in W)$.

The vector of all values of h can be computed in time $2^{m'}N$, so (13) can be computed for all cosets in time $m2^m + N2^{m'}$ and space 2^m, rather than time $N2^m$ and negligible space : quite a remarkable result!

6.4 Reformulation for reduced key space

In this case, $\overline{g}_i(x) = \overline{f \circ A^i}_U(x)$, and expression (12) is

$$\begin{aligned}&\sum_{i=1}^{N} E((-1)^{\overline{f\circ A^i}_U(x)\oplus b_i})\\ &= 2^{-n}T_V(T_V(\sum_{i=1}^{N} E((-1)^{\overline{f\circ A^i}_U\oplus b_i})))(x) \qquad \text{by section 3.3, point 2}\\ &= 2^{-n}T_V(\sum_{i=1}^{N} T_V(E((-1)^{\overline{f\circ A^i}_U\oplus b_i})))(x) \qquad \text{by the linearity of } T_V\\ &= \sum_{u\in U}(-1)^{x\cdot u}(\sum_{i=1}^{N}(-1)^{b_i}c_{(A^*)^{-i}u}) \qquad \text{using lemma 1 and (8).} \qquad (14)\end{aligned}$$

6.5 More cryptanalytic applications

In this section, we show that (14) can provide a practical vehicle for cryptanalytic attack if the correlations c_v of f vanish outside some subspace X of dimension r which is not too large. (This will be the case if f depends only on a small number r of state bits.)

We assume $V = U \oplus U^{\perp}$, so that each coset of $U^{\perp}$ has a unique representative in U, and we can perform a maximum likelihood attack by maximising (14) over $x \in U$. Thus the outer sum in (14) is a Walsh-Hadamard transform on U, which we can compute in time $m'2^{m'}$ and space $2^{m'}$.

Terms of the inner sum in (14) contribute only when $c_{(A^*)^{-i}u} \neq 0$, so we need only compute $c_{(A^*)^{-i}u}$ for those $u \in U$ for which $(A^*)^{-i}u \in X$ i.e. for $u \in U \cap A^{*i}X$. To see how many such u there are for each i, we apply the following lemma.

Lemma 3. *For a random r-dimensional subspace S of V,* $\dim(S \cap U) \approx r + m' - n$.

Proof. The expected size of $S \cap U\backslash\{0\}$ is $(2^n - 1)\times$ the probability that a random element of $V\backslash\{0\}$ is in both $S\backslash\{0\}$ and $U\backslash\{0\}$, i.e.

$$(2^n - 1) \times (2^r - 1)/(2^n - 1) \times (2^{m'} - 1)/(2^n - 1) \approx 2^{r+m'-n}.$$

□

Thus if we model each $A^{*i}X$ as a random r-dimensional subspace of V, dim $(U \cap A^{*i}X) \approx r + m' - n$, and its elements can be efficiently computed[2] in time $\max\{1, 2^{r+m'-n}\}$.

Combining all this, we see that we can compute (14) for all cosets of x with

- time complexity $= m'2^{m'} + N \max\{1, 2^{r+m'-n}\}$;
- space complexity $= 2^{m'}$;
- number N of required keystream bits given by (10).

For $r \leq n/2$, and $W = V$, the value for dim U which minimises the time complexity is $m' = n/2$, when

- time complexity $= n2^{n/2}$;
- space complexity $= 2^{n/2}$;
- number N of required keystream bits $= (n/2)2^{n/2}$ (by equation (11)).

6.6 Why less keystream may be required

Suppose $r < n/2$ and let s be maximal such that $rs \leq n/2$. Given N' bits of keystream, we can construct $\binom{N'}{s}$ "reduced" equations for the initial key :

$$\overline{g_{(i_1,\ldots,i_s)}}_U(x) = b_{i_1} \oplus \cdots \oplus b_{i_s}, \qquad (1 \leq i_j \leq N', j = 1, \ldots, s)$$

where we have defined

$$g_{(i_1,\ldots,i_s)}(x) := \bigoplus_{j=1}^{s} f(A^{i_j}x).$$

The transform of $(-1)^{g_{(i_1,\ldots,i_s)}}$ is the convolution of the transforms of the $(-1)^{f \circ A^{i_j}}$; consequently it vanishes outside a subspace of dimension rs. Now essentially the method of the previous section applies, with required number N' of required known keystream bits satisfying

$$\binom{N'}{s} \approx N,$$

i.e., for $s \ll N$,

$$N' \approx (N.s!)^{1/s}.$$

6.7 Example

To perform a known keystream attack on the sequence generator illustrated in figure 1, we can choose U to be the set of states whose first 32 bits are 0. Then $U^{\perp}$ is the set of states whose last 32 bits are 0, $W = V$, $n = m = 64$, $m' = 32$, $N = 32.2^{64-32}$, $r = 8$, $s = 4$, and the attack will determine the last 32 bits of the initial state with

- time complexity $\approx 64.2^{32} = 2^{38}$;
- space complexity $= 2^{32}$;
- number N' of required keystream bits $\approx (32.2^{32}.4!)^{1/4} \approx 2^{10 \cdot 4}$.

6.8 A brief observation concerning probabilistic f

The techniques of this section may be applicable even when some inputs to f are not known linear functions of the initial state, but instead can be modelled as independent random bits : their effect may then be absorbed by a suitable choice of U.

[2] In the case when U is a subspace whose elements are precisely those with 0 entries in certain coordinate positions, straightforward Gaussian elimination can be used.

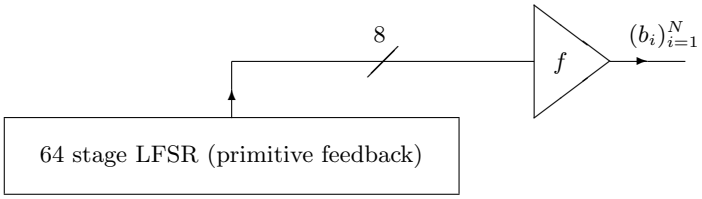

Fig. 1. Example keystream generator

References

1. Brynielsson, L. (1989). *Below the Unicity Distance*. Proceedings of the E.I.S.S. (European Institute for System Security) Workshop on Stream Ciphers held at Karlsruhe, Germany.
2. Dodd, M.W. (to appear). Ph.D. Thesis, University of London.
3. MacWilliams, F.J. and Sloane, N.J.A. (1977). *The Theory of Error-Correcting Codes* (1st edition), North Holland. 406–32.
4. Mund, S., Gollman, D., and Beth, T. (1987). Some remarks on the cross correlation analysis of pseudo-random generators. *Advances in Cryptology — Eurocrypt '87*, 25–35.
5. Siegenthaler, T. (1984). Correlation-immunity of nonlinear combining functions for cryptographic applications. *IEEE Transaction on Information Theory, IT-30*, no. 5, 776–80.
6. Siegenthaler, T. (1985). Decrypting a class of stream ciphers using ciphertext only. *IEEE Transaction on Computers, C-34*, no. 1, 81–5.