D. E. KNUTH THE ART OF COMPUTER PROGRAMMING V.I 2
ADDISON-WESLEY, 1998

that abstracts this situation: Let $U_{n+1} = \{\alpha + 2U_n - U_{n-1}\}$. As in exercise 26, divide the unit square into parts that show the relative order of U_1, U_2, and U_3 for each pair (U_1, U_2). Are there any values of α for which all six possible orders are achieved with probability $\frac{1}{6}$, assuming that U_1 and U_2 are chosen at random in the unit square?

3.3.4. The Spectral Test

In this section we shall study an especially important way to check the quality of linear congruential random number generators. Not only do all good generators pass this test, all generators now known to be bad actually *fail* it. Thus it is by far the most powerful test known, and it deserves particular attention. Our discussion will also bring out some fundamental limitations on the degree of randomness that we can expect from linear congruential sequences and their generalizations.

The spectral test embodies aspects of both the empirical and theoretical tests studied in previous sections: It is like the theoretical tests because it deals with properties of the full period of the sequence, and it is like the empirical tests because it requires a computer program to determine the results.

A. Ideas underlying the test. The most important randomness criteria seem to rely on properties of the joint distribution of t consecutive elements of the sequence, and the spectral test deals directly with this distribution. If we have a sequence $\langle U_n \rangle$ of period m, the basic idea is to analyze the set of all m points

$$\{\, (U_n, U_{n+1}, \ldots, U_{n+t-1}) \mid 0 \le n < m \,\} \tag{1}$$

in t-dimensional space.

For simplicity we shall assume that we have a linear congruential sequence (X_0, a, c, m) of maximum period length m (so that $c \neq 0$), or that m is prime and $c = 0$ and the period length is $m - 1$. In the latter case we shall add the point $(0, 0, \ldots, 0)$ to the set (1), so that there are always m points in all; this extra point has a negligible effect when m is large, and it makes the theory much simpler. Under these assumptions, (1) can be rewritten as

$$\left\{ \frac{1}{m} \big(x, s(x), s(s(x)), \ldots, s^{[t-1]}(x)\big) \;\middle|\; 0 \le x < m \right\}, \tag{2}$$

where

$$s(x) = (ax + c) \bmod m \tag{3}$$

is the successor of x. We are considering only the set of all such points in t dimensions, not the order in which those points are actually generated. But the order of generation is reflected in the dependence between components of the vectors; and the spectral test studies such dependence for various dimensions t by dealing with the totality of all points (2).

For example, Fig. 8 shows a typical small case in 2 and 3 dimensions, for the generator with

$$s(x) = (137x + 187) \bmod 256. \tag{4}$$

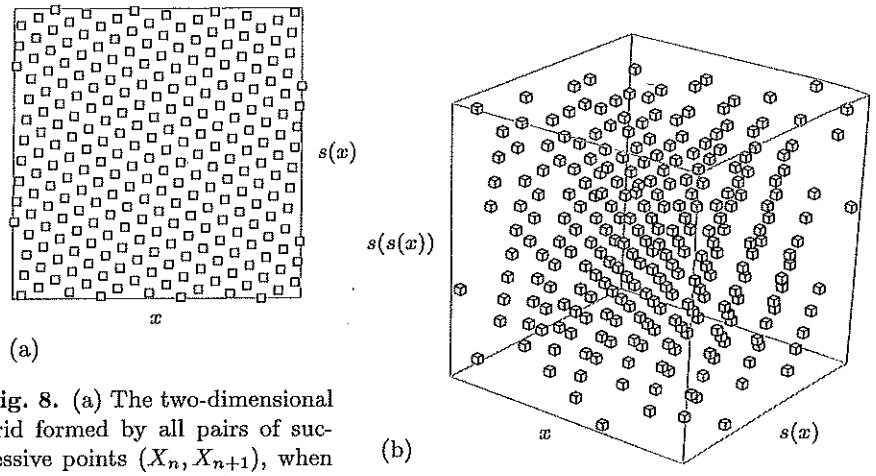


Fig. 8. (a) The two-dimensional grid formed by all pairs of successive points (X_n, X_{n+1}), when $X_{n+1} = (137X_n + 187) \bmod 256$. (b) The three-dimensional grid of triplets (X_n, X_{n+1}, X_{n+2}).

Of course a generator with period length 256 will hardly be random, but 256 is small enough that we can draw the diagram and gain some understanding before we turn to the larger m's that are of practical interest.

Perhaps the most striking thing about the pattern of boxes in Fig. 8(a) is that we can cover them all by a fairly small number of parallel lines; indeed, there are many different families of parallel lines that will hit all the points. For example, a set of 20 nearly vertical lines will do the job, as will a set of 21 lines that tilt upward at roughly a 30° angle. We commonly observe similar patterns when driving past farmlands that have been planted in a systematic manner.

If the same generator is considered in three dimensions, we obtain 256 points in a cube, obtained by appending a "height" component $s(s(x))$ to each of the 256 points $\big(x, s(x)\big)$ in the plane of Fig. 8(a), as shown in Fig. 8(b). Let's imagine that this 3-D crystal structure has been made into a physical model, a cube that we can turn in our hands; as we rotate it, we will notice various families of parallel planes that encompass all of the points. In the words of Wallace Givens, the random numbers stay "mainly in the planes."

At first glance we might think that such systematic behavior is so nonrandom as to make congruential generators quite worthless; but more careful reflection, remembering that m is quite large in practice, provides a better insight. The regular structure in Fig. 8 is essentially the "grain" we see when examining our random numbers under a high-power microscope. If we take truly random numbers between 0 and 1, and round or truncate them to finite accuracy so that each is an integer multiple of $1/\nu$ for some given number ν, then the t-dimensional points (1) we obtain will have an extremely regular character when viewed through a microscope.

Let $1/\nu_2$ be the maximum distance between lines, taken over all families of parallel straight lines that cover the points $\big\{\big(x/m,\ s(x)/m\big)\big\}$ in two dimen-

sions. We shall call ν_2 the two-dimensional *accuracy* of the random number generator, since the pairs of successive numbers have a fine structure that is essentially good to one part in ν_2. Similarly, let $1/\nu_3$ be the maximum distance between planes, taken over all families of parallel planes that cover all points $\{(x/m,\, s(x)/m,\, s(s(x))/m)\}$; we shall call ν_3 the accuracy in three dimensions. The t-dimensional accuracy ν_t is the reciprocal of the maximum distance between hyperplanes, taken over all families of parallel $(t-1)$-dimensional hyperplanes that cover all points $\{(x/m,\, s(x)/m,\, \ldots,\, s^{[t-1]}(x)/m)\}$.

The essential difference between periodic sequences and truly random sequences that have been truncated to multiples of $1/\nu$ is that the accuracy of truly random sequences is the same in all dimensions, while that of periodic sequences decreases as t increases. Indeed, since there are only m points in the t-dimensional cube when m is the period length, we can't achieve a t-dimensional accuracy of more than about $m^{1/t}$.

When the independence of t consecutive values is considered, computer-generated random numbers will behave essentially as if we took truly random numbers and truncated them to $\lg \nu_t$ bits, where ν_t decreases with increasing t. In practice, such varying accuracy is usually all we need. We don't insist that the 10-dimensional accuracy be 2^{32}, in the sense that all $(2^{32})^{10}$ possible 10-tuples $(U_n, U_{n+1}, \ldots, U_{n+9})$ should be equally likely on a 32-bit machine; for such large values of t we want only a few of the leading bits of $(U_n, U_{n+1}, \ldots, U_{n+t-1})$ to behave as if they were independently random.

On the other hand when an application demands high resolution of the random number sequence, simple linear congruential sequences will necessarily be inadequate. A generator with longer period should be used instead, even though only a small fraction of the period will actually be generated. Squaring the period length will essentially square the accuracy in higher dimensions; that is, it will double the effective number of bits of precision.

The spectral test is based on the values of ν_t for small t, say $2 \le t \le 6$. Dimensions 2, 3, and 4 seem to be adequate to detect important deficiencies in a sequence, but since we are considering the entire period it is wise to be somewhat cautious and go up into another dimension or two; on the other hand the values of ν_t for $t \ge 10$ seem to be of no practical significance whatever. (This is fortunate, because it appears to be rather difficult to calculate the accuracy ν_t precisely when $t \ge 10$.)

There is a vague relation between the spectral test and the serial test; for example, a special case of the serial test, taken over the entire period as in exercise 3.3.3–19, counts the number of boxes in each of 64 subsquares of Fig. 8(a). The main difference is that the spectral test rotates the dots so as to discover the least favorable orientation. We shall return to the serial test later in this section.

It may appear at first that we should apply the spectral test only for one suitably high value of t; if a generator passes the test in three dimensions, it seems plausible that it should also pass the 2-D test, hence we might as well omit the latter. The fallacy in this reasoning occurs because we apply more stringent conditions in lower dimensions. A similar situation occurs with the serial test:

Consider a generator that (quite properly) has almost the same number of points in each subcube of the unit cube, when the unit cube has been divided into 64 subcubes of size $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$; this same generator might yield completely *empty* subsquares of the unit square, when the unit square has been divided into 64 subsquares of size $\frac{1}{8} \times \frac{1}{8}$. Since we increase our expectations in lower dimensions, a separate test for each dimension is required.

It is not always true that $\nu_t \le m^{1/t}$, although this upper bound is valid when the points form a rectangular grid. For example, it turns out that $\nu_2 = \sqrt{274} > \sqrt{256}$ in Fig. 8, because a nearly hexagonal structure brings the m points closer together than would be possible in a strictly rectangular arrangement.

In order to develop an algorithm that computes ν_t efficiently, we must look more deeply at the associated mathematical theory. Therefore a reader who is not mathematically inclined is advised to skip to part D of this section, where the spectral test is presented as a "plug-in" method accompanied by several examples. But the mathematics behind the spectral test requires only some elementary manipulations of vectors.

Some authors have suggested using the minimum number N_t of parallel covering lines or hyperplanes as the criterion, instead of the maximum distance $1/\nu_t$ between them. However, this number N_t does not appear to be as important as the concept of accuracy defined above, because it is biased by how nearly the slope of the lines or hyperplanes matches the coordinate axes of the cube. For example, the 20 nearly vertical lines that cover all the points of Fig. 8(a) are actually $1/\sqrt{328}$ units apart, according to Eq. (14) below with $(u_1, u_2) = (18, -2)$; this might falsely imply an accuracy of one part in $\sqrt{328}$, or perhaps even an accuracy of one part in 20. The true accuracy of only one part in $\sqrt{274}$ is realized only for the larger family of 21 lines with a slope of $7/15$; another family of 24 lines, with a slope of $-11/13$, also has a greater inter-line distance than the 20-line family, since $1/\sqrt{290} > 1/\sqrt{328}$. The precise way in which families of lines act at the boundaries of the unit hypercube does not seem to be an especially "clean" or significant criterion. However, for those people who prefer to count hyperplanes, it is possible to compute N_t using a method quite similar to the way in which we shall calculate ν_t (see exercise 16).

