A class of arbitrarily ill-conditioned floating-point matrices*

Siegfried M. Rump

Abstract

Let $\mathbb{F}$ be a floating-point number system with basis $\beta \geq 2$ and an exponent range consisting at least of 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 als "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. E.g. 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 $4 \cdot 10^{32n}$.

AMS classification: 15A12, 65F05, 65G05

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

0 Introduction

It is a trivial fact that there are arbitrarily ill-conditioned *real* matrices. In this paper we concentrate on matrices which are exactly representable in som 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, despite infinity, a worst condition for given n.

*SIAM J. Matrix Anal. Appl. (SIMAX), 12(4):645–653, 1991

The well-known schemes for constructing ill-conditioned matrices suffer from the fact that for given $\mathbb{F}$ only 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

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

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)

$$n_{\max}(Z_n) = 10 \quad \text{with} \quad \|Z_{10}\| \cdot \|Z_{10}^{-1}\| \approx 2.1014.$$

From Pascal's triangle we get

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

with

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

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

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

and

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

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

$$\chi = \pm 0.m_1 m_2 \ldots m_\lambda \cdot \beta^e \tag{1}$$

with

$$0 \leq m_i < \beta \quad \text{for} \quad 1 \leq i \leq \lambda \quad \text{and} \quad e_{\min} \leq e \leq e_{\max}. \tag{2}$$

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

$$m_1 \neq 0 \quad \text{if} \quad \chi \neq 0. \tag{3}$$

Let $\mathbb{F}$ consist at least of all real numbers $\chi \in \mathbb{R}$ with a representation satisfying (1.1), (1.2) and (1.3) and assume $e_{\min} \leq 1,\ 2 \leq e_{\max}$.

Consider Pell's equation (see [1])

$$p^2 - k \cdot Q^2 = 1 \tag{4}$$

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

$$\begin{array}{l} P = \sum\limits_{\nu=0}^{n} p_i \cdot \sigma^i \quad \text{and} \quad Q = \sum\limits_{\nu=0}^{n} q_i \cdot \sigma^i \\ \text{with } p_n \neq 0 \text{ or } q_n \neq 0 \text{ for some } \sigma \in \mathbb{N} \text{ and} \\ |p_i|, |q_i| < \sigma,\ i = 0 \dots n. \end{array} \tag{5}$$

Furthermore we assume for this section $0 \leq p_i,\ q_i < \sigma$ for $i = 0 \dots n$.

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 $e_{\min} \leq 1 \leq e_{\max}$ and so is $k \cdot q_i$ because $k \cdot q_i < k^2 \leq \beta$. To store the number 1 requires 1 to be an admissable exponent, to store σ requires 1 or 2 to be admissable exponents. Therefore

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

and the matrix

$$C_n := \begin{pmatrix} p_n & p_{n-1} & \dots & & p_1 & p_0 & kq_n & kq_{n-1} & & & kq_1 & kq_0 \\ q_n & q_{n-1} & \dots & & q_1 & q_0 & p_n & p_{n-1} & & & p_1 & p_0 \\ 1 & -\sigma & & & & & & & & & & \\ & 1 & -\sigma & & & & & & & & & \\ & & & \dots & & & & & & & & \\ & & & & 1 & -\sigma & & & & & & \\ & & & & & & 1 & -\sigma & & & & \\ & & & & & & & 1 & -\sigma & & & \\ & & & & & & & & & \dots & & \\ & & & & & & & & & & 1 & -sigma \end{pmatrix} \tag{6}$$

consists only of components being 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 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 w.r.t. β are necessary; our only assumptions are (1.5) and (1.4). In the following especially the assumption $0 \leq p_i,\ q_i < \sigma$ for $i = 0 \ldots n$ is not necessary.

