MATHEMATICS OF COMPUTATION
VOLUME 53, NUMBER 187
JULY 1989, PAGES 343–354

The Exact Determination of Rectangle Discrepancy for Linear Congruential Pseudorandom Numbers

By Lothar Afflerbach and Rainer Weilbächer

Abstract. Up to now, the rectangle discrepancy of linear congruential pseudorandom number generators could be exactly calculated only in some simple cases for a small number of generated points. Here an algorithm for the exact determination of the two-dimensional rectangle discrepancy is presented which is practicable for large generators and requires less computation time. The algorithm is based on special properties of linear congruential generators.

1. Introduction. For a set $G_k^{(n)}$ containing exactly k points of $[0,1)^n$ the n-dimensional (rectangle) discrepancy is defined by

$$D_k^{(n)} = \sup_{R \in \mathscr{R}_{[0,1)}^{(n)}} \left| \frac{1}{k} N(R) - V(R) \right|, \tag{1}$$

where

$$\mathscr{R}_{[0,1)}^{(n)} = \{[s_1, t_1] \times \cdots \times [s_n, t_n] : 0 \le s_i \le t_i < 1,\ i = 1, \ldots, n\}$$

is the set of all closed n-dimensional rectangles with sides parallel to the axes lying in $[0,1)^n$. $N(R)$ and $V(R)$ denote the number of points of $G_k^{(n)}$ lying in R and the volume of R, respectively. We always have $0 \le D_k^{(n)} \le 1$. The discrepancy of a sequence of random numbers used in Monte-Carlo-Integrations is of interest because it appears in upper error bounds (see [6]). In practice, usually pseudorandom numbers generated by linear congruential generators are employed. Such a generator produces a sequence $\{x_i\}$ of m different integers with an integral initial value x_0 by the recurrence

$$x_i \equiv a \cdot x_{i-1} + b \pmod{m}, \qquad 0 \le x_i \le m-1,\ i = 1, 2, 3, \ldots, \tag{2}$$

if the three integers m (modulus), a (multiplier) and b (increment) are properly chosen, which is assumed in the following (see [7, Chapter 3.2, Theorem A]). The fractions $y_i = x_i/m$ are used as random numbers uniformly distributed on $[0,1)$.

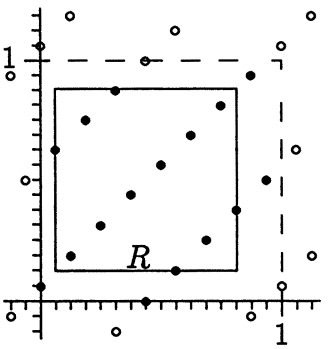
Up to now, the discrepancy of those generators could be exactly calculated only in some simple cases for a small number of points in $G_k^{(n)}$. This was done by the examination of all open and closed rectangles with points of $G_k^{(n)}$ on the borders. Taking the set $G_m^{(2)}$ of all generated pairs $\binom{y_i}{y_{i+1}}$ from a generator (2), U. Dieter determined the deviation $|\frac{1}{k}N(R) - V(R)|$ for the two-dimensional rectangles R

Received April 5, 1988; revised September 16, 1988.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65C10; Secondary 65C05, 11H06.

using Dedekind sums (see [4], [5]). In this way, the search for the rectangle with maximal deviation is expensive, so that estimates for the discrepancy were given instead of the exact value. H. Niederreiter calculated estimates for the discrepancy for $n \geq 2$ as well as estimates for the discrepancy of subsequences of linear congruential generators (see [10], [11]).

Here, an algorithm for the exact determination of the two-dimensional rectangle discrepancy is presented. We examine the set $G_m^{(2)}$ of all generated pairs of a linear congruential generator with period length m. As is well known, this set has a lattice structure (see [2], [5], [8], [9] and [1], [12] for the sublattice structure). A change of the increment b of the generator yields only a shift of the lattice (that is the periodic continuation of the generated pairs). The assessment of the lattice structure with the help of reduced bases or the spectral test (see [3], [7]) thereby does not change. The bounds of discrepancy are independent of the increment, too. But the discrepancy may depend on the increment as the following example shows.

Example 1. The linear congruential generator with $m = 16$, $a = 9$ and $b = 1$ has discrepancy $\frac{48}{256}$, which is obtained for the rectangle R shown below. If we change the increment b to 3, the lattice is shifted by the vector $\binom{0}{2/16}$ so that the rectangle R lies at the border of $[0,1)^2$. Therefore, we have a smaller discrepancy, namely $\frac{47}{256}$ in this case.



In order to avoid these difficulties at the borders of $[0,1)^2$, we first examine the "lattice discrepancy" $\hat{D}_m^{(2)}$, that is, the maximal discrepancy which appears among all discrepancies of the lattice points which lie in $[0,1)^2$ when the lattice is shifted by integral multiples of $\binom{0}{1/m}$. It is convenient to transform the coordinates by the factor m so that we get m integral lattice points in $[0,m)^2$. We will then examine

$$m^2 \hat{D}_m^{(2)} := \sup_{R \in \hat{\mathscr{R}}_{[0,m)}^{(2)}} |mN(R) - V(R)| \tag{3}$$

with

$$\hat{\mathscr{R}}_{[0,m)}^{(2)} = \{[s_1,t_1] \times [s_2,t_2]\colon\ 0 \leq s_1 \leq t_1 < m,\ -m < s_2 \leq t_2 < m,\ t_2 - s_2 < m\},$$

where the lattice is fixed in such a way that a lattice point lies at the origin. In Section 4 the algorithms developed in Sections 2 and 3 are modified to calculate $D_m^{(2)}$ of a given linear congruential generator.

