Generalized Correlation Analysis of Vectorial Boolean Functions

Claude Carlet, Khoongming Khoo,
Chu-Wee Lim and Chuan-Wen Loe

Introduction

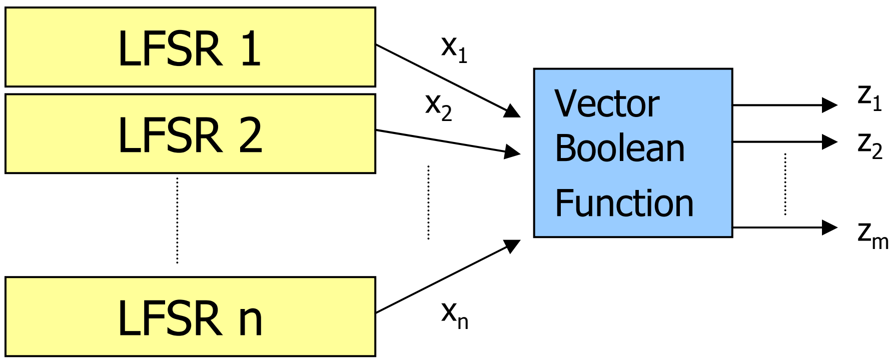
Correlation Attack of Vectorial Stream Ciphers

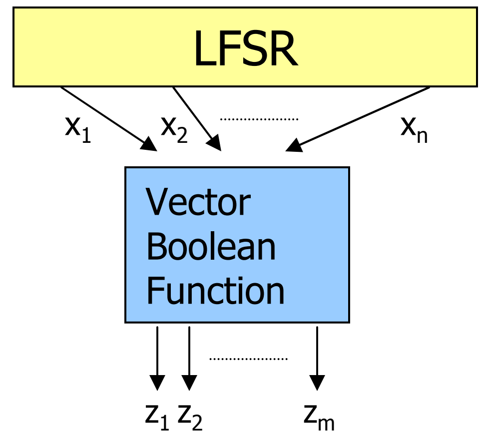
- In this talk, we shall improve correlation attacks on vectorial stream ciphers.

- Will consider vectorial Boolean functions in combinatorial and filtering generators.
  - Will not go into the details of the correlation attack.

- Focus on how to obtain good linear approximation.

Correlation Attack of Vectorial Stream Ciphers





- In standard correlation attack of vectorial Boolean functions, we form linear approximation of the form:

$$\Pr(b_1 z_1 \oplus \cdots \oplus b_m z_m = w_1 x_1 \oplus \cdots \oplus w_n x_n) = \Pr(b \cdot z = w \cdot x).$$

Linear Bias and Nonlinearity

- For correlation attack to succeed, we require

$$\text{Bias} = \left|\Pr(b \cdot z = w \cdot x) - 1/2\right| \text{ to be high.}$$

where $z = F(x)$ is the output. I.e. probability far away from ½.

- This is equivalent to the condition that nonlinearity

$$N_F = 2^{n-1} - \frac{1}{2} \max_{w \neq 0, b} \left| \sum_{x \in GF(2)^n} (-1)^{b \cdot F(x) + w \cdot x} \right| \text{ is low,}$$

Zhang-Chan Attack

- At Crypto 2000, Zhang and Chan noticed that $z=F(x)$ is known, therefore we can consider

$$\Pr(g(z) = w_1 x_1 \oplus \cdots \oplus w_n x_n) = \Pr(g(z) = w \cdot x)$$

which is linear in x for any Boolean function $g(\cdot)$.

- Because approximation of $b{\cdot}z$ is a particular case of approximation of $g(z)$. It is easier to get a better linear approximation, i.e. get $\Pr(g(z)= w{\cdot}x)$ further away from ½ than $\Pr(b{\cdot}z = w{\cdot}x)$.

Zhang-Chan Attack

- For Zhang-Chan attack to succeed, we require

$$\text{Bias} = \left|\Pr(g(z) = w \cdot x) - 1/2\right| \text{ to be high.}$$

  where $z = F(x)$ is known.