Throughout this paper we use componentwise ordering of matrices, i.e. $A \leq B :\Leftrightarrow a_{ij} \leq b_{ij}$ 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 and the sensitivity of C_n. Rohn gave in his paper [3] a nice definition of the sensitivity of a matrix C w.r.t. inversion: Let B be a matrix of relative distance $\leq \alpha$ to C, i.e.

$$|B - C| \leq \alpha \cdot |C|$$

then

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

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 $S = (s_{ij}(C))$:

$$s_{ij}(C) = \frac{(|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{ij}}{|C^{-1}|_{ij}} \quad \text{for } C_{ij}^{-1} \neq 0. \tag{7}$$

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

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

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

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

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

For the rest of the paper we frequently use

$$C := C_n \in \mathrm{IR}^{(2n+2)\times(2n+2)} \text{ with components} c_{ij}, 0 \leq i, j \leq 2n+1. \tag{8}$$

The indices of matrices start with = with the exception of A and B to be defined later on. 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; it is $\det(C_n) = (-1)^n$.

Proof. Define

$$s := (\sigma^n, \sigma^{n-1}, \dots, \sigma, 1)^t \in \mathrm{IR}^{n+1} \tag{9}$$

and

$$x := \begin{pmatrix} P \cdot s \\ -Q \cdot s \end{pmatrix} = \begin{pmatrix} p & \cdot & \sigma^n \\ & \vdots & \\ p & \cdot & 1 \\ -Q & \cdot & \sigma^n \\ & \vdots & \\ -Q & \cdot & 1 \end{pmatrix} \in \mathrm{IR}^{2n+2}. \tag{10}$$

Then

$$(p_n, \dots, p_0) \cdot s = P \text{ and } (q_n, \dots, q_0) \cdot s = Q \tag{11}$$

and using (2.2)

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

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

That means x is the first column of C^{-1} and especially

$$(C^{-1})_{2n+1,0} = -Q. \tag{12}$$

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

$$\overline{C} := \begin{pmatrix} q_n \dots q_0 & p_n \dots p_1 \\ \Sigma & 0 \\ 0 & \Sigma^* \end{pmatrix} \text{ and}$$

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

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

$$\overline{\overline{C}} := \begin{pmatrix} q_n \dots q_0 \\ \Sigma \end{pmatrix}$$

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

$$(\overline{\overline{C}}^{-1})_{00} = \sigma^n / Q = \det(\widehat{C}) / \det(\overline{\overline{C}}) \text{ with}$$

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

Therefore

$$\det(C) = \frac{\det(\overline{C})}{Q} = \frac{\det(\overline{\overline{C}})}{Q} = \frac{\det(\widehat{C}) \cdot Q}{\sigma^n \cdot Q} = (-1)^n. \qquad \blacksquare$$

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

$$y := \begin{pmatrix} -k \cdot Q \cdot c \\ \hline p \cdot s \end{pmatrix} \in \mathbb{R}^{2n+2}, \quad C \cdot y = (0, 1, 0, \dots, 0)^t. \tag{13}$$

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

$$(-Q \; P \; \alpha_n \dots \alpha_1 \; \beta_n \dots \beta_1) \in \mathbb{R}^{2n+2} \tag{14}$$

be the last row of C^{-1}.

Then multiplication with the first $n+1$ columns of C yields

$$\begin{array}{lclclclcl} -Q\cdot p_n & + & p\cdot q_n & + & \alpha_n & & & = & 0 \\ -Q\cdot p_{n-1} & + & p\cdot q_{n-1} & - & \sigma\cdot\alpha_n & + & \alpha_{n-1} & = & 0 \\ & & & \cdots & & & & & \\ -Q\cdot p_1 & + & p\cdot q_1 & - & \sigma\cdot\alpha_2 & + & \alpha_1 & = & 0 \\ -Q\cdot p_0 & + & p\cdot q_0 & - & \sigma\cdot\alpha_1 & & & = & 0 \end{array} \tag{15}$$

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

$$-Q\cdot p_i + p\cdot q_i - \sigma\cdot\alpha_i + 1 + \alpha_i = 0 \quad \text{for } i = 0\dots n \tag{16}$$

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

$$\alpha_i = Q\cdot\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} - p\cdot\sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} \quad f \tag{17}$$