***B. Theory behind the test.** In order to analyze the basic set (2), we start with the observation that

$$\frac{1}{m} s^{[j]}(x) = \left(\frac{a^j x + (1 + a + \cdots + a^{j-1})c}{m} \right) \bmod 1. \tag{5}$$

We can get rid of the "mod 1" operation by extending the set periodically, making infinitely many copies of the original t-dimensional hypercube, proceeding in all directions. This gives us the set

$$\begin{aligned} L &= \left\{ \left(\frac{x}{m} + k_1, \frac{s(x)}{m} + k_2, \ldots, \frac{s^{[t-1]}(x)}{m} + k_t \right) \;\middle|\; \text{integer } x, k_1, k_2, \ldots, k_t \right\} \\ &= \left\{ V_0 + \left(\frac{x}{m} + k_1, \frac{ax}{m} + k_2, \ldots, \frac{a^{t-1}x}{m} + k_t \right) \;\middle|\; \text{integer } x, k_1, k_2, \ldots, k_t \right\}, \end{aligned}$$

where

$$V_0 = \frac{1}{m}\big(0,\, c,\, (1+a)c,\, \ldots,\, (1+a+\cdots+a^{t-2})c\big) \tag{6}$$

is a constant vector. The variable k_1 is redundant in this representation of L, because we can change $(x, k_1, k_2, \ldots, k_t)$ to $(x+k_1m,\, 0,\, k_2-ak_1,\, \ldots,\, k_t-a^{t-1}k_1)$, reducing k_1 to zero without loss of generality. Therefore we obtain the comparatively simple formula

$$L = \{V_0 + y_1V_1 + y_2V_2 + \cdots + y_tV_t \mid \text{integer } y_1, y_2, \ldots, y_t\}, \tag{7}$$

where

$$V_1 = \frac{1}{m}(1, a, a^2, \ldots, a^{t-1}); \tag{8}$$

$$V_2 = (0,1,0,\ldots,0), \quad V_3 = (0,0,1,\ldots,0), \quad \ldots, \quad V_t = (0,0,0,\ldots,1). \tag{9}$$

The points $(x_1, x_2, \ldots, x_t)$ of L that satisfy $0 \le x_j < 1$ for all j are precisely the m points of our original set (2).

Notice that the increment c appears only in V_0, and the effect of V_0 is merely to shift all elements of L without changing their relative distances; hence *c does not affect the spectral test in any way*, and we might as well assume that $V_0 = (0,0,\ldots,0)$ when we are calculating ν_t. When V_0 is the zero vector we have a *lattice* of points

$$L_0 = \{y_1V_1 + y_2V_2 + \cdots + y_tV_t \mid \text{integer } y_1, y_2, \ldots, y_t\}, \tag{10}$$

and our goal is to study the distances between adjacent $(t-1)$-dimensional hyperplanes, in families of parallel hyperplanes that cover all the points of L_0.

A family of parallel $(t-1)$-dimensional hyperplanes can be defined by a nonzero vector $U = (u_1, \ldots, u_t)$ that is perpendicular to all of them; and the set of points on a particular hyperplane is then

$$\{(x_1, \ldots, x_t) \mid x_1u_1 + \cdots + x_tu_t = q\}, \tag{11}$$

where q is a different constant for each hyperplane in the family. In other words, each hyperplane is the set of all vectors X for which the *dot product* $X \cdot U$ has a given value q. In our case the hyperplanes are all separated by a fixed distance, and one of them contains $(0,0,\ldots,0)$; hence we can adjust the magnitude of U so that the set of all *integer* values q gives all the hyperplanes in the family. Then the distance between neighboring hyperplanes is the minimum distance from $(0,0,\ldots,0)$ to the hyperplane for $q = 1$, namely

$$\min_{\text{real } x_1,\ldots,x_t} \left\{ \sqrt{x_1^2 + \cdots + x_t^2} \;\middle|\; x_1u_1 + \cdots + x_tu_t = 1 \right\}. \tag{12}$$

Cauchy's inequality (see exercise 1.2.3–30) tells us that

$$(x_1u_1 + \cdots + x_tu_t)^2 \le (x_1^2 + \cdots + x_t^2)\,(u_1^2 + \cdots + u_t^2), \tag{13}$$

hence the minimum in (12) occurs when each $x_j = u_j/(u_1^2+\cdots+u_t^2)$; the distance between neighboring hyperplanes is

$$1/\sqrt{u_1^2+\cdots+u_t^2} = 1/\text{length}(U). \tag{14}$$

In other words, the quantity ν_t that we seek is precisely the length of the shortest vector U that defines a family of hyperplanes $\{X\cdot U = q \mid \text{integer } q\}$ containing all the elements of L_0.

Such a vector $U = (u_1,\ldots,u_t)$ must be nonzero, and it must satisfy $V\cdot U =$ integer for all V in L_0. In particular, since the points $(1,0,\ldots,0)$, $(0,1,\ldots,0)$, $\ldots$, $(0,0,\ldots,1)$ are all in L_0, all of the u_j must be integers. Furthermore since V_1 is in L_0, we must have $\frac{1}{m}(u_1+au_2+\cdots+a^{t-1}u_t) =$ integer, i.e.,

$$u_1+au_2+\cdots+a^{t-1}u_t \equiv 0 \pmod m. \tag{15}$$

Conversely, any nonzero integer vector $U=(u_1,\ldots,u_t)$ satisfying (15) defines a family of hyperplanes with the required properties, since all of L_0 will be covered. The dot product $(y_1V_1+\cdots+y_tV_t)\cdot U$ will be an integer for all integers $y_1,\ldots,y_t$. We have proved that

$$\begin{aligned}\nu_t^2 &= \min_{(u_1,\ldots,u_t)\neq(0,\ldots,0)}\{u_1^2+\cdots+u_t^2 \mid u_1+au_2+\cdots+a^{t-1}u_t\equiv 0 \pmod m\}\\ &= \min_{(x_1,\ldots,x_t)\neq(0,\ldots,0)}\big((mx_1-ax_2-a^2x_3-\cdots-a^{t-1}x_t)^2+x_2^2+x_3^2+\cdots+x_t^2\big).\end{aligned} \tag{16}$$

C. Deriving a computational method. We have now reduced the spectral test to the problem of finding the minimum value (16); but how on earth can we determine that minimum value in a reasonable amount of time? A brute-force search is out of the question, since m is very large in cases of practical interest.

It will be interesting and probably more useful if we develop a computational method for solving an even more general problem: *Find the minimum value of the quantity*

$$f(x_1,\ldots,x_t) = (u_{11}x_1+\cdots+u_{t1}x_t)^2+\cdots+(u_{1t}x_1+\cdots+u_{tt}x_t)^2 \tag{17}$$

over all nonzero integer vectors $(x_1,\ldots,x_t)$, given any nonsingular matrix of coefficients $U=(u_{ij})$. The expression (17) is called a "positive definite quadratic form" in t variables. Since U is nonsingular, (17) cannot be zero unless the x_j are all zero.

Let us write $U_1,\ldots,U_t$ for the rows of U. Then (17) may be written

$$f(x_1,\ldots,x_t) = (x_1U_1+\cdots+x_tU_t)\cdot(x_1U_1+\cdots+x_tU_t), \tag{18}$$

the square of the length of the vector $x_1U_1+\cdots+x_tU_t$. The nonsingular matrix U has an inverse, which means that we can find uniquely determined vectors $V_1,\ldots,V_t$ such that

$$U_i\cdot V_j = \delta_{ij}, \qquad 1\le i,j\le t. \tag{19}$$

For example, in the special form (16) that arises in the spectral test, we have

$$\begin{aligned} U_1 &= (\ \ m, 0, 0, \ldots, 0), & V_1 &= \tfrac{1}{m}(1, a, a^2, \ldots, a^{t-1}), \\ U_2 &= (\ -a, 1, 0, \ldots, 0), & V_2 &= \ (0, 1, \ 0, \ldots, \ \ 0), \\ U_3 &= (\ -a^2, 0, 1, \ldots, 0), & V_3 &= \ (0, 0, \ 1, \ldots, \ \ 0), \\ & \ldots\ldots\ldots\ldots & & \ldots \\ U_t &= (-a^{t-1}, 0, 0, \ldots, 1), & V_t &= \ (0, 0, \ 0, \ldots, \ \ 1). \end{aligned} \tag{20}$$

These V_j are precisely the vectors (8), (9) that we used to define our original lattice L_0. As the reader may well suspect, this is not a coincidence—indeed, if we had begun with an arbitrary lattice L_0, defined by any set of linearly independent vectors $V_1, \ldots, V_t$, the argument we have used above can be generalized to show that the maximum separation between hyperplanes in a covering family is equivalent to minimizing (17), where the coefficients u_{ij} are defined by (19). (See exercise 2.)

Our first step in minimizing (18) is to reduce it to a finite problem, namely to show that we won't need to test infinitely many vectors $(x_1, \ldots, x_t)$ when finding the minimum. This is where the vectors $V_1, \ldots, V_t$ come in handy; we have

$$x_k = (x_1 U_1 + \cdots + x_t U_t) \cdot V_k,$$

and Cauchy's inequality tells us that

$$\big((x_1 U_1 + \cdots + x_t U_t) \cdot V_k\big)^2 \le f(x_1, \ldots, x_t)(V_k \cdot V_k).$$

Hence we have derived a useful upper bound on each coordinate x_k:

Lemma A. *Let $(x_1, \ldots, x_t)$ be a nonzero vector that minimizes* (18) *and let $(y_1, \ldots, y_t)$ be any nonzero integer vector. Then*

$$x_k^2 \le f(y_1, \ldots, y_t)(V_k \cdot V_k), \qquad \text{for } 1 \le k \le t. \tag{21}$$

In particular, letting $y_i = \delta_{ij}$ for all i,

$$x_k^2 \le (U_j \cdot U_j)(V_k \cdot V_k), \qquad \text{for } 1 \le j, k \le t. \ \blacksquare \tag{22}$$

Lemma A reduces the problem to a finite search, but the right-hand side of (21) is usually much too large to make an exhaustive search feasible; we need at least one more idea. On such occasions, an old maxim provides sound advice: "If you can't solve a problem as it is stated, change it into a simpler problem that has the same answer." For example, Euclid's algorithm has this form; if we don't know the gcd of the input numbers, we change them into smaller numbers having the same gcd. (In fact, a slightly more general approach probably underlies the discovery of nearly all algorithms: "If you can't solve a problem directly, change it into one or more simpler problems, from whose solution you can solve the original one.")

In our case, a simpler problem is one that requires less searching because the right-hand side of (22) is smaller. The key idea we shall use is that it is possible to change one quadratic form into another one that is equivalent for all practical

purposes. Let j be any fixed subscript, $1 \le j \le t$; let $(q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_t)$ be any sequence of $t-1$ integers; and consider the following transformation of the vectors:

$$\begin{aligned} V_i' &= V_i - q_i V_j, \qquad & x_i' &= x_i - q_i x_j, \qquad & U_i' &= U_i, \qquad \text{for } i \ne j; \\ V_j' &= V_j, & x_j' &= x_j, & U_j' &= U_j + \textstyle\sum_{i \ne j} q_i U_i. \end{aligned} \tag{23}$$

It is easy to see that the new vectors $U_1', \ldots, U_t'$ define a quadratic form f' for which $f'(x_1', \ldots, x_t') = f(x_1, \ldots, x_t)$; furthermore the basic orthogonality condition (19) remains valid, because it is easy to check that $U_i' \cdot V_j' = \delta_{ij}$. As $(x_1, \ldots, x_t)$ runs through all nonzero integer vectors, so does $(x_1', \ldots, x_t')$; hence the new form f' has the same minimum as f.