2. A Theorem About $\hat{D}_m^{(2)}$. The special lattice structure of a linear congruential generator allows us to replace a shift of the lattice in vertical direction by a shift in horizontal direction. So we can replace $\hat{\mathscr{R}}_{[0,m)}^{(2)}$ in Eq. (3) by

$$\mathscr{R}_{(-m,m)}^{(2)} = \{[s_1,t_1]\times[s_2,t_2]\colon -m < s_1 \le t_1 < m,\ 0 \le s_2 \le t_2 < m,\ t_1 - s_1 < m\}.$$

Then we follow U. Dieter and J. H. Ahrens [5], replacing the supremum taken over $\mathscr{R}_{(-m,m)}^{(2)}$ by a maximum taken over all open and all closed rectangles with lattice points on their borders. Furthermore, we can suppose that the lattice points on opposite borders lie symmetrically to the midpoint of the rectangle. Otherwise, corresponding rectangles could be found with greater deviations. If we shift any rectangle by a linear combination of lattice vectors, the deviation $|mN(\cdot)-V(\cdot)|$ has the same value. Thus we examine rectangles $[-l,r]\times[0,h]$ and $(-l,r)\times(0,h)$, where l,r,h are integers with $l,r \ge 0$, $l+r < m$ and $0 \le h < m$. Special cases, as the case of open rectangles R with $l+r=m$ or $h=m$ which yield $|mN(R)-V(R)| \le 2m$, are negligible. Setting $l_1 = l$ and $r_1 = r$, we can calculate the lattice points $\binom{-l_1}{l_2}$ and $\binom{r_1}{r_2}$ on the left and right boundary by

$$l_2 \equiv -al_1 \pmod m \quad \text{and} \quad r_2 \equiv ar_1 \pmod m, \tag{4}$$

respectively (note that the lattice is shifted by $\binom{0}{-b}$ so that a lattice point lies at the origin). Because of the supposed symmetry we have the lattice point $\binom{h_1}{h_2}$ on the upper border of the rectangle with

$$h_1 = r_1 - l_1 \quad \text{and} \quad h_2 = l_2 + r_2.$$

This determines the height h_2 of the rectangle. We can get $h_2 > m$, but we will show in the proof of Theorem 1 that this does not lead to any falsification. With relationship (4) we examine the sets

$$\mathscr{A} := \{[-l_1,r_1]\times[0,l_2+r_2], 1 \le l_1,r_1 \le m-1,\ l_1+r_1 < m\}$$

of all closed rectangles with four lattice points (these are $\binom{0}{0}$, $\binom{-l_1}{l_2}$, $\binom{r_1}{r_2}$ and $\binom{r_1-l_1}{l_2+r_2}$) on their borders,

$$\mathscr{B} := \{[0,r_1]\times[0,r_2],\ [-l_1,0]\times[0,l_2],\ 1 \le l_1,r_1 \le m-1\}$$

of all closed rectangles with two lattice points lying on two opposite edges, and the set

$$\mathscr{C} := \{(-l_1,r_1)\times(0,l_2+r_2),\ 1 \le l_1,r_1 \le m-1,\ l_1+r_1 < m\}$$

of all open rectangles with four lattice points on their borders. If we have a rectangle R in $\mathscr{A}$ (or $\mathscr{B}$) for which $mN(R)-V(R)$ is negative, we examine the corresponding rectangle (or a somewhat bigger one) in $\mathscr{C}$ and get a greater deviation. On the other hand, for $R \in \mathscr{C}$ with $mN(R)-V(R) > 0$ we examine the corresponding rectangle in $\mathscr{A}$ to get a greater deviation. Because of this, we are interested in

$$\max\left(\max_{R\in\mathscr{A}\cup\mathscr{B}}(mN(R)-V(R)),\ \max_{R\in\mathscr{C}}(V(R)-mN(R))\right). \tag{5}$$

In the following we determine the number $P(l_1,r_1)$ of lattice points in a rectangle $R=[-l_1,r_1]\times[0,l_2+r_2]$ as a function of l_1 and r_1. Because of the lattice structure of the linear congruential generator the lattice points lie on the "lines"

$$g_\mu = \left\{\vec{p}\colon \vec{p} = \lambda\begin{pmatrix}1\\a\end{pmatrix} - \mu\begin{pmatrix}0\\m\end{pmatrix}, \lambda \text{ integer}\right\}, \quad \mu \text{ integer}.$$

The smallest and the largest parameter μ_1 and μ_2 of the lines g_μ which cut the rectangle $[-l_1, r_1] \times [0, l_2 + r_2]$ are given by

$$\mu_1 = \left[-\frac{al_1}{m}\right] \quad \text{and} \quad \mu_2 = \left[\frac{ar_1}{m}\right],$$

respectively, where $[x]$ denotes the greatest integer less than or equal to x. For fixed μ $(\mu_1 \le \mu \le \mu_2)$ the bounds $\lambda_1(\mu)$ and $\lambda_2(\mu)$ for which all lattice points $\vec{p} = \lambda\binom{1}{a} - \mu\binom{0}{m}$ with $\lambda_1(\mu) \le \lambda \le \lambda_2(\mu)$ lie in the rectangle are given by

