SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 645–653, October 1991

003

A CLASS OF ARBITRARILY ILL CONDITIONED FLOATING-POINT MATRICES*

SIEGFRIED M. RUMP†

Abstract. Let $\mathbb{F}$ be a floating-point number system with basis $\beta \geqq 2$ and an exponent range consisting of at least the exponents 1 and 2. A class of arbitrarily ill conditioned matrices is described, the coefficients of which are elements of $\mathbb{F}$. Due to the very rapidly increasing sensitivity of those matrices, they might be regarded as "almost" ill posed problems.

The condition of those matrices and their sensitivity with respect to inversion is given by means of a closed formula. The condition is rapidly increasing with the dimension. For example, in the double precision of the IEEE 754 floating-point standard (base 2, 53 bits in the mantissa including implicit 1), matrices with $2n$ rows and columns are given with a condition number of approximately $4 \cdot 10^{32n}$.

Key words. condition number, sensitivity, ill conditioned, linear systems, floating-point number systems

AMS(MOS) subject classifications. 15A12, 65F05, 65G05

0. Introduction. It is a trivial fact that there are arbitrarily ill conditioned *real* matrices. In this paper we concentrate on matrices that are exactly representable in some floating-point number system $\mathbb{F}$. There is no restriction to the basis and only a trivial technical assumption on the exponent range of $\mathbb{F}$. For fixed $\mathbb{F}$ there are finitely many square matrices with n rows and a maximum condition number less than ∞ for given n.

The well-known schemes for constructing ill-conditioned matrices suffer from the fact that for given $\mathbb{F}$ only a few matrices are exactly representable in $\mathbb{F}$, say up to $n_{\max}$ rows. For $n > n_{\max}$ rows the entries are getting "too big." For example, let

$$(Z_n)_{ij} := \frac{\binom{n+i-1}{i-1} \cdot n \cdot \binom{n-1}{n-j}}{i+j-1},$$

as proposed by Zielke. For single precision in the IEEE 754 floating-point format (base 2 with 24 bit in the mantissa including implicit 1), we have (using infinity norm)

$$n_{\max}(Z_n) = 10 \quad \text{with } \|Z_{10}\| \cdot \|Z_{10}^{-1}\| \approx 2 \cdot 10^{14}.$$

From Pascal's triangle we get

$$(P_n)_{ij} := \binom{i+j-1}{i-1}$$

with

$$n_{\max}(P_n) = 15 \quad \text{with } \|P_{15}\| \cdot \|P_{15}^{-1}\| \approx 1 \cdot 10^{16}.$$

The classical example for ill-conditioned matrices is Hilbert matrices, the ijth component of which is $1/(i + j - 1)$. In order to make them exactly representable in a binary floating-point format, we may use their inverses, or we may multiply the entire matrix by $lcm(1, 2, \cdots, 2n - 1)$. We call the latter matrix H_n^*. Then

$$n_{\max}(H_n^{-1}) = 7 \quad \text{with } \|H_7\| \cdot \|H_7^{-1}\| \approx 5 \cdot 10^8$$

* Received by the editors August 31, 1989; accepted for publication (in revised form) March 13, 1990.
† Informatik III, Technische Universität, 2100 Hamburg 90, Federal Republic of Germany.

and

$$n_{\max}(H_n^*) = 10 \quad \text{with } \|H_{10}^*\| \cdot \|H_{10}^{*-1}\| \approx 2 \cdot 10^{13}.$$

The second method is obviously much more effective with respect to generating exactly representable ill-conditioned matrices. The class of matrices to be described in the following has no restriction in the dimension. In the single precision IEEE 754 floating-point number system, there are 10×10-matrices with condition number $1.1 \cdot 10^{78}$.