Our goal is to use transformation (23), replacing U_i by U_i' and V_i by V_i' for all i, in order to make the right-hand side of (22) small; and the right-hand side of (22) will be small when both $U_j \cdot U_j$ and $V_k \cdot V_k$ are small. Therefore it is natural to ask the following two questions about the transformation (23):

a) *What choice of q_i makes $V_i' \cdot V_i'$ as small as possible?*

b) *What choice of $q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_t$ makes $U_j' \cdot U_j'$ as small as possible?*

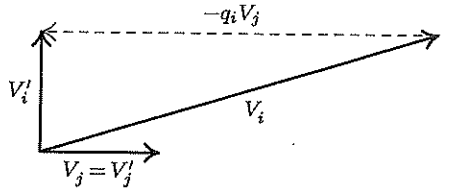
It is easiest to solve these questions first for *real* values of the q_i. Question (a) is quite simple, since

$$\begin{aligned} (V_i - q_i V_j) \cdot (V_i - q_i V_j) &= V_i \cdot V_i - 2q_i\, V_i \cdot V_j + q_i^2\, V_j \cdot V_j \\ &= (V_j \cdot V_j)\left(q_i - (V_i \cdot V_j / V_j \cdot V_j)\right)^2 + V_i \cdot V_i - (V_i \cdot V_j)^2 / V_j \cdot V_j, \end{aligned}$$

and the minimum occurs when

$$q_i = V_i \cdot V_j \,/\, V_j \cdot V_j. \tag{24}$$

Geometrically, we are asking what multiple of V_j should be subtracted from V_i so that the resulting vector V_i' has minimum length, and the answer is to choose q_i so that V_i' is perpendicular to V_j (that is, to make $V_i' \cdot V_j = 0$); the following diagram makes this plain.



(25)

Turning to question (b), we want to choose the q_i so that $U_j + \sum_{i \ne j} q_i U_i$ has minimum length; geometrically, we want to start with U_j and add some vector in the $(t-1)$-dimensional hyperplane whose points are the sums of multiples of $\{U_i \mid i \ne j\}$. Again the best solution is to choose things so that U_j' is perpendicular to the hyperplane, making $U_j' \cdot U_k = 0$ for all $k \ne j$:

$$U_j \cdot U_k + \sum_{i \ne j} q_i (U_i \cdot U_k) = 0, \qquad 1 \le k \le t, \qquad k \ne j. \tag{26}$$

(See exercise 12 for a rigorous proof that a solution to question (b) must satisfy these $t-1$ equations.)

Now that we have answered questions (a) and (b), we are in a bit of a quandary; should we choose the q_i according to (24), so that the $V_i' \cdot V_i'$ are minimized, or according to (26), so that $U_j' \cdot U_j'$ is minimized? Either of these alternatives makes an improvement in the right-hand side of (22), so it is not immediately clear which choice should get priority. Fortunately, there is a very simple answer to this dilemma: Conditions (24) and (26) are exactly the same! (See exercise 7.) Therefore questions (a) and (b) have the same answer; we have a happy state of affairs in which we can reduce the length of both the U's and the V's simultaneously. Indeed, we have just rediscovered the *Gram–Schmidt orthogonalization process* [see *Crelle* **94** (1883), 41–73].

Our joy must be tempered with the realization that we have dealt with questions (a) and (b) only for *real* values of the q_i. Our application restricts us to integer values, so we cannot make V_i' exactly perpendicular to V_j. The best we can do for question (a) is to let q_i be the *nearest integer* to $V_i \cdot V_j \,/\, V_j \cdot V_j$ (see (25)). It turns out that this is *not* always the best solution to question (b); in fact U_j' may at times be longer than U_j. However, the bound (21) is never increased, since we can remember the smallest value of $f(y_1, \ldots, y_t)$ found so far. Thus a choice of q_i based solely on question (a) is quite satisfactory.

If we apply transformation (23) repeatedly in such a way that none of the vectors V_i gets longer and at least one gets shorter, we can never get into a loop; that is, we will never be considering the same quadratic form again after a sequence of nontrivial transformations of this kind. But eventually we will get stuck, in the sense that none of the transformations (23) for $1 \le j \le t$ will be able to shorten any of the vectors $V_1, \ldots, V_t$. At that point we can revert to an exhaustive search, using the bounds of Lemma A, which will now be quite small in most cases. Occasionally these bounds (21) will be poor, and another type of transformation will usually get the algorithm unstuck again and reduce the bounds (see exercise 18). However, transformation (23) by itself has proved to be quite adequate for the spectral test; in fact, it has proved to be amazingly powerful when the computations are arranged as in the algorithm discussed below.

***D. How to perform the spectral test.** Here now is an efficient computational procedure that follows from our considerations. R. W. Gosper and U. Dieter have observed that it is possible to use the results of lower dimensions to make the spectral test significantly faster in higher dimensions. This refinement has been incorporated into the following algorithm, together with Gauss's significant simplification of the two-dimensional case (exercise 5).

Algorithm S (*The spectral test*). This algorithm determines the value of

$$\nu_t = \min\left\{ \sqrt{x_1^2 + \cdots + x_t^2} \;\middle|\; x_1 + ax_2 + \cdots + a^{t-1}x_t \equiv 0 \pmod{m} \right\} \tag{27}$$

for $2 \le t \le T$, given a, m, and T, where $0 < a < m$ and a is relatively prime to m. (The minimum is taken over all nonzero integer vectors $(x_1, \ldots, x_t)$, and the

number ν_t measures the t-dimensional accuracy of random number generators as discussed in the text above.) All arithmetic within this algorithm is done on integers whose magnitudes rarely if ever exceed m^2, except in step S7; in fact nearly all of the integer variables will be less than m in absolute value during the computation.

When ν_t is being calculated for $t \geq 3$, the algorithm works with two $t \times t$ matrices U and V, whose row vectors are denoted by $U_i = (u_{i1}, \ldots, u_{it})$ and $V_i = (v_{i1}, \ldots, v_{it})$ for $1 \leq i \leq t$. These vectors satisfy the conditions

$$u_{i1} + au_{i2} + \cdots + a^{t-1}u_{it} \equiv 0 \text{ (modulo } m), \qquad 1 \leq i \leq t; \tag{28}$$

$$U_i \cdot V_j = m\delta_{ij}, \qquad 1 \leq i, j \leq t. \tag{29}$$

(Thus the V_j of our previous discussion have been multiplied by m, to ensure that their components are integers.) There are three other auxiliary vectors, $X = (x_1, \ldots, x_t)$, $Y = (y_1, \ldots, y_t)$, and $Z = (z_1, \ldots, z_t)$. During the entire algorithm, r will denote $a^{t-1} \bmod m$ and s will denote the smallest upper bound for ν_t^2 that has been discovered so far.

S1. [Initialize.] Set $t \leftarrow 2$, $h \leftarrow a$, $h' \leftarrow m$, $p \leftarrow 1$, $p' \leftarrow 0$, $r \leftarrow a$, $s \leftarrow 1 + a^2$. (The first steps of this algorithm handle the case $t = 2$ by a special method very much like Euclid's algorithm; we will have

$$h - ap \equiv h' - ap' \equiv 0 \text{ (modulo } m) \qquad \text{and} \qquad hp' - h'p = \pm m \tag{30}$$

during this phase of the calculation.)

S2. [Euclidean step.] Set $q \leftarrow \lfloor h'/h \rfloor$, $u \leftarrow h' - qh$, $v \leftarrow p' - qp$. If $u^2 + v^2 < s$, set $s \leftarrow u^2 + v^2$, $h' \leftarrow h$, $h \leftarrow u$, $p' \leftarrow p$, $p \leftarrow v$, and repeat step S2.

S3. [Compute ν_2.] Set $u \leftarrow u - h$, $v \leftarrow v - p$; and if $u^2 + v^2 < s$, set $s \leftarrow u^2 + v^2$, $h' \leftarrow u$, $p' \leftarrow v$. Then output $\sqrt{s} = \nu_2$. (The validity of this calculation for the two-dimensional case is proved in exercise 5. Now we will set up the U and V matrices satisfying (28) and (29), in preparation for calculations in higher dimensions.) Set

$$U \leftarrow \begin{pmatrix} -h & p \\ -h' & p' \end{pmatrix}, \qquad V \leftarrow \pm \begin{pmatrix} p' & h' \\ -p & -h \end{pmatrix},$$

where the $-$ sign is chosen for V if and only if $p' > 0$.

S4. [Advance t.] If $t = T$, the algorithm terminates. (Otherwise we want to increase t by 1. At this point U and V are $t \times t$ matrices satisfying (28) and (29), and we must enlarge them by adding an appropriate new row and column.) Set $t \leftarrow t + 1$ and $r \leftarrow (ar) \bmod m$. Set U_t to the new row $(-r, 0, 0, \ldots, 0, 1)$ of t elements, and set $u_{it} \leftarrow 0$ for $1 \leq i < t$. Set V_t to the new row $(0, 0, 0, \ldots, 0, m)$. Finally, for $1 \leq i < t$, set $q \leftarrow \text{round}(v_{i1}r/m)$, $v_{it} \leftarrow v_{i1}r - qm$, and $U_t \leftarrow U_t + qU_i$. (Here "round($x$)" denotes the nearest integer to x, e.g., $\lfloor x + 1/2 \rfloor$. We are essentially setting $v_{it} \leftarrow v_{i1}r$ and immediately applying transformation (23) with $j = t$, since the numbers $|v_{i1}r|$ are so large they ought to be reduced at once.) Finally set $s \leftarrow \min(s, U_t \cdot U_t)$, $k \leftarrow t$, and $j \leftarrow 1$. (In the following steps, j denotes the

current row index for transformation (23), and k denotes the last such index where the transformation shortened at least one of the V_i.)

S5. [Transform.] For $1 \le i \le t$, do the following operations: If $i \ne j$ and $2|V_i \cdot V_j| > V_j \cdot V_j$, set $q \leftarrow \text{round}(V_i \cdot V_j \,/\, V_j \cdot V_j)$, $V_i \leftarrow V_i - qV_j$, $U_j \leftarrow U_j + qU_i$, $s \leftarrow \min(s, U_j \cdot U_j)$, and $k \leftarrow j$. (We omit the transformation when $2|V_i \cdot V_j|$ exactly equals $V_j \cdot V_j$; exercise 19 shows that this precaution keeps the algorithm from looping endlessly.)

S6. [Advance j.] If $j = t$, set $j \leftarrow 1$; otherwise set $j \leftarrow j + 1$. Now if $j \ne k$, return to step S5. (If $j = k$, we have gone through $t - 1$ consecutive cycles of no transformation, so the transformation process is stuck.)