$$\lambda_1(\mu) = \begin{cases} -l_1 & \text{for } \mu = \mu_1, \\ \lceil \frac{\mu m}{a} \rceil & \text{for } \mu_1 < \mu \le \mu_2 \end{cases}$$

and

$$\lambda_2(\mu) = r_1 - l_1 - \lambda_1(\mu_1 + \mu_2 - \mu) \quad \text{for } \mu_1 \le \mu \le \mu_2,$$

where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x. For $\mu_1 \le \mu \le \mu_2$ we have $\lambda_2(\mu) - \lambda_1(\mu) + 1$ lattice points lying in the intersection of g_μ and the rectangle. Summation yields

$$\begin{aligned} P(l_1, r_1) &= \sum_{\mu=\mu_1}^{\mu_2} (\lambda_2(\mu) - \lambda_1(u) + 1) \\ &= (\mu_2 - \mu_1 + 1)(r_1 - l_1 + 1) - \sum_{\mu=\mu_1}^{\mu_2} (\lambda_1(\mu_1 + \mu_2 - \mu) + \lambda_1(\mu)) \\ &= (\mu_2 - \mu_1 + 1)(r_1 - l_1 + 1) + 2l_1 - 2 \sum_{\mu=\mu_1+1}^{\mu_2} \left\lceil \frac{\mu m}{a} \right\rceil \\ &= (\mu_2 - \mu_1 + 1)(r_1 - l_1 + 1) + 2l_1 - 2 \sum_{\mu=\mu_1+1}^{\mu_2} \left(\left[\frac{\mu m}{a} \right] + 1 \right) + 2, \end{aligned}$$

because $\lceil \frac{\mu m}{a} \rceil = [\frac{\mu m}{a}] + 1$ for $\mu \neq 0$ and $|\mu| < a$, since a and m are relatively prime. Using (4), we calculate the area $A(l_1, r_1)$ of the rectangle by

$$\begin{aligned} A(l_1, r_1) &= (r_1 + l_1)(l_2 + r_2) \\ &= a(r_1^2 - l_1^2) - m(r_1 + l_1) \left(\left[\frac{ar_1}{m} \right] + \left[-\frac{al_1}{m} \right] \right). \end{aligned}$$

Now we define the "deviation function" D by

$$D(l_1, r_1) := mP(l_1, r_1) - A(l_1, r_1). \tag{6}$$

Using the representation of $P(l_1, r_1)$ and $A(l_1, r_1)$ and selecting the terms depending on l_1 and r_1, respectively, we get

$$D(l_1, r_1) = F(l_1) + G(r_1) \tag{7}$$

with $F(0) = 0$ and

$$F(l) = m \left(\left[\frac{al}{m} \right] - l + 1 - 2l \left[\frac{al}{m} \right] + 2 \sum_{\mu=1}^{[\frac{al}{m}]} \left[\frac{\mu m}{a} \right] \right) + al^2, \qquad l = 1, \ldots, m-1, \tag{8}$$

and

$$G(r) = m\left(2r\left[\frac{ar}{m}\right] + r + 1 - \left[\frac{ar}{m}\right] - 2\sum_{\mu=1}^{[\frac{ar}{m}]}\left[\frac{\mu m}{a}\right]\right) - ar^2, \qquad r = 0, \ldots, m-1.$$

LEMMA 1. *The following equations hold:*

$$\begin{aligned} &(\alpha) \quad F(l) = F(m-l), && l = 1, \ldots, m-1, \\ &(\beta) \quad G(r) = G(m-r), && r = 1, \ldots, m-1, \\ &(\gamma) \quad F(l) + G(l) = 2m, && l = 1, \ldots, m-1. \end{aligned}$$

Proof. According to (8) we have

$$F(m-l) = m\left(\left[\frac{am-al}{m}\right] - m + l + 1 - 2(m-l)\left[\frac{am-al}{m}\right]\right) + a(m-l)^2 + 2m\sum_{\mu=1}^{[\frac{am-al}{m}]}\left[\frac{\mu m}{a}\right].$$

Because of $[\frac{am-al}{m}] = a - [\frac{al}{m}] - 1$ the sum (without the factor $2m$) can be written as

$$\sum_{\mu=1}^{a}\left[\frac{\mu m}{a}\right] - \sum_{\mu=a-[\frac{al}{m}]}^{a}\left[\frac{\mu m}{a}\right].$$

Here, $\mu m \pmod a$ runs through all residues modulo a when μ runs from 1 to a, because a and m are relatively prime. Therefore,

$$\begin{aligned} \sum_{\mu=1}^{a}\left[\frac{\mu m}{a}\right] &= \frac{1}{a}\sum_{\mu=1}^{a}(\mu m - (\mu m \bmod a)) \\ &= \frac{m(a+1)}{2} - \frac{1}{a}\sum_{\mu=1}^{a-1}\mu = \frac{m(a+1)}{2} - \frac{a-1}{2} \end{aligned}$$

and

$$\sum_{\mu=a-[\frac{al}{m}]}^{a}\left[\frac{\mu m}{a}\right] = \sum_{\mu=0}^{[\frac{al}{m}]}\left[\frac{(a-\mu)m}{a}\right] = m\left(\left[\frac{al}{m}\right] + 1\right) - \left[\frac{al}{m}\right] - \sum_{\mu=1}^{[\frac{al}{m}]}\left[\frac{\mu m}{a}\right]$$

(note that $[-\frac{\mu m}{a}] = -[\frac{\mu m}{a}] - 1$ for $1 \le \mu \le [\frac{al}{m}]$). Equation (γ) can easily be shown. (β) follows from (α) and (γ). □

With the help of Lemma 1 we will prove the following theorem.