1. The class of matrices. Let $\mathbb{F}$ be a floating-point number system with base β, i.e., $\mathbb{F}$ consists of real numbers of the form

$$\chi = \pm 0 \cdot m_1 m_2 \cdots m_\lambda \cdot \beta^e \tag{1.1}$$

with

$$0 \leqq m_i < \beta \quad \text{for } 1 \leqq i \leqq \lambda \quad \text{and} \quad e_{\min} \leqq e \leqq e_{\max}. \tag{1.2}$$

We do not require numbers in the gradual underflow range and assume

$$m_1 \neq 0 \quad \text{if } \chi \neq 0. \tag{1.3}$$

Let $\mathbb{F}$ consist at least of all real numbers $\chi \in \mathbb{R}$ with a representation satisfying (1.1)–(1.3) and assume $e_{\min} \leqq 1$, $2 \leqq e_{\max}$.

Consider Pell's equation (see [1])

$$P^2 - k \cdot Q^2 = 1 \tag{1.4}$$

for positive integers P, Q, and k. If β is a square, let k be the smallest prime factor of β, otherwise set $k = \beta$. Then (1.4) has infinitely many solutions (P, Q) (see [1]).

Let P, Q be numbers satisfying Pell's equation (1.4) for some k and let

$$P = \sum_{\nu=0}^{n} p_i \cdot \sigma^i \quad \text{and} \quad Q = \sum_{\nu=0}^{n} q_i \cdot \sigma^i \tag{1.5}$$

with $p_n \neq 0$ or $q_n \neq 0$ for some $\sigma \in \mathbb{N}$ and $|p_i|, |q_i| < \sigma$, $i = 0 \cdots n$. Furthermore, in this section we assume that $0 \leqq p_i, q_i < \sigma$ for $i = 0 \cdots n$.

In practical applications a typical choice for σ is β^λ. However, in this section we are interested in minimum requirements for the floating-point number system $\mathbb{F}$. Therefore we set $\sigma = k$.

For $\sigma = k$ the numbers p_i, q_i are of $\mathbb{F}$ if $e_{\min} \leqq 1 \leqq e_{\max}$ and so is $k \cdot q_i$ because $k \cdot q_i < k^2 \leqq \beta^2$. To store the number 1 requires 1 to be an admissible exponent; to store $k \cdot q_i$ requires 1 or 2 to be admissible exponents. Therefore

$$p_i, k \cdot q_i, 1, \sigma \in \mathbb{F} \quad \text{if } e_{\min} \leqq 1 \quad \text{and} \quad 2 \leqq e_{\max}$$

and the matrix

$$c_n := \begin{bmatrix}
p_n & p_{n-1} & \cdots & p_1 & p_0 & kq_n & kq_{n-1} & & kq_1 & kq_0 \\
q_n & q_{n-1} & \cdots & q_1 & q_0 & p_n & p_{n-1} & & p_1 & p_0 \\
1 & -\sigma & & & & & & & & \\
 & 1 & -\sigma & & & & & & & \\
 & & \cdots & & & & & & & \\
 & & & 1 & -\sigma & & & & & \\
 & & & & & 1 & -\sigma & & & \\
 & & & & & & 1 & -\sigma & & \\
 & & & & & & & \cdots & & \\
 & & & & & & & & 1 & -\sigma
\end{bmatrix} \tag{1.6}$$

consists only of components that are exactly representable in $\mathbb{F}$. Since (1.4) has infinitely many solutions, the class of matrices C_n defined by (1.6) consists of elements with an arbitrarily large number of rows.

2. Properties of the matrices. In this section some properties of the matrices defined by (1.6) will be studied. Here, no restrictions on k or σ with respect to β are necessary; our only assumptions are (1.5) and (1.4). In the following, especially, the assumption $0 \leqq p_i, q_i < \sigma$ for $i = 0 \cdots n$ is not necessary.

Throughout this paper we use componentwise ordering of matrices, i.e., $A \leqq B : \langle = \rangle$ $a_{ij} \leqq b_{ij}$ and the componentwise absolute value $|A| = (|A_{ij}|)$, which is again a matrix.

