2.8 A General Congruential Generator

The prediction procedure becomes somewhat more involved when the module of a congruential generator is unknown. We abandon the general setting of commutative algebra and use special properties of the rings $\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$, in particular the "canonical" representation of the residue classes of $\mathbb{Z}/m\mathbb{Z}$ by the subset $\{0,\ldots,m-1\} \subseteq \mathbb{Z}$.

Let $X = \mathbb{Z}^r$, $\bar{X} = (\mathbb{Z}/m\mathbb{Z})^r$, $Z = \mathbb{Z}^k$, $\bar{Z} = (\mathbb{Z}/m\mathbb{Z})^k$. The generator uses maps

$$\Phi^{(i)} : X^i \longrightarrow Z \quad \text{for } i \geq h,$$

$$\alpha : \bar{Z} \longrightarrow \bar{X} \quad \text{linear},$$

where α and m are unknown to the cryptanalyst. Identifying the residue classes with their canonical representants we consider $\bar{X}$ as the subset $\{0,\ldots,m-1\}^r$ of X. Then we generate a sequence by the same algorithm as in the previous Section 2.6, and call this procedure a **general congruential generator**, if the evaluation of the maps $\Phi^{(i)}$ is efficient with costs that depend at most polynomially on r, k, and $\log(m)$. In particular there is a bound M for the values of the $\Phi^{(i)}$ on $\{0,\ldots,m-1\}^{ri}$ that is at most polynomial in r, k, and $\log(m)$.

The cryptanalysis proceeds in two phases. In phase one we work over the ring $\mathbb{Z}$ and its quotient field $\mathbb{Q}$, and we determine a multiple $\hat{m}$ of the module m. In phase two we work over the ring $\mathbb{Z}/\hat{m}\mathbb{Z}$. Predicting x_n in this situation can trigger three different events:

- $z_n \notin Z_{n-1}$. Then the module Z_{n-1} (over $\mathbb{Q}$ or $\mathbb{Z}/\hat{m}\mathbb{Z}$) must be enlarged to Z_n, and no prediction is possible for x_n. The cryptanalyst needs some more plaintext.
- The prediction of x_n is correct.
- The prediction of x_n is false. Then the module $\hat{m}$ has to be adjusted.

In phase one Z_{n-1} is the vector space over $\mathbb{Q}$ that is spanned by $z_h,\ldots,z_{n-1}$ (omitting redundant z_i's).

Case 1: $z_n \notin Z_{n-1}$. Then set $Z_n = Z_{n-1} + \mathbb{Q}z_n$. This case can occur at most k times.

Case 2: [Linear relation] $z_n = t_h z_h + \cdots + t_{n-1}z_{n-1}$. Then predict $x_n = t_h x_h + \cdots + t_{n-1}x_{n-1}$ (as element of $\mathbb{Q}^r$).

Case 3: We have an analogous linear relation, but $\hat{x}_n = t_h x_h + \cdots + t_{n-1}x_{n-1}$ differs from x_n. Let $d \in \mathbb{N}$ be the common denominator of $t_h,\ldots,t_{n-1}$. Then

$$d\hat{x}_n = \alpha(dt_h z_h + \cdots + dt_{n-1}z_{n-1}) = \alpha(dz_n) = dx_n$$

in $\bar{X}$, that is $\bmod m$. This shows:

Lemma 8 *(*BOYAR) *The greatest common divisor $\hat{m}$ of the components of $d\hat{x}_n - dx_n$ in case 3 is a multiple of the module m.*

The result of phase one is a multiple $\hat{m} \neq 0$ of the true module m. The expense is:

- at most $k+1$ trials of solving a system of linear equations for up to k unknowns over $\mathbb{Q}$,
- one determination of the greatest common divisor of r integers.

Along the way the procedure correctly predicts a certain number of elements x_n, each time solving a system of linear equations of the same type.

How large can $\hat{m}$ be? For an estimate we need an upper bound M for all components of all $\Phi^{(i)}$ on $\{0, \ldots, m-1\}^{ri} \subseteq X^i$. We use HADAMARD's inequality: For arbitrary vectors $x_1, \ldots, x_k \in \mathbb{R}^k$ we have

$$|\operatorname{Det}(x_1, \ldots, x_k)| \leq \|x_1\|_2 \cdots \|x_k\|_2$$

where $\| \bullet \|_2$ is the Euclidean norm.