THEOREM 1. *The discrepancy* $\hat{D}_m^{(2)}$ *is given by*

$$m^2\hat{D}_m^{(2)} = 2m + \max_{1 \le l \le [\frac{m}{2}]} F(l) - \min_{1 \le l \le [\frac{m}{2}]} F(l), \tag{9}$$

where F *is defined in* (8).

Proof. We have $N(R) = P(l_1, r_1)$ for closed rectangles $R \in \mathscr{A} \cup \mathscr{B}$ and $N(R) = P(l_1, r_1) - 4$ for open rectangles $R \in \mathscr{C}$ and always $V(R) = A(l_1, r_1)$.

In order to determine the maximal deviation $|mN(R) - V(R)|$ for all rectangles $R \in \mathscr{A} \cup \mathscr{B} \cup \mathscr{C}$, we have to calculate

$$D_{\max} := \max_{1 \le l,r \le m-1} (D(l,r), 4m - D(l,r), D(0,r), D(l,0)).$$

We can ignore the restriction $l + r < m$ because of formula (7) and the symmetry of the functions F and G (Lemma 1). Thus $D_{\max}$ is equal to the expression in (5). Using Lemma 1 (γ), we get

$$\begin{aligned} \max_{1 \le l,r \le m-1} D(l,r) &= \max_{1 \le l \le m-1} F(l) + \max_{1 \le r \le m-1} G(r) \\ &= 2m + \max_{1 \le l \le m-1} F(l) - \min_{1 \le l \le m-1} F(l) \end{aligned}$$

and

$$4m - \min_{1 \le l,r \le m-1} D(l,r) = 2m + \max_{1 \le l \le m-1} F(l) - \min_{1 \le l \le m-1} F(l).$$

Since $F(0) = 0, F(1) = a$ and $\max_{1 \le l \le m-1} F(l) \ge F(1) > 0$, we get

$$\max_{1 \le r \le m-1} D(0,r) < 2m + \max_{1 \le l \le m-1} F(l) - \min_{1 \le l \le m-1} F(l) \tag{10}$$

and

$$\max_{1 \le l \le m-1} D(l,0) < 2m + \max_{1 \le l \le m-1} F(l) - \min_{1 \le l \le m-1} F(l), \tag{11}$$

since $\min_{1 \le l \le m-1} F(l) \le F(1) = a < m$. Using the symmetry of F once more, we can restrict l by $1 \le l \le [\frac{m}{2}]$ instead of $1 \le l \le m-1$. The formulas (10) and (11) show that the maximal deviation $D_{\max}$ is achieved for rectangles R lying in $\mathscr{A}$ or $\mathscr{C}$. To finish the proof we have to show that the rectangles $R \in \mathscr{A} \cup \mathscr{C}$ with height $h_2 = l_2 + r_2 > m$ do not falsify the calculation (the trivial case $l_2 + r_2 = m$ was examined earlier). For the rectangle $R = [-l_1, r_1] \times [0, l_2 + r_2] \in \mathscr{A}$ with $l_2 + r_2 > m$ we have the rectangle $R' := (-r_1, l_1) \times (0, 2m - (l_2 + r_2)) \in \mathscr{C}$ with height $2m - (l_2 + r_2) < m$, and the deviation $|mN(R') - V(R')|$ is equal to

$$\begin{aligned} 4m - D(r_1, l_1) &= 4m - F(r_1) - G(l_1) \\ &= F(l_1) + G(r_1) = D(l_1, r_1), \end{aligned}$$

which is the same deviation as for the rectangle R. Analogously we examine for a rectangle $R = (-l_1, r_1) \times (0, l_2 + r_2) \in \mathscr{C}$ with $l_2 + r_2 > m$ the corresponding rectangle $R' := [-r_1, l_1] \times [0, 2m - (l_2 + r_2)] \in \mathscr{A}$ with $2m - (l_2 + r_2) < m$, which yields the same deviation as R. This proves the theorem. □

3. Algorithms for Calculating $\hat{D}_m^{(2)}$. On the basis of Theorem 1 we could formulate an algorithm to compute the discrepancy $\hat{D}_m^{(2)}$. We would calculate $F(l)$ for $l = 1, \dots, [\frac{m}{2}]$ according to formula (8) and determine the maximal and minimal value of F. Using formula (9), we would get $\hat{D}_m^{(2)}$. In order to get a faster algorithm, we examine the differences $Z(l) := F(l+1) - F(l)$.

LEMMA 2. *For $Z(l) = F(l+1) - F(l)$ with F defined in (8), the following equation holds:*

$$Z(l) = \begin{cases} 2(al \bmod m) - (m-a) & \text{if } al \bmod m < m - a, \\ 2m - a & \text{if } al \bmod m = m - a, \\ 2(al \bmod m) - (2m-a) & \text{if } al \bmod m > m - a. \end{cases} \tag{12}$$

Proof. Using (8), we have

$$Z(l) = m\left(\left[\frac{a(l+1)}{m}\right] - \left[\frac{al}{m}\right] - 1 - 2(l+1)\left[\frac{a(l+1)}{m}\right] + 2l\left[\frac{al}{m}\right]\right)$$
$$+ 2m\left(\sum_{\mu=1}^{[\frac{a(l+1)}{m}]}\left[\frac{\mu m}{a}\right] - \sum_{\mu=1}^{[\frac{al}{m}]}\left[\frac{\mu m}{a}\right]\right) + 2al + a.$$