The condition number $\|C_n\| \cdot \|C_n^{-1}\|$ for the ∞-norm will be calculated along with the sensitivity of C_n. Rohn, in [3], gave a nice definition of the sensitivity of a matrix C with respect to inversion: Let B be a matrix of relative distance less than or equal to α to C, i.e., $|B - C| \leqq \alpha \cdot |C|$, then

$$s_{ij}^{\alpha}(C) := \max \left\{ \frac{|B_{ij}^{-1} - C_{ij}^{-1}|}{|C_{ij}^{-1}|} ; |B - C| \leqq \alpha \cdot |C| \right\},$$

provided $C_{ij}^{-1} \neq 0$ and

$$s_{ij}(C) := \lim_{\alpha \to 0_+} \frac{s_{ij}^{\alpha}(C)}{\alpha}.$$

In [3], Rohn proves an explicit formula for the sensitivity matrix $S = (s_{ij}(C))$:

$$s_{ij}(C) = \frac{(|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{ij}}{|C^{-1}|_{ij}} \quad \text{for } C_{ij}^{-1} \neq 0. \tag{2.1}$$

LEMMA 1. $\det(C_0) = 1$, $\|C_0\|_\infty \|C_0^{-1}\|_\infty = (P + kQ)^2$, *and* $s_{ij}(C_0) = 4P^2 - 3$ *for* $i = j$ *and* $s_{ij}(C_0) = 4P^2 - 1$ *for* $i \neq j$.

Proof. For $n = 0$, (1.6) writes

$$C_0 = \begin{pmatrix} P & kQ \\ Q & P \end{pmatrix} \quad \text{with } C_0^{-1} = \begin{pmatrix} P & -kQ \\ -Q & P \end{pmatrix},$$

as follows from (1.4). Then the first two statements are obvious; for the third, a short computation yields

$$(s_{ij}(C_0)) = \begin{pmatrix} \zeta & \eta \\ \eta & \zeta \end{pmatrix} \quad \text{with } \zeta = P^2 + 3kQ^2, \qquad \eta = 3P^2 + kQ^2. \qquad \square$$

In the following we will show that for $n > 0$ the condition and sensitivity of C_n increase compared to those of C_0.

For the rest of the paper we frequently use

$$C := C_n \in \mathbb{R}^{(2n+2)\times(2n+2)} \quad \text{with components } c_{ij},\ 0 \leqq i,\ j \leqq 2n+1. \tag{2.2}$$

The indices of matrices start with 0 with the exception of A and B, to be defined later. Those are $(n + 1) \times n$-matrices with row indices starting with σ and column indices starting with 1.

LEMMA 2. *The matrices* C_n *are not singular*: $\det(C_n) = (-1)^n$.

Proof. Define

$$s := (\sigma^n, \sigma^{n-1}, \cdots, \sigma, 1)^t \in \mathbb{R}^{n+1} \tag{2.3}$$

and

$$x := \begin{pmatrix} P \\ -Q \end{pmatrix} \cdot s = \begin{pmatrix} P \cdot \sigma^n \\ \vdots \\ P \cdot 1 \\ -Q \cdot \sigma^n \\ \vdots \\ -Q \cdot 1 \end{pmatrix} \in \mathbb{R}^{2n+2}. \tag{2.4}$$

Then

$$(p_n, \cdots, p_0) \cdot s = P \quad \text{and} \quad (q_n, \cdots, q_0) \cdot s = Q, \tag{2.5}$$

and using (2.2),

$$\sum_{\nu=0}^{2n+1} c_{0\nu} \cdot x_\nu = P^2 - kQ^2 = 1,$$

$$\sum_{\nu=0}^{2n+1} c_{1\nu} \cdot x_\nu = PQ - QP = 0 = \sum_{\nu=0}^{2n+1} c_{i\nu} \cdot x_\nu \quad \text{for } i \geqq 2.$$

