On the Linear Syndrome Method in Cryptanalysis

Kencheng Zeng Minqiang Huang

Data and Communications Security Research Center

Graduate School of USTC, BX 100039-08, Beijing, China

The linear syndrome (LS) method is elaborated for the purpose of solving problems encountered in cryptanalysis, which can be reduced to the following mathematical setting. Suppose the cryptanalyst has at his hand a sufficiently long segment of the binary sequence

$$B = A + X ,$$

where A is a linear sequence with known feedback polynomial f(x) and X is a sequence with unknown or very complicated algebraic structure, but is sparse in the sense that, if we denote its signals by x(i), $i > 0$, then we shall have

$$s = \text{prob}(x(i) = 1) = 1/2 - \epsilon , \quad 0 < \epsilon < 1/2 .$$

We call s the error rate of the sequence A in the sequence B, and the job of the cryptanalyst is to recover the former from the captured segment of the latter.

One way for tackling this problem is to make use of the ideas of error correction, especially when s is comparatively small. In doing this we consider, for some fixed integer $r \geq 3$, a finite collection of r-nomials of the form

$$g(x) = 1 + x^{i_1} + x^{i_2} + \ldots + x^{i_{r-1}},$$

and compute, for every $i \geq \max\{ \deg g(x) \}$ and all g(x), the syndromes

$$\sigma_{i,k}(g) = \sum_{j=0}^{r-1} b(i - i_k + i_j) , \quad 0 \leq k \leq r-1 ,$$

S. Goldwasser (Ed.): Advances in Cryptology - CRYPTO '88, LNCS 403, pp. 469-478, 1990.

b(i) , $i \geq 0$, being the signals of the sequence B. The LS method is based on the following

Lemma 1. If f(x) divides g(x), then

$$\text{prob}(\sigma_{i,k}(g) = x(i)) = 1/2 + (1 - 2s)^{r-1}/2. \qquad (1)$$

Proof. Denote the signals of A by a(i), $i \geq 0$. Since f(x) | g(x), we have

$$\begin{aligned}\sigma_{i,k}(g) &= \sum_{p=0}^{r-1} b(i - i_k + i_p) \\ &= \sum_{p=0}^{r-1} a(i - i_k + i_p) + \sum_{p=0}^{r-1} x(i - i_k + i_p) \\ &= \sum_{p=0}^{r-1} x(i - i_k + i_p) .\end{aligned}$$

Thus we see $\sigma_{i,k}(g) = x(i)$ if and only if an even number of the signals

$$x(i - i_k), \ldots, x(i - i_k + i_{k+1}), x(i - i_k + i_{k+1}), \ldots, x(i - i_k + i_{r-1})$$

are " 1 ", and hence we have

$$\begin{aligned}\text{prob}(\sigma_{i,k}(g) = x(i)) &= \sum_{2|p} C_{r-1}^{p} s^{p} (1 - s)^{r-p-1} \\ &= 1/2 + (1-2s)^{r-1}/2 \\ &= 1/2 + (2\epsilon)^{r-1}/2.\end{aligned}$$

This simple lemma suggests that it will be wise for the cryptanalyst to behave as follows. Choose the r-nomials g(x) to be multiples of the given polynomial f(x), take into consideration 2m + 1 of the syndromes provided by these r-nomials, and revise the signals of the sequence B in accordance with the following rule of majority logic decision.

$$b(i) \longrightarrow b'(i) = \begin{cases} b(i) + 1, & \text{if at least } m + 1 \text{ syndromes are "1",} \\ b(i), & \text{if otherwise,} \end{cases}$$

in the hope that the error rate s' of the sequence A in the resulting sequence B' will be less than the initial error rate s.