The difference of the two sums is nonzero if and only if $[\frac{a(l+1)}{m}] = [\frac{al}{m}] + 1$. Using $d := [\frac{a(l+1)}{m}] - [\frac{al}{m}]$, we can write $d \cdot [([\frac{al}{m}] + 1) \cdot \frac{m}{a}]$ for the difference of the sums and get

$$Z(l) = 2\left(al - m\left[\frac{al}{m}\right]\right) + 2md\left(-l + \left[\frac{m}{a} + \frac{m}{a}\left[\frac{al}{m}\right]\right]\right) - md - m + a$$
$$= 2(al \bmod m) + 2md\left[\frac{(-al) \bmod m}{a}\right] - md - m + a.$$

For $d = 1$ we have $al \bmod m \geq m - a$, which implies $(-al) \bmod m \leq a$, with equality if and only if $l = m - 1$. This completes the proof. □

With the help of Lemma 2 we can formulate the following algorithm for the determination of $\hat{D}_m^{(2)}$.

ALGORITHM 1.

1. Input modulus m and multiplier a.
2. Calculate $F(l)$ for $l = 1, \ldots, [\frac{m}{2}]$ by $F(1) = a$ and $F(l) \leftarrow F(l-1) + Z(l-1)$ with $Z(l-1)$ calculated by formula (12).
3. Output discrepancy $\hat{D}_m^{(2)} \leftarrow (2m + \max_{1 \leq l \leq [\frac{m}{2}]} F(l) - \min_{1 \leq l \leq [\frac{m}{2}]} F(l))/m^2$.

The following figures show in two examples $F(l)$ for $l = 1, \ldots, [\frac{m}{2}]$ (linearly interpolated).

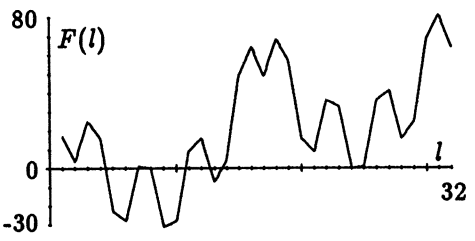


FIGURE 1
$F(l)$ *for* $m = 64, a = 17$

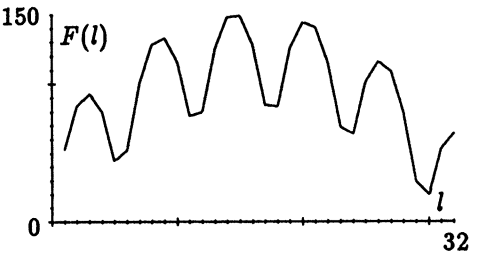


FIGURE 2
$F(l)$ *for* $m = 64, a = 53$

The examples show that the increases $Z(l)$ often have the same sign for many successive values of l. Therefore we search for those values where the sign of the function Z is changed. At these points we have local extremes of the function F. The search for these points leads to a distinction into many cases with many complex formulas which are not very useful for an algorithm. In the following theorem some simple formulas are given to describe all candidates for the values $(\leq [\frac{m}{2}])$ where the sign of the function Z changes.

THEOREM 2. *Let*

$$\underline{l}_k := \left[\frac{m+a+2mk}{2a}\right], \qquad k = 0, \dots, \underline{k}, \text{ with } \underline{k} := \left[\frac{a-1}{2}\right], \tag{13}$$

and

$$\bar{l}_k := \left[\frac{m(k+1)}{a}\right], \qquad k = 0, \dots, \bar{k}, \text{ with } \bar{k} := \left[\frac{a}{2}\right]. \tag{14}$$

Then for $l = 1, \cdots, [\frac{m}{2}]$ *we have to expect a local minimum of* $F(l)$ *only at the points* $\underline{l}_k$ *defined in* (13) *and a local maximum of* $F(l)$ *only at the points* $\bar{l}_k$ *or* $\bar{l}_k + 1$ *with* $\bar{l}_k$ *defined in* (14).

Proof. The case $al \bmod m = m - a$, which means $l = m - 1$, can be excluded in the following because the search for extremes is restricted to $l \le [\frac{m}{2}]$.

For $al \bmod m < m - a$ we have by (12) that

$$Z(l) \le 0 \quad \text{if } 0 \le al \bmod m \le \frac{m-a}{2} \tag{15}$$

and

$$Z(l) > 0 \quad \text{if } \frac{m-a}{2} < al \bmod m < m - a. \tag{16}$$

For $al \bmod m > m - a$ we have by (12) that

$$Z(l) \le 0 \quad \text{if } m - a < al \bmod m \le m - \frac{a}{2} \tag{17}$$

and

$$Z(l) > 0 \quad \text{if } m - \frac{a}{2} < al \bmod m < m. \tag{18}$$

The formulas (15) and (18) show that the function Z has a periodic behavior with "period $\frac{m}{a}$". Therefore, the function F increases and decreases with period $\frac{m}{a}$. Now let us study such a period. If we write $al \bmod m = al - km$ with $k = [\frac{al}{m}]$, all values of l which belong to the same period have the same value of k.