This means that x is the first column of C^{-1} and, especially,

$$(C^{-1})_{2n+1,0} = -Q. \tag{2.6}$$

Therefore $-Q = -\det(\bar{C})/\det(C)$ with

$$\bar{C} := \begin{pmatrix} q_n \cdots q_0 & p_n \cdots p_1 \\ \Sigma & 0 \\ 0 & \Sigma^* \end{pmatrix},$$

and

$$\Sigma := \begin{pmatrix} 1 & -\sigma & & \\ & 1 & -\sigma & \\ & & \cdots & \\ & & 1 & -\sigma \end{pmatrix}, \qquad \Sigma^* := \begin{pmatrix} 1 & -\sigma & & \\ & 1 & -\sigma & \\ & & \cdots & \\ & & & 1 \end{pmatrix}.$$

But $\det(\bar{C}) = \det(\bar{\bar{C}})$ with

$$\bar{\bar{C}} := \begin{pmatrix} q_n \cdots q_0 \\ \Sigma \end{pmatrix}$$

and $\bar{\bar{C}} \cdot s = Q \cdot e$ with $e = (1, 0, \cdots, 0)^t$. This implies that

$$(\bar{\bar{C}}^{-1})_{00} = \sigma^n / Q = \det(\hat{C})/\det(\bar{\bar{C}})$$

with

$$\hat{C} := \begin{pmatrix} -\sigma & & & \\ 1 & -\sigma & & \\ & \cdots & & \\ & & 1 & -\sigma \end{pmatrix} \in \mathbb{R}^{n \times n}, \qquad \det(\hat{C}) = -(1)^n \cdot \sigma^n.$$

Therefore

$$\det(C) = \frac{\det(\bar{C})}{Q} = \frac{\det(\bar{\bar{C}})}{Q} = \frac{\det(\hat{C}) \cdot Q}{\sigma^n \cdot Q} = (-1)^n. \qquad \square$$

Next we calculate the inverse of $C = C_n$ explicitly. The first column is already given by (2.4), the second is given by

$$y := \begin{pmatrix} -k \cdot Q \cdot s \\ P \cdot s \end{pmatrix} \in \mathbb{R}^{2n+2}, \qquad C \cdot y = (0, 1, 0, \cdots, 0)^t. \tag{2.7}$$

Formulas (2.4) and (2.7) imply, especially, that $-Q$ and P are the first two elements of the last row of C^{-1}. Let

$$(-QP\alpha_n \cdots \alpha_1 \beta_n \cdots \beta_1) \in \mathbb{R}^{2n+2} \tag{2.8}$$

be the last row of C^{-1}. Then multiplication with the first $n+1$ columns of C yields

$$\begin{aligned} &-Q \cdot p_n + p \cdot q_n + \alpha_n = 0, \\ &-Q \cdot p_{n-1} + P \cdot q_{n-1} - \sigma \cdot \alpha_n + \alpha_{n-1} = 0, \\ &\qquad \cdots \\ &-Q \cdot p_1 + P \cdot q_1 - \sigma \cdot \alpha_2 + \alpha_1 = 0, \\ &-Q \cdot p_0 + P \cdot q_0 - \sigma \cdot \alpha_1 = 0. \end{aligned} \tag{2.9}$$

Setting $\alpha_0 = \alpha_{n+1} = 0$ by definition gives

$$-Q \cdot p_i + P \cdot q_i - \sigma \cdot \alpha_{i+1} + \alpha_i = 0 \quad \text{for } i = 0 \cdots n \tag{2.10}$$

and by successively adding the equations in (2.9), multiplied by σ, yields

$$\alpha_i = Q \cdot \sum_{\nu=i}^{n} p_\nu \cdot \sigma^{\nu-i} - P \cdot \sum_{\nu=i}^{n} q_\nu \cdot \sigma^{\nu-i} \quad \text{for } i = 1 \cdots n. \tag{2.11}$$