S7. [Prepare for search.] (Now the absolute minimum will be determined, using an exhaustive search over all $(x_1, \ldots, x_t)$ satisfying condition (21) of Lemma A.) Set $X \leftarrow Y \leftarrow (0, \ldots, 0)$, set $k \leftarrow t$, and set

$$z_j \leftarrow \left\lfloor \sqrt{\lfloor (V_j \cdot V_j)s/m^2 \rfloor} \right\rfloor, \qquad \text{for } 1 \le j \le t. \tag{31}$$

(We will examine all $X = (x_1, \ldots, x_t)$ with $|x_j| \le z_j$ for $1 \le j \le t$. Usually $|z_j| \le 1$, but L. C. Killingbeck noticed in 1999 that larger values occur for about 0.00001 of all multipliers when $m = 2^{64}$. During the exhaustive search, the vector Y will always be equal to $x_1U_1 + \cdots + x_tU_t$, so that $f(x_1, \ldots, x_t) = Y \cdot Y$. Since $f(-x_1, \ldots, -x_t) = f(x_1, \ldots, x_t)$, we shall examine only vectors whose first nonzero component is positive. The method is essentially that of counting in steps of one, regarding $(x_1, \ldots, x_t)$ as the digits in a balanced number system with mixed radices $(2z_1+1, \ldots, 2z_t+1)$; see Section 4.1.)

S8. [Advance x_k.] If $x_k = z_k$, go to S10. Otherwise increase x_k by 1 and set $Y \leftarrow Y + U_k$.

S9. [Advance k.] Set $k \leftarrow k+1$. Then if $k \le t$, set $x_k \leftarrow -z_k$, $Y \leftarrow Y - 2z_kU_k$, and repeat step S9. But if $k > t$, set $s \leftarrow \min(s, Y \cdot Y)$.

S10. [Decrease k.] Set $k \leftarrow k - 1$. If $k \ge 1$, return to S8. Otherwise output $\nu_t = \sqrt{s}$ (the exhaustive search is completed) and return to S4. ∎

In practice Algorithm S is applied for $T = 5$ or 6, say; it usually works reasonably well when $T = 7$ or 8, but it can be terribly slow when $T \ge 9$ since the exhaustive search tends to make the running time grow as 3^T. (If the minimum value ν_t occurs at many different points, the exhaustive search will hit them all; hence we typically find that all $z_k = 1$ for large t. As remarked above, the values of ν_t are generally irrelevant for practical purposes when t is large.)

An example will help to make Algorithm S clear. Consider the linear congruential sequence defined by

$$m = 10^{10}, \qquad a = 3141592621, \qquad c = 1, \qquad X_0 = 0. \tag{32}$$

Six cycles of the Euclidean algorithm in steps S2 and S3 suffice to prove that the minimum nonzero value of $x_1^2 + x_2^2$ with

$$x_1 + 3141592621x_2 \equiv 0 \pmod{10^{10}}$$

occurs for $x_1 = 67654$, $x_2 = 226$; hence the two-dimensional accuracy of this generator is

$$\nu_2 = \sqrt{67654^2 + 226^2} \approx 67654.37748.$$

Passing to three dimensions, we seek the minimum nonzero value of $x_1^2 + x_2^2 + x_3^2$ such that

$$x_1 + 3141592621x_2 + 3141592621^2 x_3 \equiv 0 \pmod{10^{10}}. \tag{33}$$

Step S4 sets up the matrices

$$U = \begin{pmatrix} -67654 & -226 & 0 \\ -44190611 & 191 & 0 \\ 5793866 & 33 & 1 \end{pmatrix}, \quad V = \begin{pmatrix} -191 & -44190611 & 2564918569 \\ -226 & 67654 & 1307181134 \\ 0 & 0 & 10000000000 \end{pmatrix}.$$

The first iteration of step S5, with $q = 1$ for $i = 2$ and $q = 4$ for $i = 3$, changes them to

$$U = \begin{pmatrix} -21082801 & 97 & 4 \\ -44190611 & 191 & 0 \\ 5793866 & 33 & 1 \end{pmatrix}, \quad V = \begin{pmatrix} -191 & -44190611 & 2564918569 \\ -35 & 44258265 & -1257737435 \\ 764 & 176762444 & -259674276 \end{pmatrix}.$$

(The first row U_1 has actually gotten longer in this transformation, although eventually the rows of U should get shorter.)

The next fourteen iterations of step S5 have $(j, q_1, q_2, q_3) = (2, -2, *, 0)$, $(3, 0, 3, *)$, $(1, *, -10, -1)$, $(2, -1, *, -6)$, $(3, -1, 0, *)$, $(1, *, 0, 2)$, $(2, 0, *, -1)$, $(3, 3, 4, *)$, $(1, *, 0, 0)$, $(2, -5, *, 0)$, $(3, 1, 0, *)$, $(1, *, -3, -1)$, $(2, 0, *, 0)$, $(3, 0, 0, *)$. Now the transformation process is stuck, but the rows of the matrices have become significantly shorter:

$$U = \begin{pmatrix} -1479 & 616 & -2777 \\ -3022 & 104 & 918 \\ -227 & -983 & -130 \end{pmatrix}, \quad V = \begin{pmatrix} -888874 & 601246 & -2994234 \\ -2809871 & 438109 & 1593689 \\ -854296 & -9749816 & -1707736 \end{pmatrix}. \tag{34}$$

The search limits (z_1, z_2, z_3) in step S7 turn out to be $(0, 0, 1)$, so U_3 is the shortest solution to (33); we have

$$\nu_3 = \sqrt{227^2 + 983^2 + 130^2} \approx 1017.21089.$$

Only a few iterations were needed to find this value, although condition (33) looks quite formidable at first glance. Our computation has proved that all points (U_n, U_{n+1}, U_{n+2}) produced by the random number generator (32) lie on a family of parallel planes about 0.001 units apart, but not on any family of planes that differ by more than 0.001 units.

The exhaustive search in steps S8–S10 reduces the value of s only rarely. One such case, found in 1982 by R. Carling and K. Levine, occurs when $a = 464680339$, $m = 2^{29}$, and $t = 5$; another case arose when the author calculated ν_6^2 for line 21 of Table 1, later in this section.

E. Ratings for various generators. So far we haven't really given a criterion that tells us whether or not a particular random number generator passes or flunks the spectral test. In fact, spectral success depends on the application, since some applications demand higher resolution than others. It appears that

$\nu_t \geq 2^{30/t}$ for $2 \leq t \leq 6$ will be quite adequate for most purposes (although the author must admit choosing this criterion partly because 30 is conveniently divisible by 2, 3, 5, and 6).

For some purposes we would like a criterion that is relatively independent of m, so we can say that a particular multiplier is good or bad with respect to the set of all other multipliers for the given m, without examining any others. A reasonable figure of merit for rating the goodness of a particular multiplier seems to be the volume of the ellipsoid in t-space defined by the relation

$$(x_1 m - x_2 a - \cdots - x_t a^{t-1})^2 + x_2^2 + \cdots + x_t^2 \leq \nu_t^2,$$

since this volume tends to indicate how likely it is that nonzero *integer* points $(x_1, \ldots, x_t)$ — corresponding to solutions of (15) — are in the ellipsoid. We therefore propose to calculate this volume, namely

$$\mu_t = \frac{\pi^{t/2}\, \nu_t^t}{(t/2)!\, m}, \tag{35}$$

as an indication of the effectiveness of the multiplier a for the given m. In this formula,

$$\left(\frac{t}{2}\right)! = \left(\frac{t}{2}\right)\left(\frac{t}{2}-1\right)\cdots\left(\frac{1}{2}\right)\sqrt{\pi}, \qquad \text{for } t \text{ odd}. \tag{36}$$

Thus, in six or fewer dimensions the merit is computed as follows:

$$\mu_2 = \pi\nu_2^2/m, \qquad \mu_3 = \tfrac{4}{3}\pi\nu_3^3/m, \qquad \mu_4 = \tfrac{1}{2}\pi^2\nu_4^4/m,$$
$$\mu_5 = \tfrac{8}{15}\pi^2\nu_5^5/m, \qquad \mu_6 = \tfrac{1}{6}\pi^3\nu_6^6/m.$$

We might say that the multiplier a passes the spectral test if μ_t is 0.1 or more for $2 \leq t \leq 6$, and it "passes with flying colors" if $\mu_t \geq 1$ for all these t. A low value of μ_t means that we have probably picked a very unfortunate multiplier, since very few lattices will have integer points so close to the origin. Conversely, a high value of μ_t means that we have found an unusually good multiplier for the given m; but it does not mean that the random numbers are necessarily very good, since m might be too small. Only the values ν_t truly indicate the degree of randomness.

Table 1 shows what sorts of values occur in typical sequences. Each line of the table considers a particular generator, and lists ν_t^2, μ_t, and the "number of bits of accuracy" $\lg \nu_t$. Lines 1 through 4 show the generators that were the subject of Figs. 2 and 5 in Section 3.3.1. The generators in lines 1 and 2 suffer from too small a multiplier; a diagram like Fig. 8 will have a nearly vertical "stripes" when a is small. The terrible generator in line 3 has a good μ_2 but very poor μ_3 and μ_4; like nearly all generators of potency 2, it has $\nu_3 = \sqrt{6}$ and $\nu_4 = 2$ (see exercise 3). Line 4 shows a "random" multiplier; this generator has satisfactorily passed numerous empirical tests for randomness, but it does not have especially high values of $\mu_2, \ldots, \mu_6$. In fact, the value of μ_5 flunks our criterion.

Line 5 shows the generator of Fig. 8. It passes the spectral test with very high-flying colors, when μ_2 through μ_6 are considered, but of course m is so small that the numbers can hardly be called random; the ν_t values are terribly low.

Table 1

SAMPLE RESULTS OF THE SPECTRAL TEST

Line	a	m	ν_2^2	ν_3^2	ν_4^2	ν_5^2	ν_6^2
1	23	10^8+1	530	530	530	530	447
2	2^7+1	2^{35}	16642	16642	16642	15602	252
3	$2^{18}+1$	2^{35}	34359738368	6	4	4	4
4	3141592653	2^{35}	2997222016	1026050	27822	1118	1118
5	137	256	274	30	14	6	4
6	3141592621	10^{10}	4577114792	1034718	62454	1776	542
7	3141592221	10^{10}	4293881050	276266	97450	3366	2382
8	4219755981	10^{10}	10721093248	2595578	49362	5868	820
9	4160984121	10^{10}	9183801602	4615650	16686	6840	1344
10	$2^{24}+2^{13}+5$	2^{35}	8364058	8364058	21476	16712	1496
11	5^{13}	2^{35}	33161885770	2925242	113374	13070	2256
12	$2^{16}+3$	2^{29}	536936458	118	116	116	116
13	1812433253	2^{32}	4326934538	1462856	15082	4866	906
14	1566083941	2^{32}	4659748970	2079590	44902	4652	662
15	69069	2^{32}	4243209856	2072544	52804	6990	242
16	2650845021	2^{32}	4938969760	2646962	68342	8778	1506
17	314159269	$2^{31}-1$	1432232969	899290	36985	3427	1144
18	62089911	$2^{31}-1$	1977289717	1662317	48191	6101	1462
19	16807	$2^{31}-1$	282475250	408197	21682	4439	895
20	48271	$2^{31}-1$	1990735345	1433881	47418	4404	1402
21	40692	$2^{31}-249$	1655838865	1403422	42475	6507	1438
22	44485709377909	2^{46}	5.6×10^{13}	1180915002	1882426	279928	26230
23	31167285	2^{48}	3.2×10^{14}	4111841446	17341510	306326	59278
24	see (38)		2.4×10^{18}	4.7×10^{11}	1.9×10^{9}	3194548	1611610
25	see (39)		$(2^{31}-1)^2$	1.4×10^{12}	643578623	12930027	837632
26	see the text	2^{64}	8.8×10^{18}	6.4×10^{12}	4.1×10^{9}	45662836	1846368
27	see the text	$\approx 2^{78}$	$2^{62}+1$	4281084902	2.2×10^{9}	1.8×10^{9}	1862407
28	$2^{-24.389}$	$\approx 2^{576}$	1.8×10^{173}	3.5×10^{115}	4.4×10^{86}	2×10^{69}	5×10^{57}
29	$(2^{32}-5)^{-400}$	$\approx 2^{1376}$	1.6×10^{414}	8.6×10^{275}	1×10^{207}	2×10^{165}	8×10^{137}

Line 6 is the generator discussed in (32) above. Line 7 is a similar example, having an abnormally low value of μ_3. Line 8 shows a nonrandom multiplier for the same modulus m; all of its partial quotients are 1, 2, or 3. Such multipliers have been suggested by I. Borosh and H. Niederreiter because the Dedekind sums are likely to be especially small and because they produce best results in the two-dimensional serial test (see Section 3.3.3 and exercise 30). The particular example in line 8 has only one '3' as a partial quotient; there is no multiplier congruent to 1 modulo 20 whose partial quotients with respect to 10^{10} are only 1s and 2s. The generator in line 9 shows another multiplier chosen with malice aforethought, following a suggestion by A. G. Waterman that guarantees a reasonably high value of μ_2 (see exercise 11). Line 10 is interesting because it has high μ_3 in spite of very low μ_2 (see exercise 8).