- This is equivalent to the condition that unrestricted nonlinearity

$$UN_F = 2^{n-1} - \frac{1}{2} \max_{w \neq 0, g(\cdot)} \sum_{\mathrm{x} \in \mathrm{GF}(2)^{\mathrm{n}}} (-1)^{g(F(x)) + w \cdot x} \text{ is low,}$$

Generalized Correlation

Generalized Correlation Attack

- We still want to get approximations which are linear in x.

- The most general approximation which is linear in x:

$$\Pr(g(z) = w_1(z)x_1 \oplus \cdots \oplus w_n(z)x_n) = \Pr(g(z) = w(z) \cdot x)$$

where $w_i(z)$ are Boolean functions of the known output z and $w(z)=(w_1(z),\ldots, w_n(z))$

Generalized Correlation Attack

- For generalized correlation attack to succeed, we require

$$\text{Bias} = \left|\Pr(g(z) = w(z)\cdot x) - 1/2\right| \text{ to be high.}$$

where $z = F(x)$ is known.

- This is equivalent to the condition that generalized nonlinearity

$$GN_F = 2^{n-1} - \frac{1}{2}\max_{w(\cdot)\neq 0,\, g(\cdot)} \sum_{x\in GF(2)^n} (-1)^{g(F(x)) + w(F(x))\cdot x} \text{ is low,}$$

Generalized Correlation Attack

- $g(z) = w(z) \cdot x$ is a more general approximation than $g(z) = w \cdot x$, which in turn is a more general approximation than $b \cdot z = w \cdot x$.

- Therefore $\Pr(g(z) = w(z) \cdot x)$ can be chosen to be further away from ½ than the other two approximations.

- In terms of nonlinearities,

$$GN_F \leq UN_F \leq N_F$$

From a Cipher Designer's Viewpoint

- From the viewpoint of a stream cipher designer, he needs to ensure generalized nonlinearity GN_F is high for protection against correlation attack. Then automatically, UN_F and N_F will be high.

Comparison of Generalized Correlation Attack with Known Methods

An Example on Bent Functions

$x=x_1x_2x_3x_4$	0000	0001	0010	0011	0100	0101	0110	0111
$F(x)=(z_1z_2)$	00	00	00	00	00	01	10	11
$x=x_1x_2x_3x_4$	1000	1001	1010	1011	1100	1101	1110	1111
$F(x)=(z_1z_2)$	11	00	10	01	11	01	00	10

- $F(x)$ is a bent function from $\mathrm{GF}(2)^4$ to $\mathrm{GF}(2)^2$. We have N_F=6 and UN_F=5. This means the best affine approximation has probability 0.63 and 0.69 for usual and Zhang-Chan.
- For generalized correlation attack, we have GN_F=2. The best generalized approximation has probability:

$$\Pr(z_1 + z_2 = (z_1 + 1)(z_2 + 1)x_2 + z_1 x_3 + z_2 x_4) = 0.88$$

How much better is Generalized Correlation Attack?

- Below is a table comparing average nonlinearities of 10000 randomly generated balanced functions from n-bits to $n/2$-bits:

n	6	8	10	12	14
N_F	18	100	443	1897	7856
UN_F	16	88	407	1768	7454
GN_F	6	36	213	1101	5224

GN_F is much lower than N_F and UN_F

How much better is Generalized Correlation Attack?

- Here's the table for average best approximation probability of the previous functions from n-bits to $n/2$-bits:

n	6	8	10	12	14
Probability (usual)	0.72	0.61	0.57	0.54	0.52
Probability (Zhang-Chan)	0.75	0.66	0.60	0.57	0.55
Probability (generalized)	0.91	0.86	0.79	0.73	0.68

Probability of generalized attack much further away from 0.5 than the other attacks

Another Example on Inverse Function

- Let us compare the various approximation probability for x^{-1} on GF(2^8) restricted to m output bits.