By treating the last $n+1$ columns of C in the same way, we obtain

$$\begin{aligned} &-k \cdot Q \cdot q_i + P \cdot p_i - \sigma \cdot \beta_{i+1} + \beta_i = 0 \quad \text{for } i = 1 \cdots n, \\ &-k \cdot Q \cdot q_0 + P \cdot p_0 - \sigma \cdot \beta_1 = 1, \end{aligned} \tag{2.12}$$

setting $\beta_0 = \beta_{m+1} = 0$ by definition, and

$$\beta_i = P \cdot \sum_{\nu=i}^{n} p_\nu \cdot \sigma^{\nu-i} - k \cdot Q \cdot \sum_{\nu=i}^{n} q_\nu \cdot \sigma^{\nu-i} \quad \text{for } i = 1 \cdots n. \tag{2.13}$$

According to our assumption (1.5), $p_n \neq 0$ or $q_n \neq 0$ and

$$\sum_{\nu=i}^{n} p_\nu \cdot \sigma^{\nu-i} < \sigma^n \leqq P \quad \text{or} \quad \sum_{\nu=i}^{n} q_\nu \cdot \sigma^{\nu-i} < Q \quad \text{for } i \geqq 1.$$

Moreover, gcd $(P, kQ) = 1$ such that (2.11) and (2.13) imply

$$\alpha_i \neq 0 \quad \text{and} \quad \beta_i \neq 0 \quad \text{for } i = 1 \cdots n. \tag{2.14}$$

Let $\iota_i \in \mathbb{R}^{n+1,n+1}$ be a matrix with 1 in the ith upper diagonal and 0 elsewhere such that

$$\iota_i \cdot s = (\sigma^{n-i}, \cdots, \sigma, 1, 0, \cdots, 0)^t \in \mathbb{R}^{n+1}, \tag{2.15}$$

using s from (2.3). Then we are ready to describe C^{-1} as follows.

LEMMA 3. *The inverse of* $C = C_n$ *defined by* (1.6) *is given by*

$$\left[\begin{array}{c|c|c|c} P\cdot S & -k\cdot Q\cdot S & B & k\cdot A \\ \hline -Q\cdot S & P\cdot S & A & B \end{array}\right] \begin{matrix} 0 \\ n+1 \\ n+2 \\ 2n+1 \end{matrix} \tag{2.16}$$

$$\begin{matrix} 0 & 1 & 2 & n+1 & n+2 & 2n+1 \end{matrix}$$

with

$$A := (\alpha_n s, \cdots, \alpha_1 s) \in \mathbb{R}^{n+1,n},$$

and

$$B := ((\beta_n I + \iota_n)\cdot s, \cdots, (\beta_1 I + \iota_1)\cdot s) \in \mathbb{R}^{n+1,n}.$$

Proof. For the matrices $A = (a_{ij})$ and $B = (b_{ij})$, we have

$$a_{ij} = \alpha_{n-j+1}\cdot\sigma^{n-i}, \quad \text{and}$$

$$b_{ij} = \begin{cases} \beta_{n-j+1}\cdot\sigma^{n-i}, & j \leqq i, \\ \beta_{n-j+1}\cdot\sigma^{n-i} + \sigma^{j-i+1} & j \geqq i+1 \end{cases} \tag{2.17}$$

for $i = 0\cdots n$, $j = 1\cdots n$ (the row indices start with 0, the column indices with 1). Denote the matrix defined by (2.16) by Γ. Then for $0 \leqq i, j \leqq n$, we have

$$(\Gamma\cdot C)_{ij} = P\cdot s_i\cdot p_{n-j} - k\cdot Q\cdot s_i\cdot q_{n-j} + b_{i,j+1} - \sigma\cdot b_{ij}$$

where the third summand cancels for $j = n$, the fourth for $j = 0$. Using $\beta_0 = \beta_{n+1} = 0$ and (2.17) yields

$$(\Gamma\cdot C)_{ij} = \begin{cases} t(i,j) & \text{for } j < i, \\ t(i,j) + \sigma^{j-i} & \text{for } j = i, \\ t(i,j) + \sigma^{j-i} + \sigma^{j-i-1} & \text{for } j > i \end{cases}$$

using the abbreviation

$$t(i,j) := \sigma^{n-i}\cdot(P\cdot p_{n-j} - k\cdot Q\cdot q_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}).$$