(i) In order to find candidates for a minimum of F we suppose $Z(l) > 0$. In the case of formula (16) we examine $Z(l-1)$, and if we are also in the case (16), we will take $l - 1$ instead of l and so on. In this way we search for the value $\underline{l}_k$ with the property

$$a(\underline{l}_k - 1) - km \le \frac{m-a}{2} < a\underline{l}_k - km,$$

which leads to

$$\underline{l}_k = \left[\frac{m+a+2mk}{2a}\right].$$

In the case (18) we have $Z(l-1) > 0$ (case (16)) or $Z(l-1) \le 0$ (case (15)) but never case (17) for $l - 1$. Therefore, the candidates for a minimum are always given by $\underline{l}_k$ defined in the above formula. The condition $\underline{l}_k \le \frac{m}{2}$ leads to $k \le \frac{a-1}{2}$.

(ii) In order to find candidates for a maximum of F, we suppose $Z(l) \le 0$. In the case (17) we determine $\bar{l}_k$ by

$$a(\bar{l}_k - 1) - km \le m - a < a\bar{l}_k - km,$$

which leads to

$$\bar{l}_k = \left[\frac{m}{a}(k+1)\right].$$

If we are in the case (15) with $Z(l) \le 0$ but $Z(l-1) > 0$ with $a(l-1) \bmod m > m - \frac{a}{2}$ (that is case (18)), we determine $\bar{l}_{k-1}$ according to the above formula and add 1. The condition $\bar{l}_k \le \frac{m}{2}$ leads to $k \le \frac{a}{2}$.

This completes the proof. □

In view of Theorem 2 we have to search the absolute minimum and the absolute maximum of the function F only at the points $\underline{l}_k$ and $\bar{l}_k$, or $\bar{l}_k + 1$, respectively. The number of these points is about $\frac{3}{2} \cdot a$. Let

$$S_0 := F(\underline{l}_0) - F(1) = \sum_{l=1}^{\underline{l}_0 - 1} Z(l)$$

and

$$S_k := F(\underline{l}_k) - F(\bar{l}_{k-1} + 1) = \sum_{l=\bar{l}_{k-1}+1}^{\underline{l}_k - 1} Z(l), \qquad k = 1, \dots, \underline{k},$$

the sum of all negative increases between a candidate for a local maximum and the next candidate for a local minimum, and let

$$T_k := F(\bar{l}_k) - F(\underline{l}_k) = \sum_{l=\underline{l}_k}^{\bar{l}_k - 1} Z(l), \qquad k = 0, \dots, \bar{k},$$

the sum of all positive increases between a candidate for a local minimum and the next candidate for a local maximum. Using (12), we get by easy calculations

$$S_0 = (\underline{l}_0 - 1)(a(\underline{l}_0 + 1) - m), \tag{19}$$

$$S_k = (\underline{l}_k - \bar{l}_{k-1} - 1)(a(\underline{l}_k + \bar{l}_{k-1} + 1) - m(2k+1)) \quad \text{for } k = 1, \dots, \underline{k}, \tag{20}$$

$$T_k = (\bar{l}_k - \underline{l}_k)(a(\bar{l}_k + \underline{l}_k) - m(2k+1)) \quad \text{for } k = 0, \dots, \bar{k} \tag{21}$$

and

$$Z(\bar{l}_k) = a(2\bar{l}_k + 1) - 2m(k+1). \tag{22}$$

Using these formulas, we get an algorithm which is much faster than the previous algorithm if a is small relative to m (see also Remark 1).

ALGORITHM 2.

1. Input modulus m and multiplier a.
2. $F_{\max} \leftarrow a$, $F_{\min} \leftarrow a$, $F \leftarrow a$, $\underline{k} \leftarrow [\frac{a-1}{2}]$, $\bar{k} \leftarrow [\frac{a}{2}]$.
3. For $k = 0$ to $\bar{k}$
   calculate $\underline{l}_k$, $\bar{l}_k$ according to the formulas (13), (14)
   calculate S_k, T_k, $Z(\bar{l}_k)$ according to the formulas (19)–(22)
   $F \leftarrow F + S_k$; if $F < F_{\min}$ then $F_{\min} \leftarrow F$
   $F \leftarrow F + T_k$; if $F > F_{\max}$ then $F_{\max} \leftarrow F$
   $F \leftarrow F + Z(\bar{l}_k)$; if $F > F_{\max}$ then $F_{\max} \leftarrow F$.
4. Output discrepancy $\hat{D}_m^{(2)} \leftarrow (2m + F_{\max} - F_{\min})/m^2$.

Remark 1. It is well known that the two generators $x_i \equiv a \cdot x_{i-1} + b \pmod m$ and $x_i \equiv a' \cdot x_{i-1} + b \pmod m$ with $a \cdot a' \equiv 1 \pmod m$ have the same lattice structure (similarly for $a \cdot a'' \equiv -1 \pmod m$). The "inverse multiplier" a' can be calculated quickly (for example by the Euclidean Algorithm). If we choose the smallest value of a, a' (or a'') instead of a, the second algorithm will be faster than the first in almost all cases.

4. Exact Determination of $D_m^{(2)}$. As it can be seen below, the lattice discrepancy $\hat{D}_m^{(2)}$ calculated in Algorithm 2 (or Algorithm 1) is equal to $D_m^{(2)}$ in almost all cases. But in order to include cases as presented in Example 1, we now want to modify the algorithms presented above.