m	1	2	3	4	5	6	7
Probability (usual)	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Probability (Zhang-Chan)	0.56	0.58	0.61	0.63	0.67	0.73	0.78
Probability (generalized)	0.56	0.69	0.74	0.84	**1.00**	**1.00**	**1.00**

Computation of Generalized Nonlinearity

Computation of Generalized Nonlinearity

- Since we saw that generalized correlation attack is more powerful than known attacks, it is useful to compute the generalized nonlinearity.

$$GN_F = 2^{n-1} - \frac{1}{2} \max_{w(\cdot)\neq 0, g(\cdot)} \sum_{x \in GF(2)^n} (-1)^{g(F(x)) + w(F(x)) \cdot x}$$

- We need to compute

$$\sum_{x \in GF(2)^n} (-1)^{g(F(x)) + w_1(F(x)) x_1 + \cdots + w_n(F(x)) x_n}$$

over all choices of $g, w_1, \ldots, w_n : \mathrm{GF}(2)^m \rightarrow \mathrm{GF}(2)$.

Computation of Generalized Nonlinearity

- We need to compute

Each of these n+1 functions have 2^{2^m} choices

Each sum has complexity 2^n

$$\sum_{x \in GF(2)^n} (-1)^{g(F(x)) + w_1(F(x))x_1 + \cdots + w_n(F(x))x_n}$$

over all choices of $g, w_1, \ldots, w_n : \mathrm{GF}(2)^m \to \mathrm{GF}(2)$.

- Therefore complexity is approximately

$$\left(2^{2^m}\right)^{n+1} \times 2^n = 2^{2^m(n+1)+n}$$

More Efficient Computation of Generalized Nonlinearity

- **Theorem:** The generalized nonlinearity

$$GN_F = 2^{n-1} - \frac{1}{2} \max_{w(\cdot)\neq 0, g(\cdot)} \sum_{x \in GF(2)^n} (-1)^{g(F(x)) + w(F(x))\cdot x}$$

can be computed as

$$GN_F = 2^{n-1} - \frac{1}{2} \sum_{z \in GF(2)^m} \max_{w \in GF(2)^n \setminus \{0\}} \left| \sum_{x \in F^{-1}(z)} (-1)^{w\cdot x} \right|$$

Here we do not find the optimal functions $w_1(),\ldots,w_n()$ and $g()$, instead we just find an optimal vector $w \in \mathrm{GF}(2)^n \setminus \{0\}$ at each z.

Complexity

2^n-1 choices for w

$$GN_F = 2^{n-1} - \frac{1}{2} \sum_{z \in GF(2)^m} \max_{w \in GF(2)^n \setminus \{0\}} \left| \sum_{x \in F^{-1}(z)} (-1)^{w \cdot x} \right|$$

Complexity for this sum is $|F^{-1}(z)|$

- The new complexity for computing generalized nonlinearity is $\sum_{z \in GF(2)^m} (2^n - 1) \times \left|F^{-1}(z)\right| = (2^n - 1)2^n \approx 2^{2n}$
- This is much faster compared to original complexity of $2^{2^m(n+1)+n}$

Upper Bound on Generalized Nonlinearity

Upper Bound

- **Theorem:** If $F(x)$ is balanced, then an upper bound for GN_F:

$$GN_F \le 2^{n-1} - 2^{n-1}\sqrt{\frac{2^m - 1}{2^n - 1}}$$

- This is much lower than the known upper bounds for unrestricted nonlinearity UN_F and nonlinearity N_F:

$$UN_F \le 2^{n-1} - \frac{1}{2}\left(\frac{2^{2m} - 2^m}{2^n - 1} + \sqrt{\frac{2^{2n} - 2^{2n-m}}{2^n - 1} + \left(\frac{2^{2m} - 2^m}{2^n - 1} - 1\right)^2} - 1\right)$$

$$N_F \le 2^{n-1} - 2^{n/2-1}$$

For $m \leq n/2$, the upper bound for unrestricted nonlinearity UN_F does not improve on the Covering Radius Bound $2^{n-1}-2^{n/2-1}$.

The upper bound for generalized nonlinearity GN_F does.