Line 11 of Table 1 is a reminder of the good old days—it once was used extensively, following a suggestion of O. Taussky in the early 1950s. But computers for which 2^{35} was an appropriate modulus began to fade in importance during the late 60s, and they disappeared almost completely in the 80s, as machines

($\epsilon = \frac{1}{10}$)

$\lg \nu_2$	$\lg \nu_3$	$\lg \nu_4$	$\lg \nu_5$	$\lg \nu_6$	μ_2	μ_3	μ_4	μ_5	μ_6	Line
4.5	4.5	4.5	4.5	4.4	$2\epsilon^5$	$5\epsilon^4$	0.01	0.34	4.62	1
7.0	7.0	7.0	7.0	4.0	$2\epsilon^6$	$3\epsilon^4$	0.04	4.66	$2\epsilon^3$	2
17.5	1.3	1.0	1.0	1.0	3.14	$2\epsilon^9$	$2\epsilon^9$	$5\epsilon^9$	ϵ^8	3
15.7	10.0	7.4	5.1	5.1	0.27	0.13	0.11	0.01	0.21	4
4.0	2.5	1.9	1.3	1.0	3.36	2.69	3.78	1.81	1.29	5
16.0	10.0	8.0	5.4	4.5	1.44	0.44	1.92	0.07	0.08	6
16.0	9.0	8.3	5.9	5.6	1.35	0.06	4.69	0.35	6.98	7
16.7	10.7	7.8	6.3	4.8	3.37	1.75	1.20	1.39	0.28	8
16.5	11.1	7.0	6.4	5.2	2.89	4.15	0.14	2.04	1.25	9
11.5	11.5	7.2	7.0	5.3	$8\epsilon^4$	2.95	0.07	5.53	0.50	10
17.5	10.7	8.4	6.8	5.6	3.03	0.61	1.85	2.99	1.73	11
14.5	3.4	3.4	3.4	3.4	3.14	ϵ^5	ϵ^4	ϵ^3	0.02	12
16.0	10.2	6.9	6.1	4.9	3.16	1.73	0.26	2.02	0.89	13
16.1	10.5	7.7	6.1	4.7	3.41	2.92	2.32	1.81	0.35	14
16.0	10.5	7.8	6.4	4.0	3.10	2.91	3.20	5.01	0.02	15
16.1	10.7	8.0	6.6	5.3	3.61	4.20	5.37	8.85	4.11	16
15.2	9.9	7.6	5.9	5.1	2.10	1.66	3.14	1.69	3.60	17
15.4	10.3	7.8	6.3	5.3	2.89	4.18	5.34	7.13	7.52	18
14.0	9.3	7.2	6.1	4.9	0.41	0.51	1.08	3.22	1.73	19
15.4	10.2	7.8	6.1	5.2	2.91	3.35	5.17	3.15	6.63	20
15.3	10.2	7.7	6.3	5.2	2.42	3.24	4.15	8.37	7.16	21
22.8	15.1	10.4	9.0	7.3	2.48	2.42	0.25	3.10	1.33	22
24.1	16.0	12.0	9.1	7.9	3.60	3.92	5.27	0.97	3.82	23
30.5	19.4	15.4	10.8	10.3	1.65	0.29	3.88	0.02	4.69	24
31.0	20.2	14.6	11.8	9.8	3.14	1.49	0.44	0.69	0.66	25
31.5	21.3	16.0	12.7	10.4	1.50	3.68	4.52	4.02	1.76	26
31.0	16.0	15.5	15.4	10.4	$5\epsilon^5$	$4\epsilon^9$	$8\epsilon^5$	2.56	ϵ^4	27
288.	192.	144.	115.	95.9	2.27	3.46	3.92	2.49	2.98	28
688.	458.	344.	275.	229.	3.10	2.04	2.85	1.15	1.33	29
upper bounds from (40):					3.63	5.92	9.87	14.89	23.87	

with 32-bit arithmetic began to proliferate. This switch to a comparatively small word size called for comparatively greater care. Line 12 was, alas, the generator actually used on such machines in most of the world's scientific computing centers for more than a decade; its very name RANDU is enough to bring dismay into the eyes and stomachs of many computer scientists! The actual generator is defined by

$$X_0 \text{ odd}, \qquad X_{n+1} = (65539X_n) \bmod 2^{31}, \tag{37}$$

and exercise 20 indicates that 2^{29} is the appropriate modulus for the spectral test. Since $9X_n - 6X_{n+1} + X_{n+2} \equiv 0$ (modulo 2^{31}), the generator fails most three-dimensional criteria for randomness, and it should never have been used. Almost any multiplier $\equiv 5$ (modulo 8) would be better. (A curious fact about RANDU, noticed by R. W. Gosper, is that $\nu_4 = \nu_5 = \nu_6 = \nu_7 = \nu_8 = \nu_9 = \sqrt{116}$, hence μ_9 is a spectacular 11.98.) Lines 13 and 14 are the Borosh–Niederreiter and Waterman multipliers for modulus 2^{32}. Line 16 was found by L. C. Killingbeck, who carried out an exhaustive search of all multipliers $a \equiv 1 \bmod 4$ when $m = 2^{32}$. Line 23, similarly, was found by M. Lavaux and F. Janssens in a

(nonexhaustive) computer search for spectrally good multipliers having a very high μ_2. Line 22 is for the multiplier used with $c = 0$ and $m = 2^{48}$ in the Cray X-MP library; line 26 (whose excellent multiplier 6364136223846793005 is too big to fit in the column) is due to C. E. Haynes. Line 15 was nominated by George Marsaglia as "a candidate for the best of all multipliers," after a computer search for nearly cubical lattices in dimensions 2 through 5, partly because it is easy to remember [*Applications of Number Theory to Numerical Analysis*, edited by S. K. Zaremba (New York: Academic Press, 1972), 275].

Line 17 uses a random primitive root, modulo the prime $2^{31}-1$, as multiplier. Line 18 shows the spectrally best primitive root for $2^{31}-1$, found in an exhaustive search by G. S. Fishman and L. R. Moore III [*SIAM J. Sci. Stat. Comput.* 7 (1986), 24–45]. The adequate but less outstanding multiplier $16807 = 7^5$ in line 19 is actually used most often for that modulus, after being proposed by Lewis, Goodman, and Miller in *IBM Systems J.* **8** (1969), 136–146; it has been one of the main generators in the popular IMSL subroutine library since 1971. The main reason for continued use of $a = 16807$ is that a^2 is less than the modulus m, hence $ax \bmod m$ can be implemented with reasonable efficiency in high-level languages using the technique of exercise 3.2.1.1–9. However, such small multipliers have known defects. S. K. Park and K. W. Miller noticed that the same implementation technique applies also to certain multipliers greater than $\sqrt{m}$, so they asked G. S. Fishman to find the best "efficiently portable" multiplier in this wider class; the result appears in line 20 [*CACM* **31** (1988), 1192–1201]. Line 21 shows another good multiplier, due to P. L'Ecuyer [*CACM* **31** (1988), 742–749, 774]; this one uses a slightly smaller prime modulus.

When the generators of lines 20 and 21 are combined by subtraction as suggested in Eq. 3.2.2–(15), so that the generated numbers $\langle Z_n \rangle$ satisfy

$$X_{n+1} = 48271X_n \bmod (2^{31}-1), \qquad Y_{n+1} = 40692Y_n \bmod (2^{31}-249), \qquad Z_n = (X_n - Y_n) \bmod (2^{31}-1), \tag{38}$$

exercise 32 shows that it is reasonable to rate $\langle Z_n \rangle$ with the spectral test for $m = (2^{31}-1)(2^{31}-249)$ and $a = 1431853894371298687$. (This value of a satisfies $a \bmod (2^{31}-1) = 48271$ and $a \bmod (2^{31}-249) = 40692$.) The results appear on line 24. We needn't worry too much about the low value of μ_5, since $\nu_5 > 1000$. Generator (38) has a period of length $(2^{31}-2)(2^{31}-250)/62 \approx 7 \times 10^{16}$.

Line 25 of the table represents the sequence

$$X_n = (271828183X_{n-1} - 314159269X_{n-2}) \bmod (2^{31}-1), \tag{39}$$

which can be shown to have period length $(2^{31}-1)^2 - 1$; it has been analyzed with the generalized spectral test of exercise 24.

The last three lines of Table 1 are based on add-with-carry and subtract-with-borrow methods, which simulate linear congruential sequences that have extremely large moduli (see exercise 3.2.1.1–14). Line 27 is for the generator

$$X_n = (X_{n-1} + 65430X_{n-2} + C_n) \bmod 2^{31},$$
$$C_{n+1} = \lfloor (X_{n-1} + 65430X_{n-2} + C_n)/2^{31} \rfloor,$$

which corresponds to $\mathcal{X}_{n+1} = (65430 \cdot 2^{31} + 1)\mathcal{X}_n \bmod (65430 \cdot 2^{62} + 2^{31} - 1)$; the numbers in the table refer to the "super-values"

$$\mathcal{X}_n = (65430 \cdot 2^{31} + 1)X_{n-1} + 65430X_{n-2} + C_n$$

rather than to the values X_n actually computed and used as random numbers. Line 28 represents a more typical subtract-with-borrow generator

$$X_n = (X_{n-10} - X_{n-24} - C_n) \bmod 2^{24}, \quad C_{n+1} = [X_{n-10} < X_{n-24} + C_n],$$

but modified by generating 389 elements of the sequence and then using only the first (or last) 24. This generator, called RANLUX, was recommended by Martin Lüscher after it passed many stringent tests that previous generators failed [*Computer Physics Communications* **79** (1994), 100–110]. A similar sequence,

$$X_n = (X_{n-22} - X_{n-43} - C_n) \bmod (2^{32} - 5), \quad C_{n+1} = [X_{n-22} < X_{n-43} + C_n],$$

with 43 elements used after 400 are generated, appears in line 29; this sequence is discussed in the answer to exercise 3.2.1.2–22. In both cases the table entries refer to the spectral test on multiprecision numbers $\mathcal{X}_n$ instead of to the individual "digits" X_n, but the high μ values indicate that the process of generating 389 or 400 numbers before selecting 24 or 43 is an excellent way to remove biases due to the extreme simplicity of the generation scheme.