Therefore, for $0 \leqq i, j \leqq n$,

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(P\cdot p_{n-j} - k\cdot Q\cdot q_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}) + \delta_{ij} \tag{2.18}$$

using Kronecker's delta. Since later on we will need $|C^{-1}|\cdot|C|$, we write down the explicit formulae for the other components of $\Gamma\cdot C$. For $0 \leqq i \leqq n$, $n+1 \leqq j \leqq 2n+1$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot k\cdot(P\cdot q_{n-j} - Q\cdot p_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}); \tag{2.19}$$

for $n+1 \leqq i \leqq 2n+1$, $0 \leqq j \leqq n$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(-Q\cdot p_{n-j} + P\cdot q_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}); \tag{2.20}$$

and for $n+1 \leqq i, j \leqq 2n+1$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(-k\cdot Q\cdot q_{n-j} + P\cdot p_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}) + \delta_{ij}. \tag{2.21}$$

The identities (2.10) and (2.12) prove $(\Gamma\cdot C)_{ij} = \delta_{ij}$. □

For the condition of C using the ∞-norm and $\alpha_i \neq 0$,

$$\|C_n\|_\infty \cdot \|C_n^{-1}\|_\infty > \left\{ \sum_{\nu=0}^{n} (p_\nu + k \cdot q_\nu) \right\} \cdot \{\sigma^n \cdot (P + k \cdot Q)\}$$

$$= \left\{ \sum_{\nu=0}^{n} (\sigma^n p_\nu + k\sigma^n q_\nu) \right\} \cdot (P + kQ) \geqq (P + k \cdot Q)^2. \tag{2.22}$$

We calculate the sensitivity $s_{ij}(C)$ according to (2.1) for $0 \leqq i \leqq n$, $j = 0$. By (2.18) we have

$$(|C^{-1}| \cdot |C|)_{iv} \geqq \sigma^{n-i} \cdot (P \cdot |p_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}| + |\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|),$$

for $0 \leqq \nu \leqq n$ and by (2.19) we have

$$(|C^{-1}| \cdot |C|)_{iv} \geqq \sigma^{n-i} \cdot k \cdot (P \cdot |q_{n-\nu}| + Q \cdot |p_{n-\nu}| + |\alpha_{n-\nu}| + \sigma \cdot |\alpha_{n-\nu+1}|),$$

for $n + 1 \leqq \nu \leqq 2n + 1$.

Using $\alpha_\nu, \beta_\nu \neq 0$ we get, for $0 \leqq i \leqq n$,

$$(|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{i0}$$

$$= \sum_{\nu=0}^{n} (|C^{-1}| \cdot |C|)_{iv} \cdot |C^{-1}|_{\nu 0} + \sum_{\nu=n+1}^{2n+1} (|C^{-1}| \cdot |C|)_{iv} \cdot |C^{-1}|_{\nu 0}$$

$$\geqq \sigma^{n-i} \cdot \sum_{\nu=0}^{n} \{(P \cdot |p_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}|) \cdot P \cdot \sigma^{n-\nu} + k \cdot (P \cdot |q_{n-\nu}| + Q \cdot |p_{n-\nu}|) \cdot Q \cdot \sigma^{n-\nu}\}$$

$$+ \sigma^{n-i} \cdot \left\{ \sum_{\nu=0}^{n} (|\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|) \cdot P \cdot \sigma^{n-\nu} \right.$$