Comparison of Upper Bound for N_F, UN_F and GN_F

n	6	8	10	12	14	16
$m=n/2$	3	4	5	6	7	8
Upp Bd N_F	28	120	496	2016	8128	32640
Upp Bd UN_F	29	121	497	2017	8129	32641
Upp Bd GN_F	22	97	423	1794	7471	30724

Corresponding Bound for Probability of Best Approximation

n	6	8	10	12	14	16
$m=n/2$	3	4	5	6	7	8
Probability (usual)	≥ 0.563	≥ 0.531	≥ 0.516	≥ 0.508	≥ 0.504	≥ 0.502
Probability (Zhang-Chan)	≥ 0.558	≥ 0.530	≥ 0.515	≥ 0.508	≥ 0.504	≥ 0.502
Probability (generalized)	≥ 0.667	≥ 0.621	≥ 0.587	≥ 0.562	≥ 0.544	≥ 0.531

For $m>n/2$, the upper bound for unrestricted nonlinearity UN_F does improve on the Covering Radius Bound but not by much.

The upper bound for generalized nonlinearity GN_F improves on the Covering Radius bound $2^{n-1}-2^{n/2-1}$ by much more.

Comparison of Upper Bound for N_F, UN_F and GN_F

n	6	8	10	12	14	16
$m=3n/4$	4	6	7	9	10	12
Upp Bd N_F	28	120	496	2016	8128	32640
Upp Bd UN_F	27	110	487	1972	8090	32460
Upp Bd GN_F	17	65	332	1325	6145	24577

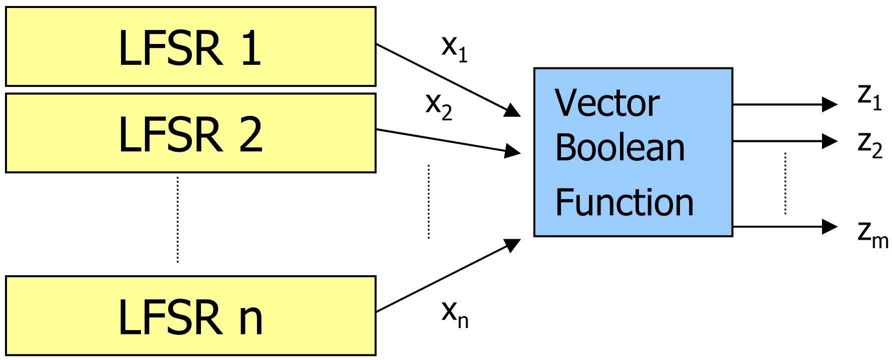
Corresponding Bound for Probability of Best Approximation

n	6	8	10	12	14	16
$m=3n/4$	4	6	7	9	10	12
Probability (usual)	≥0.563	≥0.531	≥0.516	≥0.508	≥0.504	≥0.502
Probability (Zhang-Chan)	≥0.587	≥0.571	≥0.524	≥0.519	≥0.506	≥0.505
Probability (generalized)	≥0.744	≥0.749	≥0.676	≥0.677	≥0.625	≥0.625

Thus we have further evidence that generalized correlation attack is more effective than Zhang-Chan and usual correlation attack on vector Boolean functions.

Generalized Resiliency

Siegenthaler’s Attack



- Suppose there exists a correlation $\Pr(x_1 = z_1 \oplus z_2) = 3/4$.
- Then we guess the content of LFSR1
- If our guess is correct, LFSR1 sequence matches the known keystream $z_1 \oplus z_2$ with probability ¾.
- If not, LFSR1 sequence matches the keystream with probability ½.
- Reduction in attack complexity: Instead of attacking all LFSR’s simultaneously, we attack one LFSR separately and then the others.

Resiliency

- To prevent against the previous attack, we want to avoid linear approximations which involve too few input variables.

- A function $F\text{:GF}(2)^n \rightarrow \text{GF}(2)^m$ is called correlation immune of order k if

$$\Pr(b \cdot z = w \cdot x) = \tfrac{1}{2}$$

for all $b \in \text{GF}(2)^m \backslash \{0\}$ whenever $1 \leq \text{wt}(w) \leq k$. If furthermore, $F(x)$ is balanced, then we say $F(x)$ is k-resilient.