By treating the last $n+1$ columns of C in the same way gives

$$\begin{array}{lclclclcll} -k\cdot Q\cdot q_i & + & p\cdot p_i & - & \sigma\cdot\beta_{i+1} & + & \beta_i & = & 0 & \quad\text{for } i = 1\dots n \\ -k\cdot Q\cdot q_0 & + & p\cdot p_0 & - & \sigma\cdot\beta_1 & & & = & 1 & \end{array} \tag{18}$$

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

$$\beta_i = p\cdot\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} - k\cdot Q\cdot\sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} \quad \text{for } i = 1\dots n. \tag{19}$$

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

$\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} < \sigma^n \leq p$ or $\sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} < Q$ for $i \geq 1$.

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

$$\alpha_i \neq 0 \quad \text{and} \quad \beta_i \neq 0 \quad \text{for } i = 1\dots n. \tag{20}$$

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

$$\iota_i\cdot s = (\sigma^{n-i}, \dots, \sigma, 1, 0, \dots 0)^t \in \mathbb{R}^{n+1} \tag{21}$$

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

Lemma 2. The inverse of $C = C_n$ defined by (1.6) is given by

$$\begin{array}{cc} \left(\begin{array}{c|c|c|c} P\cdot S & -k\cdot Q\cdot S & B & k\cdot A \\ \hline -Q\cdot S & P\cdot S & A & B \end{array}\right) & \begin{array}{c} 0 \\ n+1 \\ \\ n+2 \\ 2n+1 \end{array} \end{array} \tag{22}$$

with

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

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

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

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

$$b_{ij} = \left(\begin{array}{lllll} \beta_{n-j+1} & \cdot & \sigma^{n-i} & & j \le i \\ \beta_{n-j+1} & \cdot & \sigma^{n-i} & + \quad \sigma^{j-i+1} & j \ge i+1 \end{array}\right. \tag{23}$$

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

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

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 $0 \le i, j \le n$

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(P\cdot p_{n-j} - k\cdot Q\cdot q_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}) + \delta_{ij} \tag{24}$$

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

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot k\cdot(P\cdot q_{n-j} - Q\cdot p_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}), \tag{25}$$

for $n+1 \le i \le 2n+1, \quad 0 \le j \le n$

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(-Q\cdot p_{n-j} + P\cdot q_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}) \tag{26}$$

and for $n+1 \le i, \quad j \le 2n+1$

$$(\Gamma \cdot C)_{ij} = \sigma^{n-i} \cdot (-k \cdot Q \cdot q_{n-j} + P \cdot p_{n-j} + \beta_{n-j} - \sigma \cdot \beta_{n-j+1}) + \delta_{ij}. \tag{27}$$

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

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

$$\begin{aligned} \|C_n\|_\infty \cdot \|C_n^{-1}\|\|_\infty &> \Big\{\sum_{\nu=0}^{n}(p_\nu + k \cdot q_\nu)\Big\} \cdot \Big\{\sigma^n \cdot (P + k \cdot Q)\Big\} \\ &= \Big\{\sum_{\nu=0}^{n}(\sigma^n p_\nu + k\sigma^n q_\nu)\Big\} \cdot (P + kQ) \ge (P + k \cdot Q)^2. \end{aligned} \tag{28}$$

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

$$(|C^{-1}| \cdot |C|)_{i\nu} \ge \sigma^{n-i} \cdot (P \cdot |P_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}| + |\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|)$$

for $0 \le \nu \le n$ and by (2.19)