Lemma 9 $\hat{m} \leq (k+1) \cdot m \cdot \sqrt{k^k} \cdot M^k$. *In particular* $\log(\hat{m})$ *is bounded by a polynomial in* k, $\log(m)$, $\log(M)$.

Proof. The coefficient vector t is the solution of a system of at most k linear equations for the same number of unknowns. The coefficients z_i of this system are bounded by M. By HADAMARD's inequality and CRAMER's rule the numerators dt_i and denominators d of the solution are bounded by

$$\prod_{i=1}^{k} \sqrt{\sum_{j=1}^{k} M^2} = \prod_{i=1}^{k} \sqrt{kM^2} = \sqrt{k^k} \cdot M^k.$$

Hence the components of $d\hat{x}_n$ are bounded by

$$\|d\hat{x}_n\|_\infty = \|\sum dt_i x_i\|_\infty \leq \sqrt{k^k} \cdot M^k \cdot \sum \|x_i\|_\infty \leq km \cdot \sqrt{k^k} \cdot M^k$$

because m bounds the components of the x_i. We conclude

$$\|d\hat{x}_n - dx_n\|_\infty \leq km \cdot \sqrt{k^k} \cdot M^k + \sqrt{k^k} \cdot M^k \cdot m = (k+1) \cdot m \cdot \sqrt{k^k} \cdot M^k,$$

as claimed. $\diamond$

How does this procedure look in the example of an ordinary linear congruential generator? Here we have

$$z_1 = \begin{pmatrix} x_0 \\ 1 \end{pmatrix}, z_2 = \begin{pmatrix} x_1 \\ 1 \end{pmatrix}, z_3 = \begin{pmatrix} x_2 \\ 1 \end{pmatrix}, \ldots$$

If $x_1 = x_0$, then we have the trivial case of a constant sequence. Otherwise z_3 is a rational linear combination $t_1 z_1 + t_2 z_2$. Solving the system

$$\begin{aligned} x_0 t_1 + x_1 t_2 &= x_2, \\ t_1 + t_2 &= 1 \end{aligned}$$

yields

$$t = \frac{1}{d} \cdot \begin{pmatrix} -x_2 + x_1 \\ x_2 - x_0 \end{pmatrix} \quad \text{with } d = x_1 - x_0.$$

From this we derive the prediction

$$\hat{x}_3 = t_1 x_1 + t_2 x_2 = \frac{-x_2 x_1 + x_1^2 + x_2^2 - x_2 x_0}{x_1 - x_0} = \frac{(x_2 - x_1)^2}{x_1 - x_0} + x_2.$$

Hence $d(\hat{x}_3 - x_3) = (x_2 - x_1)^2 - (x_1 - x_0)(x_3 - x_2) = y_2^2 - y_1 y_3$ where (y_i) is the sequence of differences. If $\hat{x}_3 = x_3$, then we must continue this way. Otherwise we get, see Lemma 6,

$$m | \hat{m} = |y_1 y_3 - y_2^2|.$$

For our concrete standard example, where $x_0 = 2134$, $x_1 = 2160$, $x_2 = 6905$, $x_3 = 3778$, $y_1 = 26$, $y_2 = 4745$, $y_3 = -3127$, this general approach gives

$$\hat{m} = 4745^2 + 26 \cdot 3127 = 22596327.$$

A closer look, using Lemma 8 directly, even yields

$$t_1 = -\frac{365}{2},\ t_2 = \frac{367}{2},\ \hat{x}_3 = \frac{1745735}{2},\ \hat{m} = 2 \cdot (\hat{x}_3 - x_3) = 1738179.$$

In phase two of the algorithm we execute the same procedure but over the ring $\hat{R} = \mathbb{Z}/\hat{m}\mathbb{Z}$. However we can't simply reduce $\bmod \hat{m}$ the rational numbers from phase one. Hence we restart at z_h. Again we distinguish three cases for each single step:

Case 1: $z_n \notin \hat{Z}_{n-1} = \hat{R} z_h + \cdots + \hat{R} z_{n-1}$. Then set $\hat{Z}_n = \hat{Z}_{n-1} + \hat{R} z_n$ (and represent this $\hat{R}$-module by a non-redundant system $\{z_{j_1}, \ldots, z_{j_l}\}$ of generators where $z_{j_l} = z_n$). We can't predict x_n (but have to get it from somewhere else).