Generalized Siegenthaler's Attack

- Suppose for a set of output vectors, e.g. z = 0000, 0001, 0010, 0111,… there exists good approximations

$$\Pr(L_1(x,z)=0) = p_1 \neq \tfrac{1}{2},\ \Pr(L_2(x,z)=0) = p_2 \neq \tfrac{1}{2}, \ldots$$

  which are linear in x and involve only k variables $x_1,\ldots,x_k$ (where k is small) out of n variables $x_1,\ldots,x_n$.

- We can attack k LFSR's instead of all n LFSR's. E.g. guess the contents of the k LFSR's and see if they satisfy the approximations

$$\Pr(L_1(x,z)=0) = p_1 \neq \tfrac{1}{2},\ \Pr(L_2(x,z)=0) = p_2 \neq \tfrac{1}{2}, \ldots$$

Generalized Resiliency

- To prevent against the previous attack, we want to avoid linear approximations $\Pr(L(x,z)=0)=p\neq½$ which involve too few input variables $x_1,\ldots,x_n$ for any subset of output z.

- A function $F:\mathrm{GF}(2)^n\rightarrow\mathrm{GF}(2)^m$ is called generalized correlation immune of order k if for all $z\in\mathrm{GF}(2)^m$

$$\Pr(\mathrm{g(z)}\oplus w_1(\mathrm{z})x_1 \oplus\ldots \oplus w_n(\mathrm{z})x_n) = ½$$

whenever $\mathrm{wt}(w_1(z),\ldots, w_n(\mathrm{z}))\leq k$. If furthermore, $F(x)$ is balanced, then we say $F(x)$ is generalized k-resilient.

Equivalence between Resiliency and Generalized Resiliency

- **Theorem:** A function $F\text{:GF}(2)^n \rightarrow \text{GF}(2)^m$ is correlation immune of order k if and only if it is generalized correlation immune of order k.

- The above statement is true if we replace correlation immune with resilient.

Generalized Nonlinearity of Secondary Constructions

Output Composition

- It is common to form balanced highly nonlinear vectorial functions by dropping output bits of a highly nonlinear permutation, e.g. x^{-1}, $x^{2^\wedge k+1}$. The nonlinearity N_F is preserved in this case.

  We prove the following generalization.

- **Proposition:** Let F:$\mathrm{GF}(2)^n \rightarrow \mathrm{GF}(2)^m$ and G:$\mathrm{GF}(2)^m \rightarrow \mathrm{GF}(2)^k$ be balanced vector functions. Then $GN_{G \circ F} \geq GN_F$.

- If $G(x)$ is a permutation, then $GN_{G \circ F} = GN_F$.

Concatenation

- By our previous result, a resilient function is also generalized resilient.
- Therefore we would like to check that secondary constructions for resilient functions yield high generalized nonlinearity.
- A secondary construction for resilient function we will look at is concatenation.

Concatenation

- **Proposition (Zhang-Zheng):** Let $F:\mathrm{GF}(2)^n \to \mathrm{GF}(2)^m$ be t_1-resilient and $\mathrm{G}:\mathrm{GF}(2)^p \to \mathrm{GF}(2)^q$ be t_2-resilient. Then $H:\mathrm{GF}(2)^{\mathrm{n+p}} \to \mathrm{GF}(2)^{\mathrm{m+q}}$ defined by $H(x,y)=(F(x),G(y))$ is a t-resilient function where $t=\min(t_1,t_2)$.

- **Proposition:** For $H(x,y)$ as defined above:

$$GN_H \leq 2^{n+p\text{-}1} - ½(2^n\text{-}2\mathrm{G}N_F)(2^p\text{-}2\mathrm{G}N_G)$$

- Thus for $H(x,y)$ to have high generalized nonlinearity, both component functions $F(x)$ and $G(y)$ must have high generalized nonlinearity.