First we have to record in the algorithm all values l_1 and r_1 for which the function F has the maximal and minimal value, respectively. Based on the relationship between rectangles of $\mathscr{A}$ and $\mathscr{C}$ (see proof of Theorem 1), and using the symmetry of the functions F and G (see Lemma 1), we get the following rectangles for each pair of l_1 and r_1:

$$R_1 = \begin{cases} [-l_1, r_1] \times [0, l_2 + r_2] & \text{if } l_2 + r_2 < m, \\ (-r_1, l_1) \times (0, 2m - (l_2 + r_2)) & \text{if } l_2 + r_2 > m \end{cases} \tag{23}$$

and

$$R_2 = \begin{cases} [l_1 - m, r_1] \times [0, m - l_2 + r_2] & \text{if } l_1 > r_1 \text{ and } l_2 > r_2, \\ [-l_1, m - r_1] \times [0, m + l_2 - r_2] & \text{if } l_1 < r_1 \text{ and } l_2 < r_2, \\ (-r_1, m - l_1) \times (0, m + l_2 - r_2) & \text{if } l_1 > r_1 \text{ and } l_2 < r_2, \\ (r_1 - m, l_1) \times (0, m - l_2 + r_2) & \text{if } l_1 < r_1 \text{ and } l_2 > r_2, \end{cases} \tag{24}$$

which lead to the maximal deviation $m^2 \hat{D}_m^{(2)}$. The values l_2 and r_2 are determined by l_1 and r_1 according to (4). Now we have to check whether one of these rectangles can be moved into $[0, m)^2$ in such a way that the lattice point $\binom{0}{0}$ at the bottom of the rectangle fits in a lattice point of the original lattice generated by $x_i \equiv a x_{i-1} + b \pmod m$. The rectangles R_1 and R_2 can be moved in that way if and only if there exists a lattice point of the original lattice in the "control rectangles"

$$C_1 = \begin{cases} [l_1, m - r_1) \times [0, m - (l_2 + r_2)) & \text{if } l_2 + r_2 < m, \\ [r_1, m - l_1] \times [0, (l_2 + r_2) - m] & \text{if } l_2 + r_2 > m \end{cases} \tag{25}$$

and

$$C_2 = \begin{cases} [m - l_1, m - r_1) \times [0, l_2 - r_2) & \text{if } l_1 > r_1 \text{ and } l_2 > r_2, \\ [l_1, r_1) \times [0, r_2 - l_2) & \text{if } l_1 < r_1 \text{ and } l_2 < r_2, \\ [r_1, l_1] \times [0, r_2 - l_2] & \text{if } l_1 > r_1 \text{ and } l_2 < r_2, \\ [m - r_1, m - l_1] \times [0, l_2 - r_2] & \text{if } l_1 < r_1 \text{ and } l_2 > r_2, \end{cases} \tag{26}$$

respectively. Therefore, we define the function $check(\cdot, \cdot)$ by

$$check(l_1, r_1) := \begin{cases} true & \text{if there is a lattice point in } C_1 \text{ or } C_2, \\ false & \text{if there is no lattice point in } C_1 \text{ and } C_2. \end{cases} \tag{27}$$

If $check(l_1, r_1) = true$ for at least one pair l_1, r_1, then the descrepancy $D_m^{(2)}$ is equal to $\hat{D}_m^{(2)}$ as calculated by the algorithms of Section 3. Otherwise, we have to proceed as in the following general algorithm.

ALGORITHM 3.

1. Input modulus m and multiplier a.
2. Calculate $\hat{D}_m^{(2)}$ according to Algorithm 2 (or Algorithm 1) and record all values $l_{\max}$ and $r_{\min}$ for which F is maximal and minimal, respectively, $\overline{\max} \leftarrow F_{\max}$, $\underline{\min} \leftarrow F_{\min}$, $Dis \leftarrow \hat{D}_m^{(2)}$, $Dis_0 \leftarrow \frac{2}{m}$.

3. Calculate $check(l_{\max}, r_{\min})$ for each pair $l_{\max}, r_{\min}$ by (27);
   if $check(l_{\max}, r_{\min}) = true$ for at least one pair, then goto 7.
4. $F_{\max} \leftarrow \underline{\min}, F_{\min} \leftarrow \overline{\max}$
   for $l = 1$ to $[\frac{m}{2}]$
   calculate $F(l)$ by $F(1) = a$ and $F(l) \leftarrow F(l-1) + Z(l-1)$ with
   $Z(l-1)$ calculated by formula (12);
   if $F(l) > F_{\max}$ and $F(l) < \overline{\max}$
   then calculate $check(l, r_{\min})$ for all $r_{\min}$,
   if $check(l, r_{\min}) = true$ for at least one $r_{\min}$ then $F_{\max} \leftarrow F(l)$
   if $F(l) < F_{\min}$ and $F(l) > \underline{\min}$
   then calculate $check(l_{\max}, l)$ for all $l_{\max}$,
   if $check(l_{\max}, l) = true$ for at least one $l_{\max}$ then $F_{\min} \leftarrow F(l)$
5. If $F_{\max} > \underline{\min}$ then $Dis_1 \leftarrow (2m + F_{\max} - \underline{\min})/m^2$ else $Dis_1 \leftarrow 0$
   if $F_{\min} < \overline{\max}$ then $Dis_2 \leftarrow (2m + \overline{\max} - F_{\min})/m^2$ else $Dis_2 \leftarrow 0$
   $Dis_0 \leftarrow \max(Dis_0, Dis_1, Dis_2)$.
6. For $l = 1, \ldots, [\frac{m}{2}]$ calculate $F(l)$, determine all values $l_{\max}$ and $r_{\min}$ with
   $F(l_{\max}) = \max_{1 \le l \le [\frac{m}{2}]}\{F(l)\colon F(l) < \overline{\max}\}$,
   $F(r_{\min}) = \min_{1 \le l \le [\frac{m}{2}]}\{F(l)\colon F(l) > \underline{\min}\}$,
   $\overline{\max} \leftarrow F(l_{\max}), \underline{\min} \leftarrow F(r_{\min}), Dis \leftarrow (2m + \overline{\max} - \underline{\min})/m^2$,
   if $Dis \le Dis_0$ then $Dis = Dis_0$, goto 7, else goto 3.
7. Output discrepancy $D_m^{(2)} \leftarrow Dis$

Numerical tests show that for almost all increments b the discrepancy $D_m^{(2)}$ is already obtained by Algorithm 2 (and Algorithm 1).