In order to see, under which conditions this will be the case, we write

$$p = p(s) = (1 - (1 - 2s)^{r-1})/2, \quad q = 1 - p,$$

and prove the following

Theorem 1. If the number of syndromes used in making the majority logic decision is $n = 2m + 1$, then the error rate of the sequence A in the sequence B' which results from one round of revision will be

$$s' = T_m = p - (1 - 2p) \sum_{k=0}^{m-1} C_{2k+1}^{k} (pq)^{k+1}. \qquad (2)$$

Proof. It is easy to see from the revision algorithm, that $b'(i) = a(i)$ if and only if at least $m + 1$ syndrome values are different from $x(i)$. But, by lemma 1, the probability for a given syndrome value to be different from $x(i)$ is p, so we have

$$s' = T_m = p^n + C_n^1 p^{n-1} q + \ldots + C_n^m p^{m+1} q^m .$$

Further, we have

$$\begin{aligned} T_m &= T_m(p + q) \\ &= p^{m+1} + C_n^1 p^n q + \ldots + C_n^m p^{m+1} q^m \\ &\qquad + p^n q + \ldots + C_n^{m-1} p^{m+1} q^m + C_n^m p^{m+1} q \\ &= (p^{m+1} + C_{n+1}^1 p^n q + \ldots + C_{n+1}^m p^{m+1} q^m)(p + q) + C_n^m p^{m+1} q^{m+1} \\ &= (p^{n+2} + C_{n+2}^1 p^{n+1} q + \ldots + C_{n+1}^n p^{m+2} q^m) + C_{n+1}^m p^{m+2} q^{m+1} + C_n^m p^{m+1} q^{m+1} \\ &= T_{m+1} - (C_{n+2}^{m+1} - C_{n+1}^m) p^{m+2} q^{m+1} + C_n^m p^{m+1} q^{m+1}. \end{aligned}$$

But

$$C_{n+2}^{m+1} = C_{n+1}^{m+1} + C_{n+1}^m = C_{n+1}^m + 2C_n^m ,$$

so we have the following recursive relation

$$T_{m+1} = T_m - (1 - 2p) C_{2m+1}^m (pq)^{m+1} ,$$

which, together with $T_0 = p$, gives rise to (2).

Now, it is easy to see from (2) that, s being fixed, s' decreases as m increases. Furthermore, since

$$T_0 = 1/2 - (2\epsilon)^{r-1} > 1/2 - \epsilon = s$$

and

$$\lim_{m\to\infty} T_m = p - (1 - 2p) \sum_{k=0}^{\infty} C_{\frac{1}{2}}^{k} (pq)^{k+1}$$

$$= p - (1 - 2p)((1 - 4pq)^{-\frac{1}{2}} - 1)/2$$

$$= p - (1 - 2p)((1 - 2p)^{-1} - 1)/2 = 0 ,$$

we see that for each possible initial error rate s there is a critical number mc = mc(s), such that s' will be less than s if and only if m > mc. The following is a table of critical numbers computed for practically tractable values of s, for the case r = 3, where the LS method works the best.

s	mc
0.22	3
0.28	4
0.32	5
0.35	6
0.37	7
0.38	8
0.40	9

(II) Iterated revision and its convergence

The above analysis shows also,that the error rate of the sequence A can be made arbitrarily small,when we make use of a large enough number of syndromes. But such an approach is quite impractical in view of the difficulty in finding the necessary collection of r-nomials, divisible by f(x) and of degrees not too large. A better alternative is to fix the number of syndromes but apply the revision algorithm iteratedly to the segment under consideration,and the problem is that the convergence of such an iterative revision procedure has to be considered.

In order to settle the problem just raised, we consider the polynomial

$$T_m(x) = x - (1 - 2x) \sum_{k=0}^{\infty} C_{2k+1}^{k} (x(1 - x))^{k+1}$$

and prove a couple of simple lemmas about the function p(s) mentioned before as well as about the function

$$s' = f(x) = T_m(p(s)).$$

Lemma 2.The function p(s) is increasing on (0 , 1/2) and maps this interval onto itself.

Proof. In fact, we have

$$p'(s) = (r - 1)(1 - 2s)^{r-2} > 0$$

and

$$p(0) = 0 \quad , \quad p(1/2) = 1/2 \ ,$$

as expected.