Case 2: $z_n = t_h z_h + \cdots + t_{n-1} z_{n-1}$. Then predict $x_n = t_h x_h + \cdots + t_{n-1} x_{n-1}$ (as an element of $\hat{X} = (\mathbb{Z}/\hat{m}\mathbb{Z})^r$). The prediction turns out to be correct.

Case 3: The same, but now the predicted value $\hat{x}_n = t_h x_h + \cdots + t_{n-1} x_{n-1}$ differs from x_n in $\hat{X}$. Then considering $\hat{x}_n - x_n$ as an element of $\mathbb{Z}^r$ we show:

Lemma 10 *In case 3 the greatest common divisor d of the coefficients of $\hat{x}_n - x_n$ is a multiple of m, but not a multiple of $\hat{m}$.*

Proof. It is a multiple of m since $\hat{x}_n \bmod m = x_n$. It is not a multiple of $\hat{m}$ since otherwise $\hat{x}_n = x_n$ in $\hat{X}$. $\diamond$

In case 3 we replace $\hat{m}$ by the greatest common divisor of d and $\hat{m}$ and reduce mod $\hat{m}$ all the former z_j. The lemma tells us that the new $\hat{m}$ is properly smaller than the old one.

By Lemma 9 case 3 can't occur too often, the number of occurences is polynomially in k, $\log(m)$, and $\log(M)$. If we already hit the true m this case can't occur any more. Case 1 may occur at most $\log_2(\#(\mathbb{Z}/\hat{m}\mathbb{Z})^k) = k \cdot \log_2(\hat{m})$ times in phase 2 by Proposition 5, and this bound is polynomial in k, $\log(m)$, and $\log(M)$.

> **Note.** There is a common aspect of phases one and two: In both cases we use the full quotient ring. The full quotient ring of $\mathbb{Z}$ is the quotient field $\mathbb{Q}$. In a residue class ring $\mathbb{Z}/m\mathbb{Z}$ the non-zero-divisors are exactly the elements that are coprime with m, hence the units. Thus $\mathbb{Z}/m\mathbb{Z}$ is its own full quotient ring.

For the concrete standard example we had $\hat{m} = 1738179$ after phase one, and now have to solve mod $\hat{m}$ the system (1) of linear equations. Since the determinant -26 is coprime with $\hat{m}$ we already have $Z_2 = \hat{R}^2$, and know that case 1 will never occur. The inverse of -26 is 66853 (in $\mathbb{Z}/\hat{m}\mathbb{Z}$), so from $-26\,t_1 = 4745$ we get $t_1 = 868907$. Hence $t_2 = 1 - t_1 = 869273$, and $\hat{x}_3 = 1_1 x_1 + t_2 x_2 = 3778$ is a correct prediction.

In the next step we calculate new coefficients t_1 and t_2 for the linear combination $z_4 = t_1 z_1 + t_2 z_2$. We solve (in $\mathbb{Z}/\hat{m}\mathbb{Z}$)

$$\begin{aligned} 2134\,t_1 + 2160\,t_2 &= 3778, \\ t_1 + t_2 &= 1. \end{aligned}$$

Eliminating t_2 yields $-26\,t_1 = 1618$, hence $t_1 = 401056$, and thus $t_2 = 1337124$, as well as $\hat{x}_4 = 1_1 x_1 + t_2 x_2 = 302190$. Since $x_4 = 8295$ we are in case 3 and must adjust $\hat{m}$:

$$\gcd(\hat{x}_4 - x_4, \hat{m}) = \gcd(293895, 1738179) = 8397.$$

Now $\hat{m} < 2x_2$. Thus from now on only case 2 will occur. This means that we'll predict all subsequent elements correctly.