$$\left. + \sum_{\nu=0}^{n} (|\alpha_{n-\nu}| + \sigma \cdot |\alpha_{n-\nu+1}|) \cdot kQ\sigma^{n-\nu} \right\}$$

$$\geqq \sigma^{n-i} \cdot P \cdot (P^2 + kQ^2 + kQ^2 + kQ^2) + \sigma^{n-i} \cdot P \cdot 4$$

$$= \sigma^{n-i} \cdot P \cdot (4P^2 - 3 + 4) > \sigma^{n-i} \cdot P \cdot (4P^2)$$

using $k \cdot Q \geqq P$. Together with $|C^{-1}|_{i0} = \sigma^{n-i} \cdot P \neq 0$,

$$S_{i0}(C) > 4P^2 \quad \text{for } 0 \leqq i \leqq n$$

follows. This proves the following theorem.

THEOREM 4. *The matrix C defined by* (1.6) *satisfies*

$$\|C\|_\infty \cdot \|C^{-1}\|_\infty > (P + k \cdot Q)^2$$

and there are components of C of which the sensitivity defined by (2.1) *is greater than* $4 \cdot P^2$.

3. Some examples. For given k, suitable pairs (P, Q) satisfying Pell's equation $P^2 - k \cdot Q^2 = 1$ are easily generated. Given some (P_0, Q_0) unequal, the trivial solution is $(1, 0)$, and successive solutions are

$$(P_{i+1}, Q_{i+1}) = (P_i P_0 + k Q_i Q_0, Q_i P_0 + P_i Q_0).$$

For a floating-point number system given by (1.1)–(1.3), a choice for σ is β^λ. Any expansion (1.5) of P, Q is suitable. The coefficients p_i, q_i are calculated successively.

Some bits can be saved by the following observation. If some coefficient p_i is divisible by β or by a power of β, then p_i and the following p_j, $j > i$ are expressed with a corresponding exponent. If the last digit m_λ in the mantissa of p_{i+1} is equal to $\beta - 1$, then p_i can be replaced by $p_i - \sigma$ and p_{i+1} by $p_{i+1} + 1$, the latter being divisible by β.

For example, let $P = 73942$, $\beta = 10$, $\sigma = 100$. Then expanding P yields $(p_2, p_1, p_0) = (7, 39, 42)$ and this is reduced by the method just described to $(p_1, p_0) = (74 \cdot 10^1, -58)$. This method is especially useful for base 2.

For a given number P, the corresponding coefficients p_i, $i = 0 \cdots n$ can be calculated by the following algorithm:

$e = 0$; $i = 0$;
repeat
  while $P \bmod \beta = 0$ do $\{P = P/\beta; e = e+1\}$;
  $q = \lfloor P/\sigma \rfloor$; $r = P - \sigma \cdot q$;
  if $(q \bmod \beta \neq \beta - 1)$ *or* $(q < \beta)$
    then $\{p_j = r \cdot \beta^e; P = q\}$
    else $\{p_i = (r - \sigma) \cdot \beta^e; P = q+1\}$;
  $i = i+1$
until $P = 0$;

For $k = 2$, successive pairs P, Q are $(3, 2)$, $(17, 12)$, $(99, 70) \cdots$. In Table 1 we display some values for p_i, q_i for single and double precision. For the individual value of n (resulting in a $2n \times 2n$-matrix C) we choose the maximum values (P, Q) being representable by $(p_{n-1}, \cdots, p_0)$ and $(q_{n-1}, \cdots, q_0)$. In the columns of Table 1, the condition number is given followed by the coefficients p_i and q_i, both in descending order. The coefficients are given by two numbers m and e such that $m \cdot 2^e$ is the actual coefficient. For example, $q_4 = 1175 \cdot 2^{22}$ for $n = 5$ (yielding a 10×10-matrix). Our algorithm yields a higher condition than the expected maximum $4 \cdot 2^{24 \cdot 2n} \approx 7 \cdot 10^{72}$, especially for this 10×10-matrix.