Lemma 3. The derivative of the polynomial $T_m(x)$ is

$$T'_m(x) = (m + 1) C_{2m+1}^{m} (x(1 - x))^{m} .$$

Proof. First, as we have noticed before

$$x = (1 - 2x) \sum_{k=0}^{\infty} C_{2k+1}^{k} (x(1 - x))^{k+1} ,$$

whenever $|x| < 1$. So we have

$$\begin{aligned} T_m(x) &= x - (1 - 2x) \sum_{k=0}^{m} C_{2k+1}^{k} (x(1 - x)) \\ &= (1 - 2x) \sum_{k=m}^{\infty} C_{2k+1}^{k} (x(1 - x))^{k+1} \\ &= C_{2m+1}^{m} x^{m+1} \pmod{x^{m+2}} \ . \end{aligned}$$

and hence

$$T'_m(x) = (m + 1) C_{2m+1}^{m} x^{m} \pmod{x^{m+1}} .$$

Further, we have the functional relation

$$T_m(1-x) = 1 - x + (1-2x)\sum_{k=0}^{m-1} C_{2k+1}^{k}(x(1-x))^{k+1}$$

$$= 1 - T_m(x) \ .$$

By differentiation on both sides we have

$$T_m'(x) = (m+1)C_{2m+1}^{m}(1-x)^{m} \quad (\mathrm{mod}\ (1-x)^{m+1}) .$$

But $T_m'(x)$ is a polynomial of degree 2m, so we conclude that

$$T_m'(x) = (m+1)C_{2m+1}^{m}(x(1-x))^{m} \ .$$

Lemma 4. There is a number $\alpha \in (0, 1/2)$ such that

$$f(s) < s \quad , \quad \text{if} \quad 0 < s < \alpha$$

and

$$f(s) > s \quad , \quad \text{if} \quad \alpha < s < 1/2.$$

Proof. Consider the auxiliary function

$$w(s) = f(s) - s \ .$$

We see from

$$p(0) = 0 \quad , \quad p(1/2) = 1/2$$

and the expression for $T_m(x)$ that

$$w(0) = w(1/2) = 0 \ . \qquad (3)$$

Further, we see from

$$w'(s) = T_m'(p(s))p'(s) - 1$$

and

$$T'(0) = 0 \quad , \quad p'(1/2) = 0$$

that

$$w'(0) = w'(1/2) = -1 \qquad (4)$$

(3) and (4) taken together imply that w(s) has at least one zero in the interval (0 , 1/2). On the other hand, if w(s) has in this interval two or more zeroes, then as can be easily seen from (3) and the mean value theorem of differential calculus, w''(s), too, will have at least two zeroes in it. But by direct manipulation we have

$$\begin{aligned} w''(s) &= (r-1)(1-2s)^{r-3}[(r-1)T_m'(p(s))(1-2s)^2 - 2(r-2)T_m'(p(s))] \\ &= (r-1)(1-2s)^{r-3}[(r-1)T_m'(p(s))(1-2p(s)) - 2(r-2)T_m'(p(s))] \\ &= (r-1)(1-2s)^{r-3}K(p(s)), \end{aligned}$$

where

$$K(x) = (m+1)C_{2m+1}^{m}(x(1-x))^{m-1}(ax^2 - ax + b)$$

and

$$a = 4m(r-1) + 2(r-2), \quad b = m(r-1),$$

so we see, by noticing the statement of lemma 2, that w''(s) has only one zero β in the interval (0 , 1/2), satisfying

$$p(\beta) = 1/2 - (1 - 4b/a)^{\frac{1}{2}}/2 .$$

This conclusion means that the function w(s) has only one zero in (0 , 1/2). If we denote this unique zero of w(s) by α , then, by returning to (3) and (4) again, we see w(s) is negative on (0 , α) and positive on (α, 1/2). But this is just what we wanted to prove.