Theoretical upper bounds on μ_t, which can never be transcended for any m, are shown just below Table 1; it is known that every lattice with m points per unit volume has

$$\nu_t \le \gamma_t^{1/2} m^{1/t}, \tag{40}$$

where γ_t takes the respective values

$$(4/3)^{1/2}, \quad 2^{1/3}, \quad 2^{1/2}, \quad 2^{3/5}, \quad (64/3)^{1/6}, \quad 4^{3/7}, \quad 2 \tag{41}$$

for $t = 2, \ldots, 8$. [See exercise 9 and J. W. S. Cassels, *Introduction to the Geometry of Numbers* (Berlin: Springer, 1959), 332; J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups* (New York: Springer, 1988), 20.] These bounds hold for lattices generated by vectors with arbitrary real coordinates. For example, the optimum lattice for $t = 2$ is hexagonal, and it is generated by vectors of length $2/\sqrt{3m}$ that form two sides of an equilateral triangle. In three dimensions the optimum lattice is generated by vectors V_1, V_2, V_3 that can be rotated into the form $(v, v, -v)$, $(v, -v, v)$, $(-v, v, v)$, where $v = 1/\sqrt[3]{4m}$.

***F. Relation to the serial test.** In a series of important papers published during the 1970s, Harald Niederreiter showed how to analyze the distribution of the t-dimensional vectors (1) by means of exponential sums. One of the main consequences of his theory is that the serial test in several dimensions will be passed by any generator that passes the spectral test, even when we consider only a sufficiently large part of the period instead of the whole period. We shall now turn briefly to a study of his interesting methods, in the case of linear congruential sequences (X_0, a, c, m) of period length m.

The first idea we need is the notion of *discrepancy* in t dimensions, a quantity that we shall define as the difference between the expected number and the actual number of t-dimensional vectors $(x_n, x_{n+1}, \ldots, x_{n+t-1})$ falling into a hyper-rectangular region, maximized over all such regions. To be precise, let $\langle x_n \rangle$ be a sequence of integers in the range $0 \le x_n < m$. We define

$$D_N^{(t)} = \max_R \left| \frac{\text{number of } (x_n, \ldots, x_{n+t-1}) \text{ in } R \text{ for } 0 \le n < N}{N} - \frac{\text{volume of } R}{m^t} \right| \tag{42}$$

where R ranges over all sets of points of the form

$$R = \{(y_1, \ldots, y_t) \mid \alpha_1 \le y_1 < \beta_1, \ldots, \alpha_t \le y_t < \beta_t\}; \tag{43}$$

here α_j and β_j are integers in the range $0 \le \alpha_j < \beta_j \le m$, for $1 \le j \le t$. The volume of R is clearly $(\beta_1 - \alpha_1) \ldots (\beta_t - \alpha_t)$. To get the discrepancy $D_N^{(t)}$, we imagine looking at all these sets R and finding the one with the greatest excess or deficiency of points $(x_n, \ldots, x_{n+t-1})$.

An upper bound for the discrepancy can be found by using exponential sums. Let $\omega = e^{2\pi i/m}$ be a primitive mth root of unity. If $(x_1, \ldots, x_t)$ and $(y_1, \ldots, y_t)$ are two vectors with all components in the range $0 \le x_j, y_j < m$, we have

$$\sum_{0 \le u_1, \ldots, u_t < m} \omega^{(x_1 - y_1)u_1 + \cdots + (x_t - y_t)u_t} = \begin{cases} m^t & \text{if } (x_1, \ldots, x_t) = (y_1, \ldots, y_t), \\ 0 & \text{if } (x_1, \ldots, x_t) \ne (y_1, \ldots, y_t). \end{cases}$$

Therefore the number of vectors $(x_n, \ldots, x_{n+t-1})$ in R for $0 \le n < N$, when R is defined by (43), can be expressed as

$$\frac{1}{m^t} \sum_{0 \le n < N} \sum_{0 \le u_1, \ldots, u_t < m} \omega^{x_n u_1 + \cdots + x_{n+t-1} u_t} \sum_{\alpha_1 \le y_1 < \beta_1} \cdots \sum_{\alpha_t \le y_t < \beta_t} \omega^{-(y_1 u_1 + \cdots + y_t u_t)}.$$

When $u_1 = \cdots = u_t = 0$ in this sum, we get N/m^t times the volume of R; hence we can express $D_N^{(t)}$ as the maximum over R of

$$\left| \frac{1}{Nm^t} \sum_{0 \le n < N} \sum_{\substack{0 \le u_1, \ldots, u_t < m \\ (u_1, \ldots, u_t) \ne (0, \ldots, 0)}} \omega^{x_n u_1 + \cdots + x_{n+t-1} u_t} \sum_{\alpha_1 \le y_1 < \beta_1} \cdots \sum_{\alpha_t \le y_t < \beta_t} \omega^{-(y_1 u_1 + \cdots + y_t u_t)} \right|.$$

Since complex numbers satisfy $|w + z| \le |w| + |z|$ and $|wz| = |w||z|$, it follows that

$$D_N^{(t)} \le \max_R \frac{1}{m^t} \sum_{\substack{0 \le u_1, \ldots, u_t < m \\ (u_1, \ldots, u_t) \ne (0, \ldots, 0)}} \left| \sum_{\alpha_1 \le y_1 < \beta_1} \cdots \sum_{\alpha_t \le y_t < \beta_t} \omega^{-(y_1 u_1 + \cdots + y_t u_t)} \right| g(u_1, \ldots, u_t)$$

$$\le \frac{1}{m^t} \sum_{\substack{0 \le u_1, \ldots, u_t < m \\ (u_1, \ldots, u_t) \ne (0, \ldots, 0)}} \max_R \left| \sum_{\alpha_1 \le y_1 < \beta_1} \cdots \sum_{\alpha_t \le y_t < \beta_t} \omega^{-(y_1 u_1 + \cdots + y_t u_t)} \right| g(u_1, \ldots, u_t)$$

$$= \sum_{\substack{0\le u_1,\dots,u_t<m\\(u_1,\dots,u_t)\neq(0,\dots,0)}} f(u_1,\dots,u_t)\,g(u_1,\dots,u_t), \tag{44}$$

where

$$g(u_1,\dots,u_t) = \left|\frac{1}{N}\sum_{0\le n<N}\omega^{x_nu_1+\cdots+x_{n+t-1}u_t}\right|;$$

$$f(u_1,\dots,u_t) = \max_R \frac{1}{m^t}\left|\sum_{\alpha_1\le y_1<\beta_1}\cdots\sum_{\alpha_t\le y_t<\beta_t}\omega^{-(y_1u_1+\cdots+y_tu_t)}\right|$$

$$= \max_R \left|\frac{1}{m}\sum_{\alpha_1\le y_1<\beta_1}\omega^{-u_1y_1}\right|\dots\left|\frac{1}{m}\sum_{\alpha_t\le y_t<\beta_t}\omega^{-u_ty_t}\right|.$$

Both f and g can be simplified further in order to get a good upper bound on $D_N^{(t)}$. We have

$$\left|\frac{1}{m}\sum_{\alpha\le y<\beta}\omega^{-uy}\right| = \left|\frac{1}{m}\,\frac{\omega^{-\beta u}-\omega^{-\alpha u}}{\omega^{-u}-1}\right| \le \frac{2}{m\,|\omega^u-1|} = \frac{1}{m\sin(\pi u/m)}$$

when $u\neq 0$, and the sum is ≤ 1 when $u=0$; hence

$$f(u_1,\dots,u_t) \le r(u_1,\dots,u_t), \tag{45}$$

where

$$r(u_1,\dots,u_t) = \prod_{\substack{1\le k\le t\\u_k\neq 0}}\frac{1}{m\sin(\pi u_k/m)}. \tag{46}$$

Furthermore, when $\langle x_n\rangle$ is generated modulo m by a linear congruential sequence, we have

$$\begin{aligned}x_nu_1+\cdots+x_{n+t-1}u_t &= x_nu_1+(ax_n+c)u_2+\cdots+\big(a^{t-1}x_n+c(a^{t-2}+\cdots+1)\big)u_t\\ &= (u_1+au_2+\cdots+a^{t-1}u_t)x_n+h(u_1,\dots,u_t)\end{aligned}$$

where $h(u_1,\dots,u_t)$ is independent of n; hence

$$g(u_1,\dots,u_t) = \left|\frac{1}{N}\sum_{0\le n<N}\omega^{q(u_1,\dots,u_t)x_n}\right|, \tag{47}$$

where

$$q(u_1,\dots,u_t) = u_1+au_2+\cdots+a^{t-1}u_t. \tag{48}$$

Now here is where the connection to the spectral test comes in: We will show that the sum $g(u_1,\dots,u_t)$ is rather small unless $q(u_1,\dots,u_t)\equiv 0$ (modulo m); in other words, the contributions to (44) arise mainly from the solutions to (15). Furthermore exercise 27 shows that $r(u_1,\dots,u_t)$ is rather small when $(u_1,\dots,u_t)$ is a "large" solution to (15). Hence the discrepancy $D_N^{(t)}$ will be rather small

when (15) has only "large" solutions, namely when the spectral test is passed. Our remaining task is to quantify these qualitative statements by making careful calculations.