Example 2. For $m = 2^e$, $3 \le e \le 12$, we examine all generators (2) with period length m. The discrepancy $D_m^{(2)}$ differs from the lattice discrepancy $\hat{D}_m^{(2)}$ only in the cases $a = \frac{m}{2} + 1$ and $b = \frac{m}{4} + 1$ or $b = \frac{m}{4} - 1$ and in addition for odd e in the cases $a = \frac{2m-1}{3}$ and $b = \frac{1}{3}(\frac{m}{2} - 1)$ or $b = \frac{m}{2} - 1$ as well as $a = m - 3$ and $b = \frac{m}{2} - 3$ or $b = \frac{m}{2} - 1$. This is a very small number of exceptions, and the discrepancy is relatively large in these cases so that the exceptions are not interesting in applications.

Example 3. We examine the generator (2) with modulus $m = 2^{32}$ and Marsaglia's multiplier $a = 69069$ and increment $b = 1$ (see [9]). The estimation $mD_m^{(2)} > 15545$ given in [11] was sharpened to $mD_m^{(2)} > 15546.9$ in [1] by simple considerations of the special lattice structure of the generator. Using Algorithm 3, we get the exact value

$$mD_m^{(2)} = 66800785799847 \cdot 2^{-32} = 15553.26995$$

(usually $mD_m^{(2)}$ is examined instead of $D_m^{(2)}$). On a SIEMENS MX-2 computer (32-bit CPU) the algorithm required about 110 CPU-sec. The check of other increments $b = 3, 5, \ldots, 69069$ takes additional 80 CPU-sec. In all these cases we have $D_m^{(2)} = \hat{D}_m^{(2)}$.

Remark 2. The algorithms can be used also to calculate the discrepancy of multiplicative congruential generators ($b = 0$ in (2)). For example, in the case of modulus 2^e we have to use $m = 2^{e-2}$ in the algorithms and the control rectangles have to be modified. Similarly, the discrepancy of a sublattice can be determined.

Remark 3. The considerations for the calculation of the two-dimensional rectangle discrepancy presented here can be extended to the cases of higher dimensions. But in these cases there are great difficulties to give a simple form of the corresponding deviation function D so that effective algorithms cannot easily be developed.

Acknowledgment. The authors would like to thank the referees for their careful reading of the manuscript and for valuable suggestions.

Technische Hochschule Darmstadt
D-6100 Darmstadt
Federal Republic of Germany
E-mail: xmatdb3y@ddathd21.bitnet

1. L. AFFLERBACH, *Lineare Kongruenz-Generatoren zur Erzeugung von Pseudo-Zufallszahlen und ihre Gitterstruktur*, Dissertation, Technische Hochschule Darmstadt, 1983.

2. W. A. BEYER, "Lattice structure and reduced bases of random vectors generated by linear recurrences," in *Applications of Number Theory to Numerical Analysis* (S. K. Zaremba, ed.), Academic Press, New York, 1972, pp. 361–370.

3. R. R. COVEYOU & R. D. MACPHERSON, "Fourier analysis of uniform random number generators," *J. Assoc. Comput. Mach.*, v. 14, 1967, pp. 100–119.

→ U. DIETER, "Pseudo-random numbers: The exact distribution of pairs," *Math. Comp.*, v. 25, 1971, pp. 855–883.

5. U. DIETER & J. H. AHRENS, *Uniform Random Numbers*, Inst. f. Math. Stat., Technische Hochschule Graz, 1974.

→ E. HLAWKA, "Zur angenäherten Berechnung mehrfacher Integrale," *Monatsh. Math.*, v. 66, 1962, pp. 140–151.

7. D. E. KNUTH, *The Art of Computer Programming*, Vol. II, 2nd ed., Addison-Wesley, Reading, Mass., 1981.

→ G. MARSAGLIA, "Random numbers fall mainly in the planes," *Proc. Nat. Acad. Sci. U.S.A.*, v. 61, 1968, pp. 25–28.

9. G. MARSAGLIA, "The structure of linear congruential sequences," in *Applications of Number Theory to Numerical Analysis* (S. K. Zaremba, ed.), Academic Press, New York, 1972, pp. 249–285.

10. H. NIEDERREITER, "On the distribution of pseudo-random numbers generated by the linear congruential method. III," *Math. Comp.*, v. 30, 1976, pp. 571–597.

→ H. NIEDERREITER, "Quasi-Monte Carlo methods and pseudo-random numbers," *Bull. Amer. Math. Soc.*, v. 84, 1978, pp. 957–1041.

→ B. D. RIPLEY, "The lattice structure of pseudo-random number generators," *Proc. Roy. Soc. London Ser. A*, v. 389, 1983, pp. 197–204.