Now we are in a position to prove the convergence theorem for the procedure of iterated revision.

Theorem 2.If we denote by s the error rate of the sequence A in the sequence, which results from the i-th round of revision, then the number sequence $\{s_i\}$ will decrease to 0 if m > mc, and increase to 1/2 if m < mc.

Proof. Suppose m > mc. Then we have by the definition of mc

$$s_1 = f(s_0) < s_0 .$$

So we see from lemma 4 that

$$s_1 < s_0 < \alpha .$$

By applying the same lemma to s we have

$$s_2 = f(s_1) < s_1 < \alpha .$$

By going along with the same argument we have

$$\alpha > s = s_0 > s_1 > \ldots > s_i > \ldots > 0,$$

so we must have

$$\alpha > \lim_{i \to \infty} s_i = s^* \geq 0 ,$$

and s, being a zero of the function w(s) met in the proof of lemma 4, can be nothing else than 0.

The case m < mc, where iterated revision will lead to disastrous garble, can be discussed in exactly the same manner.

(III) An example of applying the LS method

The above analysis of the LS method is by no means rigorous in view of the assumptions made tacitly in computing the probabilities. For a really convincing justification of this method we, in the last run, have to resort to its usefulness in solving concrete problems. Practical problems encountered in cryptanalysis may not yield to the LS method immediately, but can in some cases be reduced to a suitable form, so as to make the method applicable. The following example, though artificial in nature, will be sufficient as an illustration for what we say here.

In the laboratory of the DCS-center people produced a stream X of digital speech by the method of code excited linear prediction followed by vector quantization and turned it, as an experiment, into a stream

$$Y = A + X$$

of incomprehensible enciphered speech, by the help of a linear sequence A with generating polynomial

$$f(x) = 1 + x^{4} + x^{39} .$$

Now, suppose the cryptanalyst knows the polynomial f(x), but has no concept about how to make use of the specific properties of the stream X itself, then for the purpose of recovering it he has to test one after another the $2^{39} - 1$ possible initial states of A,a task far beyond the reach of today's technique.

We show,it is the specific structure of the plaintext stream X, that makes the stream Y easily breakable. In fact, as a result of the slow variation nature of the speech data flow and the imprudent way of encoding it, there exists a sort of betraying correlation between the frames

$$F_0 , F_1 , \ldots , F_i , \ldots$$

of X . A closer examination shows that if we denote the number of " 1 "s in the frame F by w(F) and denote the frame length by l, then for most of the adjacent frame pairs F_i , F_{i+1} we have

$$w(F_i + F_{i+1}) \leqslant 1/4 .$$

And here is the clue we need.In fact,if the cryptanalyst proves to be clever enough to think of going from the stream Y over to the transformed stream

$$Y' = Y + LY = (A + LA) + (X + LX) = A' + X' ,$$

L being the 1-step shift to the left,then he will find himself in the typical situation discussed in the present paper, where A' is linear with the same generating polynomial f(x) as A, while X' is sparse with s = 1/4 Experimentation shows, that by making use of the 9 syndromes provided by the trinomials

$$1 + x^{4} + x^{39} , \quad 1 + x^{8} + x^{78} , \quad 1 + x^{16} + x^{156} ,$$

four rounds of iterated revision applied to Y' suffice to recover A' from a captured segment of length $n < 1500$, and after that the plaintext X can be determined easily by

$$X = Y + (I + L)^{-1} A'.$$

A tape record has been prepared by the same lab for this simple, but instructive instance of successful codebreaking. This example reminds us, in particular,that in order to guarantee safety in communication, not only the algorithm for generating the enciphering signals, but also the data flow to be enciphered, as well as the problem about the suitable way of encoding and enciphering, should be considered carefully.

References

[1] Zeng Kencheng," Phenomena of Key-entropy Leak in Cryptosystems ", unpublished report presented to " Symposium on Problems of Cryptanalysis ", Beijing,1986.