In the first place, let's consider the size of $g(u_1,\ldots,u_t)$. When $N=m$, so that the sum (47) is over an entire period, we have $g(u_1,\ldots,u_t)=0$ except when $(u_1,\ldots,u_t)$ satisfies (15), so the discrepancy is bounded above in this case by the sum of $r(u_1,\ldots,u_t)$ taken over all the nonzero solutions of (15). But let's consider also what happens in a sum like (47) when N is less than m and $q(u_1,\ldots,u_t)$ is not a multiple of m. We have

$$\frac{1}{N}\sum_{0\le n<N}\omega^{x_n}=\frac{1}{N}\sum_{0\le n<N}\frac{1}{m}\sum_{0\le k<m}\omega^{-nk}\sum_{0\le j<m}\omega^{x_j+jk}$$
$$=\frac{1}{N}\sum_{0\le k<m}\left(\frac{1}{m}\sum_{0\le n<N}\omega^{-nk}\right)S_{k0}, \tag{49}$$

where

$$S_{kl}=\sum_{0\le j<m}\omega^{x_{j+l}+jk}. \tag{50}$$

Now $S_{kl}=\omega^{-lk}S_{k0}$, so $|S_{kl}|=|S_{k0}|$ for all l, and we can calculate this common value by further exponential-summery:

$$\begin{aligned}|S_{k0}|^2&=\frac{1}{m}\sum_{0\le l<m}|S_{kl}|^2\\&=\frac{1}{m}\sum_{0\le l<m}\sum_{0\le j<m}\omega^{x_{j+l}+jk}\sum_{0\le i<m}\omega^{-x_{i+l}-ik}\\&=\frac{1}{m}\sum_{0\le i,j<m}\omega^{(j-i)k}\sum_{0\le l<m}\omega^{x_{j+l}-x_{i+l}}\\&=\frac{1}{m}\sum_{0\le i<m}\sum_{i\le j<m+i}\omega^{(j-i)k}\sum_{0\le l<m}\omega^{(a^{j-i}-1)x_{i+l}+(a^{j-i}-1)c/(a-1)}.\end{aligned}$$

Let s be minimum such that $a^s\equiv 1$ (modulo m), and let

$$s'=(a^s-1)c/(a-1) \bmod m.$$

Then s is a divisor of m (see Lemma 3.2.1.2P), and $x_{n+js}\equiv x_n+js'$ (modulo m). The sum on l vanishes unless $j-i$ is a multiple of s, so we find that

$$|S_{k0}|^2=m\sum_{0\le j<m/s}\omega^{jsk+js'}.$$

We have $s'=q's$ where q' is relatively prime to m (see exercise 3.2.1.2–21), so it turns out that

$$|S_{k0}|=\begin{cases}0 & \text{if } k+q'\not\equiv 0 \text{ (modulo } m/s),\\ m/\sqrt{s} & \text{if } k+q'\equiv 0 \text{ (modulo } m/s).\end{cases} \tag{51}$$

3.3.4

Putting this information back into (49), and recalling the derivation of (45), shows that

$$\left|\frac{1}{N}\sum_{0\le n<N}\omega^{x_n}\right| \le \frac{m}{N\sqrt{s}}\sum_{k} r(k), \tag{52}$$

where the sum is over $0<k<m$ such that $k+q'\equiv 0$ (modulo m/s). Exercise 25 can now be used to estimate the remaining sum, and we find that

$$\left|\frac{1}{N}\sum_{0\le n<N}\omega^{x_n}\right| \le \frac{2}{\pi}\frac{\sqrt{s}}{N}\ln s + O\left(\frac{m}{N\sqrt{s}}\right). \tag{53}$$

The same bound can be used to estimate $|N^{-1}\sum_{0\le n<N}\omega^{qx_n}|$ for any $q\not\equiv 0$ (modulo m), since the effect is to replace m in this derivation by a divisor of m. In fact, the upper bound gets even smaller when q has a factor in common with m, since s and $m/\sqrt{s}$ generally become smaller. (See exercise 26.)

We have now proved that the $g(u_1,\ldots,u_t)$ part of our upper bound (44) on the discrepancy is small, if N is large enough and if $(u_1,\ldots,u_t)$ does not satisfy the spectral test congruence (15). Exercise 27 proves that the $f(u_1,\ldots,u_t)$ part of our upper bound is small, when summed over all the nonzero vectors $(u_1,\ldots,u_t)$ satisfying (15), provided that all such vectors are far away from $(0,\ldots,0)$. Putting these results together leads to the following theorem of Niederreiter:

Theorem N. *Let $\langle X_n\rangle$ be a linear congruential sequence (X_0,a,c,m) of period length m, and let s be the least positive integer such that $a^s\equiv 1$ (modulo m). Then the t-dimensional discrepancy $D_N^{(t)}$ corresponding to the first N values of $\langle X_n\rangle$, as defined in (42), satisfies*

$$D_N^{(t)} = O\left(\frac{\sqrt{s}\log s\,(\log m)^t}{N}\right) + O\left(\frac{m(\log m)^t}{N\sqrt{s}}\right) + O\big((\log m)^t\, r_{\max}\big); \tag{54}$$

$$D_m^{(t)} = O\big((\log m)^t\, r_{\max}\big). \tag{55}$$

Here $r_{\max}$ is the maximum value of the quantity $r(u_1,\ldots,u_t)$ defined in (46), taken over all nonzero integer vectors $(u_1,\ldots,u_t)$ satisfying (15).

Proof. The first two O terms in (54) come from vectors $(u_1,\ldots,u_t)$ in (44) that do not satisfy (15), since exercise 25 proves that $f(u_1,\ldots,u_t)$ summed over all $(u_1,\ldots,u_t)$ is $O\big(((2/\pi)\ln m)^t\big)$ and exercise 26 bounds each $g(u_1,\ldots,u_t)$. (These terms are missing from (55) since $g(u_1,\ldots,u_t)=0$ in that case.) The remaining O term in (54) and (55) comes from nonzero vectors $(u_1,\ldots,u_t)$ that do satisfy (15), using the bound derived in exercise 27. (By examining this proof carefully, we could replace each O in these formulas by an explicit function of t.) ∎

Eq. (55) relates to the serial test in t dimensions over the entire period, while Eq. (54) gives us useful information about the distribution of the first N generated values when N is less than m, provided that N is not too small.

Notice that (54) will guarantee low discrepancy only when s is sufficiently large; otherwise the $m/\sqrt{s}$ term will dominate. If $m = p_1^{e_1} \dots p_r^{e_r}$ and $\gcd(a-1, m) = p_1^{f_1} \dots p_r^{f_r}$, then s equals $p_1^{e_1-f_1} \dots p_r^{e_r-f_r}$ by Lemma 3.2.1.2P; thus, the largest values of s correspond to high potency. In the common case $m = 2^e$ and $a \equiv 5$ (modulo 8), we have $s = \frac{1}{4}m$, so $D_N^{(t)}$ is $O\bigl(\sqrt{m}\,(\log m)^{t+1}/N\bigr) + O\bigl((\log m)^t r_{\max}\bigr)$. It is not difficult to prove that

$$r_{\max} \le \frac{1}{\sqrt{8}\,\nu_t} \tag{56}$$

(see exercise 29). Therefore Eq. (54) says in particular that the discrepancy will be low in t dimensions if the spectral test is passed and if N is somewhat larger than $\sqrt{m}\,(\log m)^{t+1}$.

In a sense Theorem N is almost too strong, for the result in exercise 30 shows that linear congruential sequences like those in lines 8 and 13 of Table 1 have a discrepancy of order $(\log m)^2/m$ in two dimensions. The discrepancy in this case is extremely small in spite of the fact that there are parallelogram-shaped regions of area $\approx 1/\sqrt{m}$ containing no points (U_n, U_{n+1}). The fact that discrepancy can change so drastically when the points are rotated warns us that the serial test may not be as meaningful a measure of randomness as the rotation-invariant spectral test.

G. Historical remarks. In 1959, while deriving upper bounds for the error in the evaluation of t-dimensional integrals by the Monte Carlo method, N. M. Korobov devised a way to rate the multiplier of a linear congruential sequence. His rather complicated formula is related to the spectral test, since it is strongly influenced by "small" solutions to (15); but it is not quite the same. Korobov's test has been the subject of an extensive literature, surveyed by Kuipers and Niederreiter in *Uniform Distribution of Sequences* (New York: Wiley, 1974), §2.5.

The spectral test was originally formulated by R. R. Coveyou and R. D. MacPherson [*JACM* **14** (1967), 100–119], who introduced it in an interesting indirect way. Instead of working with the grid structure of successive points, they considered random number generators as sources of t-dimensional "waves." The numbers $\sqrt{x_1^2 + \cdots + x_t^2}$ such that $x_1 + \cdots + a^{t-1}x_t \equiv 0$ (modulo m) in their original treatment were the wave "frequencies," or points in the "spectrum" defined by the random number generator, with low-frequency waves being the most damaging to randomness; hence the name *spectral test.* Coveyou and MacPherson introduced a procedure analogous to Algorithm S for performing their test, based on the principle of Lemma A. However, their original procedure (which used matrices UU^T and VV^T instead of U and V) dealt with extremely large numbers; the idea of working directly with U and V was independently suggested by F. Janssens and by U. Dieter. [See *Math. Comp.* **29** (1975), 827–833.]

Several other authors pointed out that the spectral test could be understood in far more concrete terms; by introducing the study of the grid and lattice structures corresponding to linear congruential sequences, the fundamental limitations on randomness became graphically clear. See G. Marsaglia, *Proc. Nat. Acad. Sci.*

61 (1968), 25–28; W. W. Wood, *J. Chem. Phys.* **48** (1968), 427; R. R. Coveyou, *Studies in Applied Math.* **3** (Philadelphia: SIAM, 1969), 70–111; W. A. Beyer, R. B. Roof, and D. Williamson, *Math. Comp.* **25** (1971), 345–360; G. Marsaglia and W. A. Beyer, *Applications of Number Theory to Numerical Analysis*, edited by S. K. Zaremba (New York: Academic Press, 1972), 249–285, 361–370.

R. G. Stoneham showed, by using estimates of exponential sums, that $p^{1/2+\epsilon}$ or more elements of the sequence $a^k X_0 \bmod p$ have asymptotically small discrepancy, when a is a primitive root modulo the prime p [*Acta Arithmetica* **22** (1973), 371–389]. This work was extended as explained above in a number of papers by Harald Niederreiter [*Math. Comp.* **28** (1974), 1117–1132; **30** (1976), 571–597; *Advances in Math.* **26** (1977), 99–181; *Bull. Amer. Math. Soc.* **84** (1978), 957–1041]. See also Niederreiter's book *Random Number Generation and Quasi-Monte Carlo Methods* (Philadelphia: SIAM, 1992).

EXERCISES

1. [*M10*] To what does the spectral test reduce in *one* dimension? (In other words, what happens when $t = 1$?)

2. [*HM20*] Let $V_1, \ldots, V_t$ be linearly independent vectors in t-space, let L_0 be the lattice of points defined by (10), and let $U_1, \ldots, U_t$ be defined by (19). Prove that the maximum distance between $(t-1)$-dimensional hyperplanes, over all families of parallel hyperplanes that cover L_0, is $1/\min\{f(x_1, \ldots, x_t)^{1/2} \mid (x_1, \ldots, x_t) \neq (0, \ldots, 0)\}$, where f is defined in (17).

3. [*M24*] Determine ν_3 and ν_4 for all linear congruential generators of potency 2 and period length m.

▸ **4.** [*M23*] Let u_{11}, u_{12}, u_{21}, u_{22} be elements of a 2×2 integer matrix such that $u_{11} + au_{12} \equiv u_{21} + au_{22} \equiv 0$ (modulo m) and $u_{11}u_{22} - u_{21}u_{12} = m$.

a) Prove that all integer solutions (y_1, y_2) to the congruence $y_1 + ay_2 \equiv 0$ (modulo m) have the form $(y_1, y_2) = (x_1u_{11} + x_2u_{21}, x_1u_{12} + x_2u_{22})$ for integer x_1, x_2.

b) If, in addition, $2|u_{11}u_{21} + u_{12}u_{22}| \leq u_{11}^2 + u_{12}^2 \leq u_{21}^2 + u_{22}^2$, prove that $(y_1, y_2) = (u_{11}, u_{12})$ minimizes $y_1^2 + y_2^2$ over all nonzero solutions to the congruence.

5. [*M30*] Prove that steps S1 through S3 of Algorithm S correctly perform the spectral test in two dimensions. [*Hint:* See exercise 4, and prove that $(h'+h)^2 + (p'+p)^2 \geq h^2 + p^2$ at the beginning of step S2.]

