Electronic Journal: Southwest Journal of Pure and Applied Mathematics
Internet: http://rattler.cameron.edu/swjpam.html
ISSN 1083-0464
Issue 1, July 2002, pp. 33–38.
Submitted: November 13, 2001. Published: July 1 2002.

A RELIABLE METHOD TO DETERMINE THE ILL-CONDITION FUNCTIONS USING STOCHASTIC ARITHMETIC

Saeid Abbasbandy and M.A. Fariborzi

Abstract. One of the important matters in numerical methods is to find ill-condition functions which have instability in some points. In this paper it has been explained a method to estimate the number of common significant digits between the values of $f(x)$ and $f(x+\delta)$ where $|\delta|$ is small enough. Then we develop this idea for $f(x,y)$ and $f(x+\delta, y+\epsilon)$ where $|\delta|$ and $|\epsilon|$ are small enough. It has been proved two theorems which express the relationship between these values and the condition number of the function f. It has been defined in [4], the condition number of one dimensional function $f(x)$ and in [3] the number of common significant digits between two real numbers . It has been developed these definitions for two dimensional function $f(x,y)$. It has been explained in [9,10,11], one can use the CESTAC[1] method which is a method based on stochastic arithmetic in order to find $f(x)$ numerically. It is used this method to rely and validate the results and implemented the numerical examples. In the numerical examples it has been considered the Rump's function and has been shown it is ill-condition based on the CESTAC method and a kind of perturbation method which has been explained in [1].

1991 A.M.S. (MOS) Subject Classification Codes. 65C20, 65Y20, 68W20.

Key Words and Phrases. keywords Ill-posed function,Stochastic arithmetic, CESTAC method, Condition number, Common significant digits.

1. Introduction. It has mentioned in [3], the number of significant digits common between two distinct real numbers a and b , denoted by $C_{a,b}$, can be defined by

$$C_{a,b} = \log_{10} \mid \frac{a+b}{2(a-b)} \mid = \log_{10} \mid \frac{a}{a-b} - \frac{1}{2} \mid, \tag{1}$$

[1] Control et Estimation Stochastically den Almonds die Calculus

Department of Mathematics, Imam Khomeini International University
Department of Mathematics,Tehran Branch,Islamic Azad University
Sabandy@pnu.ac.ir and f-araghi@pnu.ac.ir

Typeset by $\mathcal{AMS}$-TEX

if $a = b$ then $C_{a,a} = +\infty$. Also if $\mid a - b \mid$ is small enough, one can take

$$C_{a,b} \simeq \log_{10} \mid \frac{a}{a-b} \mid . \tag{2}$$

One can use this definition in order to find the number of common significant digits between two ordered pairs (x, y) and (x', y').It will explain in the next section.

It has mentioned in [4], the condition number of the function f at x can be defined by

$$Cond(f(x)) = max\{|\frac{f(x) - f(x')}{f(x)}|/|\frac{x - x'}{x}| : |x - x'| \ is \ \ small\} \simeq |\frac{xf'(x)}{f(x)}|. \tag{3}$$

If $cond(f(x)) >> 1$ the function f is ill-condition at x. One can use this definition in order to find the condition number of two dimensional function f at (x, y).

2. Numerical accuracy of one variable functions.

Theorem 1. *If it is perturbed x to $x + \delta$ which $|\delta| << 1$ then the number of common significant digits between $f(x)$ and $f(x + \delta)$ is estimated as follows*

$$C_{f(x),f(x+\delta)} \simeq C_{x,x+\delta} - \log_{10} Cond(f(x)) + O(\delta), \tag{4}$$

Proof. According to (1)

$$C_{f(x),f(x+\delta)} = \log_{10} \mid \frac{f(x) + f(x+\delta)}{2(f(x) - f(x+\delta))} \mid = \log_{10} \mid \frac{\frac{f(x)+f(x+\delta)}{\delta}}{\frac{2(f(x)-f(x+\delta))}{\delta}} \mid$$

$$\simeq \log_{10} \mid \frac{f(x) + f(x+\delta)}{2\delta f'(x)} \mid = \log_{10} \mid \frac{2f(x) + O(\delta)}{2\delta f'(x)} \mid = \log_{10} \mid \frac{f(x)(1 + O(\delta))}{\delta f'(x)} \mid$$

$$= \log_{10} \mid \frac{x + O(\delta)}{\delta \frac{xf'(x)}{f(x)}} \mid = \log_{10} \mid \frac{\frac{x}{\delta} + O(1)}{\frac{xf'(x)}{f(x)}} \mid .$$