A **prediction method** for a general congruential generator is an algorithm that gets the initial values $x_0, \ldots, x_{h-1}$ as input, then successively produces predictions of $x_h, x_{h+1}, \ldots$, and compares them with the true values; in the case of a mistake it adjusts the parameters using the respective true value.

A prediction method is **efficient** if

1. the cost of predicting each single x_n is polynomial in r, k, and $\log(m)$,

2. the number of false predictions is bounded by a polynomial in r, k, and $\log(m)$, as is the cost of adjusting the parameters in the case of a mistake.

The Boyar/Krawczyk algorithm that we considered in this section fulfils requirement 2. It also fulfils requirement 1 since solving systems of linear equations over residue class rings $\mathbb{Z}/m\mathbb{Z}$ is efficient (as shown in Section 9.2 of Part I). Thus we have shown:

Theorem 2 *For an arbitrary (efficient) general congruential generator the* Boyar/Krawczyk *algorithm is an efficient prediction method.*

A simple concrete example shows the application to a non-linear congruential generator. Suppose a quadratic generator of the form

$$x_n = ax_{n-1}^2 + bx_{n-1} + c \bmod m$$

outputs the sequence

$$x_0 = 63,\ x_1 = 96,\ x_2 = 17,\ x_3 = 32,\ x_4 = 37,\ x_5 = 72.$$

We set $X = \mathbb{Z}$, $Z = \mathbb{Z}^3$, $h = 1$. In phase one the vectors

$$z_1 = \begin{pmatrix} 3969 \\ 63 \\ 1 \end{pmatrix} z_2 = \begin{pmatrix} 9216 \\ 96 \\ 1 \end{pmatrix} z_3 = \begin{pmatrix} 289 \\ 17 \\ 1 \end{pmatrix}$$

span $\mathbb{Q}^3$ since the coefficient matrix is the Vandermonde matrix with determinant 119922. Solving

$$z_4 = \begin{pmatrix} 1024 \\ 32 \\ 1 \end{pmatrix} = t_1z_1 + t_2z_2 + t_3z_3$$

yields

$$t_1 = \frac{160}{253}, \quad t_2 = -\frac{155}{869}, \quad t_3 = \frac{992}{1817},$$

with common denominator $d = 11 \cdot 23 \cdot 79 = 19987$. The algorithm predicts

$$\hat{x}_4 = \frac{1502019}{19987} \neq x_4.$$

Hence the first guessed module is

$$\hat{m} = d\hat{x}_4 - dx_4 = 762500,$$

and phase one is completed. Now we have to solve the same system of linear equations over $\mathbb{Z}/\hat{m}\mathbb{Z}$. Here the determinant is a zero divisor. We get two solutions, one of them being

$$t_1 = 156720, \quad t_2 = 719505, \quad t_3 = 648776.$$

Thus we predict the correct value

$$\hat{x}_4 = 156720 \cdot 96 + 719505 \cdot 17 + 648776 \cdot 32 \bmod 763500 = 37.$$

We are in case 2, and continue with predicting x_5: The system

$$z_5 = \begin{pmatrix} 1369 \\ 37 \\ 1 \end{pmatrix} = t_1 z_1 + t_2 z_2 + t_3 z_3$$

has two solutions, one of them being

$$t_1 = 2010\,, \quad t_2 = 558640\,, \quad t_3 = 201851\,,$$

hence

$$\hat{x}_5 = 136572\,, \quad \hat{x}_5 - x_5 = 136500\,.$$

We are in case 3 and adjust $\hat{m}$ to

$$\gcd(762500, 136500) = 500.$$

This exhausts the known values. Because all z_i are elements of $\hat{Z}_3 = \hat{R}z_1 + \hat{R}z_2 + \hat{R}z_3 \neq \hat{R}^3$ case 1 remains a possibility for the following steps. Since $x_0, \ldots, x_5$ are smaller than half the current module $\hat{m}$ also case 3 remains possible. In particular maybe we have to adjust the module furthermore.

Trying to predict x_6 we get $(\bmod 500)$

$$t_1 = 240\,, \quad t_2 = 285\,, \quad t_3 = 476\,, \quad x_6 = 117\,.$$

Exercise. What happens in the concrete standard example if after phase 1 we continue with the value $\hat{m} = 22596327$?