6. [*M30*] Let $a_0, a_1, \ldots, a_{t-1}$ be the partial quotients of a/m as defined in Section 3.3.3, and let $A = \max_{0 \leq j < t} a_j$. Prove that $\mu_2 > 2\pi/(A + 1 + 1/A)$.

7. [*HM22*] Prove that questions (a) and (b) following Eq. (23) have the same solution for real values of $q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_t$ (see (24) and (26)).

8. [*M18*] Line 10 of Table 1 has a very low value of μ_2, yet μ_3 is quite satisfactory. What is the highest possible value of μ_3 when $\mu_2 = 10^{-6}$ and $m = 10^{10}$?

9. [*HM32*] (C. Hermite, 1846.) Let $f(x_1, \ldots, x_t)$ be a positive definite quadratic form, defined by the matrix U as in (17), and let θ be the minimum value of f at nonzero integer points. Prove that $\theta \leq (\frac{4}{3})^{(t-1)/2} |\det U|^{2/t}$. [*Hints:* If W is any integer matrix of determinant 1, the matrix WU defines a form equivalent to f; and if S is any orthogonal matrix (that is, if $S^{-1} = S^T$), the matrix US defines a form identically equal to f. Show that there is an equivalent form g whose minimum θ occurs at

$(1, 0, \ldots, 0)$. Then prove the general result by induction on t, writing $g(x_1, \ldots, x_t) = \theta(x_1 + \beta_2 x_2 + \cdots + \beta_t x_t)^2 + h(x_2, \ldots, x_t)$ where h is a positive definite quadratic form in $t-1$ variables.]

10. [*M28*] Let y_1 and y_2 be relatively prime integers such that $y_1 + a y_2 \equiv 0$ (modulo m) and $y_1^2 + y_2^2 < \sqrt{4/3}\, m$. Show that there exist integers u_1 and u_2 such that $u_1 + a u_2 \equiv 0$ (modulo m), $u_1 y_2 - u_2 y_1 = m$, $2|u_1 y_1 + u_2 y_2| \le \min(u_1^2+u_2^2, y_1^2+y_2^2)$, and $(u_1^2 + u_2^2)(y_1^2 + y_2^2) \ge m^2$. (Hence $\nu_2^2 = \min(u_1^2+u_2^2, y_1^2+y_2^2)$ by exercise 4.)

▶ **11.** [*HM30*] (Alan G. Waterman, 1974.) Invent a reasonably efficient procedure that computes multipliers $a \equiv 1$ (modulo 4) for which there exists a relatively prime solution to the congruence $y_1 + a y_2 \equiv 0$ (modulo m) with $y_1^2 + y_2^2 = \sqrt{4/3}\, m - \epsilon$, where $\epsilon > 0$ is as small as possible, given $m = 2^e$. (By exercise 10, this choice of a will guarantee that $\nu_2^2 \ge m^2/(y_1^2 + y_2^2) > \sqrt{3/4}\, m$, and there is a chance that ν_2^2 will be near its optimum value $\sqrt{4/3}\, m$. In practice we will compute several such multipliers having small ϵ, choosing the one with best spectral values ν_2, ν_3, ...)

12. [*HM23*] Prove, without geometrical handwaving, that any solution to question (b) following Eq. (23) must also satisfy the set of equations (26).

13. [*HM22*] Lemma A uses the fact that U is nonsingular to prove that a positive definite quadratic form attains a definite, nonzero minimum value at nonzero integer points. Show that this hypothesis is necessary, by exhibiting a quadratic form (19) whose matrix of coefficients is singular, and for which the values of $f(x_1, \ldots, x_t)$ get arbitrarily near zero (but never reach it) at nonzero integer points $(x_1, \ldots, x_t)$.

14. [*24*] Perform Algorithm S by hand, for $m = 100$, $a = 41$, $T = 3$.

▶ **15.** [*M20*] Let U be an integer vector satisfying (15). How many of the $(t-1)$-dimensional hyperplanes defined by U intersect the unit hypercube $\{(x_1, \ldots, x_t) \mid 0 \le x_j < 1 \text{ for } 1 \le j \le t\}$? (This is approximately the number of hyperplanes in the family that will suffice to cover L_0.)

16. [*M30*] (U. Dieter.) Show how to modify Algorithm S in order to calculate the minimum number N_t of parallel hyperplanes intersecting the unit hypercube as in exercise 15, over all U satisfying (15). [*Hint:* What are appropriate analogs to positive definite quadratic forms and to Lemma A?]

17. [*20*] Modify Algorithm S so that, in addition to computing the quantities ν_t, it outputs all integer vectors $(u_1, \ldots, u_t)$ satisfying (15) such that $u_1^2 + \cdots + u_t^2 = \nu_t^2$, for $2 \le t \le T$.

18. [*M30*] This exercise is about the worst case of Algorithm S.

a) By considering "combinatorial matrices," whose elements have the form $y + x\delta_{ij}$ (see exercise 1.2.3–39), find 3×3 matrices of integers U and V satisfying (29) such that the transformation of step S5 does nothing for any j, but the corresponding values of z_k in (31) are so huge that exhaustive search is out of the question. (The matrix U need not satisfy (28); we are interested here in *arbitrary* positive definite quadratic forms of determinant m.)

b) Although transformation (23) is of no use for the matrices constructed in (a), find another transformation that does produce a substantial reduction.

▶ **19.** [*HM25*] Suppose step S5 were changed slightly, so that a transformation with $q = 1$ would be performed when $2V_i \cdot V_j = V_j \cdot V_j$. (Thus, $q = \lfloor (V_i \cdot V_j / V_j \cdot V_j) + \frac{1}{2} \rfloor$ whenever $i \ne j$.) Would it be possible for Algorithm S to get into an infinite loop?

20. [*M23*] Discuss how to carry out an appropriate spectral test for linear congruential sequences having $c = 0$, X_0 odd, $m = 2^e$, $a \bmod 8 = 3$ or 5. (See exercise 3.2.1.2–9.)

21. [*M20*] (R. W. Gosper.) A certain application uses random numbers in batches of four, but "throws away" the second of each set. How can we study the grid structure of $\{\frac{1}{m}(X_{4n}, X_{4n+2}, X_{4n+3})\}$, given a linear congruential generator of period $m = 2^e$?

22. [*M46*] What is the best upper bound on μ_3, given that μ_2 is very near its maximum value $\sqrt{4/3}\,\pi$? What is the best upper bound on μ_2, given that μ_3 is very near its maximum value $\frac{4}{3}\pi\sqrt{2}$?

23. [*M46*] Let U_i, V_j be vectors of real numbers with $U_i \cdot V_j = \delta_{ij}$ for $1 \le i, j \le t$, and such that $U_i \cdot U_i = 1$, $2|U_i \cdot U_j| \le 1$, $2|V_i \cdot V_j| \le V_j \cdot V_j$ for $i \ne j$. How large can $V_1 \cdot V_1$ be? (This question relates to the bounds in step S7, if both (23) and the transformation of exercise 18(b) fail to make any reductions. The maximum value known to be achievable is $(t+2)/3$, which occurs when $U_1 = I_1$, $U_j = \frac{1}{2}I_1 + \frac{1}{2}\sqrt{3}\,I_j$, $V_1 = I_1 - (I_2 + \cdots + I_t)/\sqrt{3}$, $V_j = 2I_j/\sqrt{3}$, for $2 \le j \le t$, where $(I_1, \ldots, I_t)$ is the identity matrix; this construction is due to B. V. Alexeev.)

▸ **24.** [*M28*] Generalize the spectral test to second-order sequences of the form $X_n = (aX_{n-1} + bX_{n-2}) \bmod p$, having period length $p^2 - 1$. (See Eq. 3.2.2–(8).) How should Algorithm S be modified?

25. [*HM24*] Let d be a divisor of m and let $0 \le q < d$. Prove that $\sum r(k)$, summed over all $0 \le k < m$ such that $k \bmod d = q$, is at most $(2/d\pi)\ln(m/d) + O(1)$. (Here $r(k)$ is defined in Eq. (46) when $t = 1$.)

26. [*M22*] Explain why the derivation of (53) leads to a similar bound on

$$\left| N^{-1} \sum_{0 \le n < N} \omega^{qx_n} \right|$$

for $0 < q < m$. Where does the derivation of (53) break down when $m = 1$?

27. [*HM39*] (E. Hlawka, H. Niederreiter.) Let $r(u_1, \ldots, u_t)$ be the function defined in (46). Prove that $\sum r(u_1, \ldots, u_t)$, summed over all $0 \le u_1, \ldots, u_t < m$ such that $(u_1, \ldots, u_t) \ne (0, \ldots, 0)$ and (15) holds, is at most $2\big((\pi + 2\pi \lg m)^t\, r_{\max}\big)$, where $r_{\max}$ is the maximum term $r(u_1, \ldots, u_t)$ in the sum.

▸ **28.** [*M28*] (H. Niederreiter.) Find an analog of Theorem N for the case $m =$ prime, $c = 0$, $a =$ primitive root modulo m, $X_0 \not\equiv 0$ (modulo m). [*Hint:* Your exponential sums should involve $\zeta = e^{2\pi i/(m-1)}$ as well as ω.] Prove that in this case the "average" primitive root has discrepancy $D_{m-1}^{(t)} = O\big(t(\log m)^t/\varphi(m-1)\big)$, hence good primitive roots exist for all m.

29. [*HM22*] Prove that the quantity $r_{\max}$ of exercise 27 is never larger than $1/\sqrt{8}\,\nu_t$.

30. [*M33*] (S. K. Zaremba.) Prove that $r_{\max} = O(\max(a_1, \ldots, a_s)/m)$ in two dimensions, where $a_1, \ldots, a_s$ are the partial quotients obtained when Euclid's algorithm is applied to m and a. [*Hint:* We have $a/m = //a_1, \ldots, a_s//$, in the notation of Section 4.5.3; apply exercise 4.5.3–42.]

31. [*HM47*] (I. Borosh.) Prove that for all sufficiently large m there exists a number a relatively prime to m such that all partial quotients of a/m are ≤ 3. Furthermore the set of all m satisfying this condition but with all partial quotients ≤ 2 has positive density.

▸ **32.** [*M21*] Let $m_1 = 2^{31} - 1$ and $m_2 = 2^{31} - 249$ be the moduli of generator (38).

a) Show that if $U_n = (X_n/m_1 - Y_n/m_2) \bmod 1$, we have $U_n \approx Z_n/m_1$.

b) Let $W_0 = (X_0 m_2 - Y_0 m_1) \bmod m$ and $W_{n+1} = aW_n \bmod m$, where a and m have the values stated in the text following (38). Prove that there is a simple relation between W_n and U_n.

In the next edition of this book, I plan to introduce a new Section 3.3.5, entitled "The L^3 Algorithm." It will be a digression from the general topic of Random Numbers, but it will continue the discussion of lattice basis reduction in Section 3.3.4. Its main topic will be the now-classic algorithm of A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász [Math. Annalen **261** *(1982), 515–534] for finding a near-optimum set of basis vectors, and improvements to that algorithm made subsequently by other researchers. Examples of the latter can be found in the following papers and their bibliographies: M. Seysen, Combinatorica* **13** *(1993), 363–375; C. P. Schnorr and H. H. Hörner, Lecture Notes in Comp. Sci.* **921** *(1995), 1–12.*