We conclude from (2) and (3)

$$C_{f(x),f(x+\delta)} \simeq \log_{10} \mid \frac{x}{\delta} \mid - \log_{10} Cond(f(x)) + O(\delta) = \log_{10} \mid \frac{x}{x + \delta - x} \mid$$

$$- \log_{10} Cond(f(x)) + O(\delta) \simeq C_{x,x+\delta} - \log_{10} Cond(f(x)) + O(\delta).$$

This result shows that if $Cond(f(x)) >> 1$ then $f(x)$ and $f(x + \delta)$ have not any common significant digits. In other words if f is ill-conditioned at x then $C_{f(x),f(x+\delta)} \leq 0$. Also if $0 \leq Cond(f(x)) \leq 1$ then $C_{f(x),f(x+\delta)} > 0$.It means that in this case f is well-conditioned. We observe that if $Cond(f(x)) \simeq 1$ then $C_{f(x),f(x+\delta)} \simeq C_{x,x+\delta}$.

3. Numerical accuracy of two variable functions.

At first we define the common significant digits between two distinct ordered pairs and the condition number of the function $f(x, y)$ as follows

Definition 1. We suppose that (x, y) and (x', y') are two distinct ordered pairs. Then

$$C_{(x,y),(x',y')} = \log_{10}(\frac{\| (x + x', y + y') \|_2}{2 \| (x - x', y - y') \|_2}), \tag{5}$$

if $(x, y) \simeq (x', y')$ then

$$C_{(x,y),(x',y')} \simeq \log_{10}(\frac{\| (x, y) \|_2}{\| (x - x', y - y') \|_2}). \tag{6}$$

Definition 2. The condition number of function $f : \Re^2 \to \Re$ at (x, y) is

$$Cond(f(x, y)) = max\{| \frac{f(x, y) - f(x', y')}{f(x, y)} | / \frac{\| (x - x', y - y') \|_2}{\| (x, y) \|_2} : |x - x'|, | y - y' | \ \ are \ \ small\},$$

(7)

$$= max\{\frac{| f(x, y) - f(x', y') |}{\| (x - x', y - y') \|_2} \| (x, y) \|_2 \ / \ | f(x, y) |: |x - x'|, | y - y' | \ \ are \ \ small\}.$$

We know that $f(x, y) - f(x', y') \simeq (x - x')\frac{\partial f}{\partial x} + (y - y')\frac{\partial f}{\partial y}$, therefore

$$Cond(f(x, y)) \simeq \frac{\sqrt{x^2 + y^2}}{| f(x, y) |}(| \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} |). \tag{8}$$

Theorem 2. *If it is perturbed (x, y) to $(x + \delta, y + \epsilon)$ which $|\delta| << 1$ and $|\epsilon| << 1$ then the number of common significant digits between $f(x, y)$ and $f(x + \delta, y + \epsilon)$ is estimated as follows*

$$C_{f(x,y),f(x+\delta,y+\epsilon)} \simeq C_{(x,y),(x+\delta,y+\epsilon)} - \log_{10} Cond(f(x, y)) + O(\delta, \epsilon), \tag{9}$$

Proof. Let $max\{| \delta |, | \epsilon |\} =| \eta |$. According to (5)

$$C_{f(x,y),f(x+\delta,y+\epsilon)} = \log_{10} | \frac{f(x, y) + f(x + \delta, y + \epsilon)}{2(f(x, y) - f(x + \delta, y + \epsilon))} |= \log_{10} | \frac{f(x, y) + f(x, y) + O(\eta)}{2(\delta\frac{\partial f}{\partial x} + \epsilon\frac{\partial f}{\partial y} + O(\eta^2))} |$$

$$= \log_{10} | \frac{f(x, y)(1 + O(\eta))}{\delta\frac{\partial f}{\partial x} + \epsilon\frac{\partial f}{\partial y} + O(\eta^2)} |= \log_{10} | \frac{\frac{1+O(\eta)}{|\delta|+|\epsilon|}}{\frac{\delta\frac{\partial f}{\partial x}+\epsilon\frac{\partial f}{\partial y}+O(\eta^2)}{f(x,y)(|\delta|+|\epsilon|)}} |\simeq \log_{10} | \frac{\frac{\sqrt{x^2+y^2}}{|\delta|+|\epsilon|} + O(1)}{\frac{\sqrt{x^2+y^2}}{f(x,y)}(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + O(\eta^2))} |$$

$$= \log_{10} \frac{\sqrt{x^2+y^2}}{\mid \delta \mid + \mid \epsilon \mid} - \log_{10} \mid \frac{\sqrt{x^2+y^2}}{f(x,y)} (\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}) \mid + O(\eta)$$