For double precision we choose different values for k yielding the coefficients in Table 2. These coefficients are, of course, only samples used to construct matrices of the general form (1.6). We conclude by writing the 6×6-matrix for single precision explicitly.

TABLE 1
p_i, q_i for binary format, 24 bit precision; $k \equiv 2$.

Cond	1.3E+030	2.2E+044	6.5E+060	1.1E+078	4.8E+090	1.7E+107
p_i	15248163 2	3527199 3	6929233 6	425393 14	2161033 8	8490761 10
	11171905 0	6746489 1	9763077 3	6127903 11	5075327 7	15520103 6
		−8816797 0	12608263 1	−10707825 7	8241033 6	6855055 5
q_i	84235 9		−6160127 0	7194379 1	−9934673 5	6997339 4
	−3559681 3	1247053 4		−2285085 0	−5752371 1	−11831695 3
		13508351 2	1224927 8		12291875 0	9051609 1
		−14061827 1	−5131195 6	1175 22		−11093871 0
			14870387 5	−14199789 15	47753 13	
			−7145793 4	12492253 13	−15523515 12	3001937 11
				9093109 10	−1620555 9	12103369 10
				10074835 1	14867027 6	−13213329 9
					14366575 3	−9497253 7
					−4879973 1	−3241495 4
						8507481 3
						−1367575 2

TABLE 2
p_i, q_i for binary format, 53 bit precision.

Cond k	7.0E+066 32	3.4E+097 2	2.1E+131 32	1.4E+164 2
p_i	8384758637032543 5	119071610094027 9	1838140087490775 8	1217131843483323 9
	−3529290569461695 0	−3183251058136493 3	−6618243915631817 2	5555590710757647 8
		−8183182949466111 0	−7698164339527309 1	−1048381871128883 4
q_i	5928919690858185 3		5251261634103185 0	4113071334050663 3
	−6097772977423311 1	84196342944287 9		−3228782923936605 0
		891386017353869 8	162470165079445 9	
		−1900818942150157 7	6774769086897599 6	1721284360250283 8
			4831599480133437 3	292142371452983 6
			−5900891544265983 0	−4351444206118847 4
				1403045714199203 2
				−2787903664869301 1

It is exactly storable with only 24 bits in the mantissa (and therefore in almost any floating-point number system) but matrix inversion will "fail" in almost any floating-point format available because, due to the condition number $2.2 \cdot 10^{44}$, an equivalent of approximately 44 decimal digits precision would be necessary:

$$\begin{pmatrix} 3527199\cdot 2^3 & 6746489\cdot 2^1 & -8816797\cdot 2^0 & 1247053\cdot 2^5 & 13508351\cdot 2^3 & -14061827\cdot 2^2 \\ 1247053\cdot 2^4 & 13508351\cdot 2^2 & -14061827\cdot 2^1 & 3527199\cdot 2^3 & 6746489\cdot 2^1 & -8816797\cdot 2^0 \\ 1 & -2^{24} & & & & \\ & 1 & -2^{24} & & & \\ & & & 1 & -2^{24} & \\ & & & & 1 & -2^{24} \end{pmatrix}.$$

To generate this matrix, the values $P = 7942546277405390632803$ and $Q = 5616228332641321147898$ have been used.

MATLAB [2] delivers as an estimation for the condition number of the matrix the (almost) correct answer ∞.

REFERENCES

[1] G. H. HARDY AND E. WRIGHT, *An Introduction to the Theory of Numbers*, Fifth Edition, Oxford Science Publications, Oxford, 1980, 1981, p. 442.

[2] PRO-MATLAB *User's Guide*, Vers. 32-SUN, The MathWorks, Inc., Sherborn, MA, 1987.

[3] J. ROHN, *New condition numbers for matrices and linear systems*, Computing, 41 (1989), pp. 167–169.