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

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

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

$$\begin{aligned} (|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{i0} &= \sum_{\nu=0}^{n}(|C1{-}1| \cdot |C|)_{i\nu} \cdot |C^{-1}|_{\nu 0} + \sum_{\nu=n+1}^{2n+1}(|C^{-1}| \cdot |C|)_{i\nu} \cdot |C^{-1}|_{\nu 0} \\ &\ge \sigma^{n-i} \cdot \sum_{\nu=0}^{n}\Big\{(P \cdot |P_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}|) \cdot p \cdot \sigma^{n-\nu} + k \cdot (P \cdot |q_{n-\nu}| + Q \cdot |p_{n-\nu}| \cdot Q \cdot \sigma^{n-\nu})\Big\} \\ &+ \sigma^{n-i} \cdot \Big\{\sum_{\nu=0}^{n}(|\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|) \cdot P \cdot \sigma^{n-\nu} + \sum_{\nu=0}^{n}(|\alpha_{n-\nu}| + \sigma \cdot |\alpha_{n-\nu+1}| \cdot k \cdot Q \cdot \sigma^{n-\nu})\Big\} \\ &\ge \sigma^{n-i} \cdot p \cdot (P^2 + kQ^2 + kQ^2 + kQ^2) + \sigma^{n-i} \cdot P \cdot 4 \\ &= \sigma^{n-i} \cdot P \cdot (4P^2 - 3 + 4) > \sigma^{n-i} \cdot P \cdot (4P^2). \end{aligned}$$

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

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

This proves

Theorem 3. 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 the sensitivity defined by (2.1) of which 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 (1.0) 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), (1.2), (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 observing the following. If some coefficient p_i ist 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 is reduced by the method just described to $(p_1, p_0) = (74 \cdot 10^1, -58)$. Especially for base 2 this method is useful.

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

$e = 0; \quad i = 0;$
$\underline{\text{repeat}}$
    $\underline{\text{while}}\ P \bmod \beta = 0 \text{ do } \{P = P/\beta;\ e = e + 1\};$
    $q = [P/\sigma];\ r = P - \sigma \cdot q;$
    $\underline{\text{if}}\ (q \text{mod} \beta \neq \beta - 1)\ \underline{\text{or}}\ (q < \beta)$
        $\underline{\text{then}}\ \{p_i = r \cdot \beta^e; p = q\}$
        $\underline{\text{else}}\ \{p_i = (r - \sigma) \cdot \beta^e; P = q + 1\};$
    $i = i + 1$
$\underline{\text{until}}\ P = 0;$

For $k = 2$ successive pairs P, Q are $(3, 2), (17, 12), (99, 70) \ldots$. In the following we display some values for p_i, q_i for IEEE 754 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 $(p_{n-1}, \ldots, p_0)$ and $(q_{n-1}, \ldots, q_0)$.

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

In the columns of the table 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. E.g. $q_4 = 1175 \cdot 2^{22}$ for $n = 5$ (yielding a 10×10-matrix). Especially for this 10×10-matrix our algorithm yields a higher condition than the expected maximum $4 \cdot 2^{24.2n} \approx 7 \cdot 10^{72}$.

For double precision we choose different values for k yielding the following coefficients:

Table 2. p_i, q_i for binary format, 53 bits precision

These coefficients are, of course, only samples to construct matrices of the general form (1.6). We conclude with writing down the 6×6-matrix for single precision explicitely. 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:

$$\begin{pmatrix} 3527199 \cdot 2^3 & 6746489 \cdot 2^1 & -8816797 \cdot 2^0 & 1247053 \cdot 2^5 & 13508351 \cdot 2^3 & -14061827 \cdot 2^2 \\ 1247053 \cdot 2^4 & 13508351 \cdot 2^2 & -140061827 \cdot 2^1 & 3527199 \cdot 2^3 & 6746489 \cdot 2^1 & -8816797 \cdot 2^0 \\ 1 & -2^{24} & & & & \\ & 1 & -2^{24} & & & \\ & & & 1 & -2^{24} & \\ & & & & 1 & -2^{24} \end{pmatrix}$$

To generate this matrix the values

$P = 794254627740539063 2803$ and $Q = 561622833264132114 7898$

have been used.

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

References

[1] Hardy, G.H. and Edward Wright: An Introduction to the Theory of Numbers, 5th Edition, Oxford Science Publications, p. 442 (1980, 1981)

[2] PRO-MATLAB User's Guide, Vers. 32-SUN, The MathWorks. Inc. (1987)

[3] Rohn, J.: New Condition Numbers for Matrices and Linear Systems, COMPUTING 41, p. 167–169 (1989)