Since δ and ϵ are small hence

$$C_{(x,y),(x+\delta,y+\epsilon)} \simeq \log_{10} \mid \frac{\| (x,y) \|_2}{\| (\delta,\epsilon) \|_2} \mid \simeq \log_{10} \frac{\sqrt{x^2+y^2}}{\sqrt{\delta^2+\epsilon^2}} \simeq \log_{10} \frac{\sqrt{x^2+y^2}}{\mid \delta \mid + \mid \epsilon \mid}$$

We conclude from (6) and (8)

$$C_{f(x,y),f(x+\delta,y+\epsilon)} \simeq C_{(x,y),(x+\delta,y+\epsilon)} - \log_{10} Cond(f(x,y)) + O(\eta),$$

This result shows that if $Cond(f(x,y)) >> 1$ then $f(x,y)$ and $f(x+\delta,y+\epsilon)$ have not any common significant digits. In other words if f is ill-conditioned at (x,y) then $C_{f(x,y),f(x+\delta,y+\epsilon)} \leq 0$. Also if $0 \leq Cond(f(x,y)) \leq 1$ then $C_{f(x,y),f(x+\delta,y+\epsilon)} > 0$. It means that in this case f is well-conditioned.

In the next section we remind the main idea of the CESTAC method and stochastic arithmetic [9,10,11].

4. Stochastic Arithmetic-CESTAC Method.

Let F be the set of all the values representable in the computer. Thus any value $r \in \mathbb{R}$ is represented in the form of $R \in F$ in the computer. It has been mentioned in [11] that in a binary floating-point arithmetic with P mantissa bits, the rounding error stems from assignment operator is

$$R = r - \epsilon 2^{E-p} \alpha. \tag{10}$$

In this relation ϵ is the sign of r and $2^{-p}\alpha$ is the lost part of the mantissa due to round-off error and E is the binary exponent of the result. In single precision case, $P = 24$ and in double precision case, $P = 53$. Also if the floating-point arithmetic is as rounding to $+\infty$ or $-\infty$ then $-1 \leq \alpha \leq 1$.

According to (10) , if you want to perturb the last mantissa bit of the value r, it is sufficient that you change α in the interval $[-1,1]$. In CESTAC method if the arithmetic is considered as rounding to $+\infty$ or $-\infty$, α can be considered as a random variable uniformly distributed on $[-1,1]$. Thus R, the calculated result, is a random variable and its precision depends on its mean (μ) and its standard deviation (σ).

In practice the samples R_i are obtained by perturbation of the last mantissa bit of every result R, then the mean of random samples R_i, that is $\overline{R} = \frac{\sum_{i=1}^{N} R_i}{N}$, is considered as the result of an arithmetic operation. If $N = 3$, it has been proved in [5,11] that the number of decimal significant digits common to $\overline{R}$ and to the exact value r can be estimated by

$$C_{\overline{R},r} = \log_{10} \frac{\mid \overline{R} \mid}{\sigma} - 0.39. \tag{11}$$

In this formula σ,the standard deviation of the samples R_i, is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(R_i - \overline{R}\right)^2}{N-1}}.$$

In CESTAC method if $C_{\overline{R},r} \leq 0$, the informational result R is insignificant and it means a numerical instability exists in its related result.

In order to simultaneous implementation of the CESTAC method we should substitute a stochastic arithmetic in place of the floating-point arithmetic. In this way every arithmetic operation is run N times synchronously before running the next operation. The term set of stochastic numbers, denoted S, is applied to the set of Gassing random variables. An element $R \in S$ is denoted $R = (\mu, \sigma^2)$ where μ is the mean value of R and σ its standard deviation. The definitions of stochastic zero and arithmetical operations and comparative operators have been defined in [9,11].

Definition 3. $R \in S$ is a "stochastic zero", if and only if

$$C_{\overline{R},r} \leq 0 \quad or \quad \overline{R} = 0.$$

Definition 4. If $R \in S$ is a "stochastic zero" then the notation @0 is used to show the detection of this case in implementation of the CESTAC method. Hence if $C_{\overline{R},r} \leq 0$ or $\overline{R} = 0$ then we write $R = @0$.

Let $a, b, c \in F$, in order to implement the arithmetical operator $c = a\omega b$, first for any of the values a and b, it is obtained $N = 3$ random samples as it was mentioned. In this case the operation ω is in the form of $c_i = a_i\omega b_i;\ i = 1, 2, 3$. Let μ_a, μ_b and μ_c be the means and σ_a^2, σ_b^2 and σ_c^2 be the variance of the random samples a_i, b_i and c_i respectively. The result c from implementation the stochastic operation S^ω between two random variables $a, b \in S$ is an element of S. Its mean and variance can be obtained directly using the samples a_i and b_i [1,2]. At last μ_c is considered as the result of the stochastic operation S^ω. σ_c is used for estimating the number of decimal significant digits of this result using relation (11).

5. Numerical Examples.

In this section three examples have been considered. At first it is shown the instability of the functions at the given points using theorems 1 and 2, then these instabilities are obtained in the stochastic arithmetic using the CESTAC method. The programs have been provided with Fortran 90.

example 1. In this example it has been considered the function $f = x^2 + x - 1150$ [7]. It is shown for the values near the roots this is an ill-conditioned function. One of the exact roots is $\alpha \simeq 33.415336$. Let $x = 33.4153$, then $Cond(f(x)) = 895796.60000$ and therefore $\log_{10} Cond(f(x)) = 5.952209$ and $C_{x,\alpha} = 5.9424851$. Hence according to the theorem 1, $C_{f(x),f(\alpha)} < 0$. It means that f is ill-conditioned at x. Now if it is used the stochastic arithmetic and is found $f(x)$ at $x = 33.4153$, it has been detected the stochastic zero $f(x) = @0$.

example 2. In this example the Wilkinson polynomial $p(x) = \prod_{i=1}^{20}(x - i)$ [8]. One can estimate the value of $p(x)$ at points near the roots and find the instability of this polynomial at them. We know that $p'(x) = \sum_{i=1}^{20} \frac{p(x)}{x-i}$. Thus if we consider

$x = 1.000001$ then $Cond(p(x)) = 1048573.0000$ and $C_{1,x} = 6.0206001$ and therefore $\log_{10} Cond(p(x)) = 6.020599$. Hence according to the theorem 1, $C_{p(1),p(x)} \simeq 0$. In the floating-point arithmetic with single precision the value of polynomial is $p(x) = -1.160094E + 11$. Using the CESTAC method in order to find $p(x)$ at $x = 1$ it is obtained $p(x) = 4.833746e+09$ in company with the notation $p(x) = @0$. It means that the solution at this point is not valid and there is an instability in finding the exact root $x = 1$.

example 3. In this example we consider the Rump's function[10]

$$f(x,y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + 5.5y^8 + \frac{x}{2y}$$

We evaluate this function at $(x,y) = (77617, 33096)$. In the floating point arithmetic with double precision the value of function is $f(x,y) = -1.390612133896486E + 021$. Also using the relation (8) the condition number of f is $Cond(f(x,y)) = 1.665159112522263E + 016$ thus $\log_{10} Cond(f(x,y)) \simeq 16.2215$. Also if we suppose $\delta = \epsilon = 0.0001$ then $C_{(x,y),(x+\delta,y+\epsilon)} = 8.62520220154603$. According to theorem 2 we conclude $C_{f(x,y),f(x+\delta,y+\epsilon)} < 0$. It means this function is ill-conditioned at this point. Using the stochastic arithmetic we can validate this result. In this arithmetic the value f is a stochastic zero. In other words $f(x,y) = @0$. Consequently the result of the floating-point arithmetic is not reliable and this function has an instability at this point.

6. Conclusion.

There is a relationship between the condition number of a function at a given point and the significant digits common to this point and its perturbation. If it is considered an ill-conditioned function because it has a large condition number, it occurs an instability at the final result. But in many cases, the calculation of the condition number is difficult. Therefore one can use the CESTAC method to find this instability. Using the stochastic arithmetic one can validate the result and find whether it is reliable or not.

References

1. S. Abbasbandy and M.A. Fariborzi Araghi, *The Usage of the Stochastic Arithmetic in the Accuracy Estimation of the Numerical Algorithms.*
2. S. Abbasbandy and M.A. Fariborzi Araghi, *The valid implementation of numerical integration method using Stochastic arithmetic.*
3. J.M. Chesneaux and F. Jezequel, *Dynamical Control of Computations Using the Trapezoidal and Simpson's rules*, JUCS **4** (1998), 2–10.
4. S.D. Conte and C. De Boor, *Elementary numerical analysis*, International student edition, 1981.
5. M. Maille, *Some methods to estimate accuracy of measurements or numerical Computations,* (1982), 495–503.
6. R.E. Moore, *Methods and Applications of Interval Analysis*, SIAM studies in Applied Mathematics, Philadelphia, Pennsylvania, 1979.
7. G.M. Phillips and P.J. Taylor, *Theory and applications of Numerical Analysis*, 1980.
8. J.M. Ortega, *Numerical Analysis : a second course*, 1990.
9. M. Pichat, *Chaotic evolution and stochastic arithmetic.*
10. F. Toutounian, *The use of the CADNA library for validating the numerical results of the hybrid GMRES algorithm*, Applied Numerical Mathematics **23** (1997), 275–289.
11. J. Vignes, *A stochastic arithmetic for reliable scientific computation*, Math. and Comp. in Sim. **35** (1993), 233-261.