Journal of Inequalities in Pure and Applied Mathematics

http://jipam.vu.edu.au/

Volume 3, Issue 3, Article 44, 2002

BOUNDING THE MAXIMUM VALUE OF THE REAL-VALUED SEQUENCE

EUGENE V. DULOV AND NATALIA A. ANDRIANOVA

DEPARTAMENTO DE MATHEMÁTICAS,
FACULTAD DE CIENCIAS
UNIVERSIDAD NACIONAL DE COLOMBIA,
BOGOTÁ, COLOMBIA
edulov@matematicas.unal.edu.co

DEPARTMENT OF MATHEMATICS AND MECHANICS,
ULYANOVSK STATE UNIVERSITY
432700, LEO TOLSTOY ST., 42
ULYANOVSK, RUSSIA
nandrian2000@yahoo.com

Received 12 June, 2001; accepted 14 April, 2002
Communicated by A. Rubinov

ABSTRACT. For the given arbitrary sequence of real numbers $\{x_i\}_{i=1}^n$ we construct several lower and upper bound converging sequences. Our goal is to localize the absolute value of the sequence maximum. Also we can calculate the value of such numbers. Since the proposed algorithms are iterative, asymptotical convergence theorems are proved.

The presented task seems to be pointless from the ordinary point of view, but we illustrate its importance for a set of applied problems: matrix analysis, measurement data processing and Monte Carlo methods. According to the modern conception of fault tolerant computations, also known as "interval analysis", these results could also be treated as a part of interval mathematics.

Key words and phrases: Interval analysis, Maximum value, Data processing.

1991 *Mathematics Subject Classification.* 11K31, 65Gxx, 15A42, 11K45.

1. INTRODUCTION

We deal with an arbitrary sequence of real numbers $\{x_i\}_{i=1}^n$. If all the sequence numbers are explicitly given, an exact maximum (or it's absolute value) along with a quantity of such values are searched directly.

The problem becomes harder if the sequence is not explicitly given and we are supplied only it's mean value or in general — by power sums

$$s_k = s(k) = \sum_{i=1}^{n} x_i^k, \ k - \text{natural}.$$

ISSN (electronic): 1443-5756

049-02

For a variety of tasks we must also calculate the quantity of numbers, which are, by modulus equivalent to the maximal one. Thus, we define *multiplicity* as a quantity of numbers whose modulus is equal to the absolute value of a sequence maximum.

Moreover, if $\{x_i\}_{i=1}^n$ is stochastic, the usual meaning of the "maximum" becomes quite arbitrary. Therefore considering a sequence of lower and upper bounds for the maximum (as embedded intervals) seems to be reasonable. This idea leads us to the well-known estimation, given for example in [9]:

Lemma 1.1. *If $x_1, \ldots, x_n$ are real numbers, such that $0 \le x_n \le x_{n-1} \le \ldots \le x_1$, then*

$$\frac{\sum_{i=1}^n x_i}{n} + \sqrt{\frac{1}{n(n-1)} \sum_{j=1}^n \left(x_j - \frac{\sum_{i=1}^n x_i}{n} \right)^2} \le x_1. \tag{1.1}$$

This lemma, in modified form, see [7], may be used for estimating its maximal value by the absolute value of the number in the sample. (In the following work we use the standard notion of sample when referring to a sequence $\{x_i\}_{i=1}^n$).

Lemma 1.2. *Considering the real valued sample $\{y_i\}_{i=1}^n$, with $k \ge 1$ some integer,*

$$\left[\frac{\sum_{i=1}^n y_i^{2k}}{n} + \sqrt{\frac{1}{n(n-1)} \sum_{j=1}^n \left(y_j^{2k} - \frac{\sum_{i=1}^n y_i^{2k}}{n} \right)^2} \right]^{\frac{1}{2k}} \le \max_i |y_i|. \tag{1.2}$$

The above lemmas have an evident connection in statistics:

$$M\{x\} = \frac{\sum_{i=1}^n x_i}{n},\ D\{x\} = \frac{1}{n} \sum_{j=1}^n \left(x_j - \frac{\sum_{i=1}^n x_i}{n} \right)^2 .$$

Here $x \stackrel{\text{def}}{=} \{x_i\}_{i=1}^n$. According to one cornerstone theorem in statistics (see [3] for example):

$$\mathrm{P}\left(M\{x\} + \sqrt{D\{x\}} \le \max_i |x_i| \right) \to 1 \,. \tag{1.3}$$

Moreover, one can directly check, that the correctness of the inequality depends only on the multiplicity.

2. ESTIMATION OF PROPER BOUNDS

Taking into account that

$$D\{x\} = M\{(x - M\{x\})^2\} = M\{x^2\} - (M\{x\})^2 = \frac{s_2}{n} - \frac{s_1^2}{n^2} \,, \tag{2.1}$$

we will investigate the properties of the generalized sequence

$$f_k(x,p) = \left[\frac{\sum_{i=1}^n x_i^k}{n} + \sqrt{p \left(\frac{\sum_{i=1}^n x_i^{2k}}{n} - \frac{\left(\sum_{i=1}^n x_i^k \right)^2}{n^2} \right)} \right]^{\frac{1}{k}} = \left[\frac{s_k}{n} + \sqrt{p \left(\frac{s_{2k}}{n} - \frac{s_k^2}{n^2} \right)} \right]^{\frac{1}{k}} . \tag{2.2}$$

The left hand side of expression (1.2) is equivalent to (2.2) for $p = \frac{1}{n-1}$. In formula (2.2) we directly use the power sums s_k, mentioned in the introduction. In what follows we shall take $k = 2^j$ in (2.2), which is equivalent to the consequent squaring of each number in the sample. Under this supposition, sequence (2.2) is proved to be at least linearly convergent, depending on the parameter p.

The fact that the generalised sequence $f_k\left(x, \frac{1}{n-1}\right)$ converges to $\max_i |x_i|$ from below can be found in [7]. Here we investigate a more general result of (2.2) using other techniques.

Theorem 2.1. *For k a natural number,*

$$f_k^{\mathrm{up}}(x,m) = \left[\frac{s(2^k)}{n} + \sqrt{\frac{n-m}{m}\left(\frac{s(2^{k+1})}{n} - \frac{s^2(2^k)}{n^2}\right)}\right]^{2^{-k}}$$

is a decreasing sequence such that

$$f_1^{\mathrm{up}}(x,m) \geq f_2^{\mathrm{up}}(x,m) \geq \ldots \geq f_k^{\mathrm{up}}(x,m) \geq \ldots \geq \max_i |x_i|$$

of upper bound estimations for the modulus of the largest value in the sample $x = \{x_i\}_{i=1}^n$. Here $m, m < n$ is a multiplicity of $\max_i |x_i|$.

Proof. Assume without loss of generality that values $x_{\mathrm{m}} = \max_i |x_i|$ are the first numbers in the sample. Hence $s(\cdot)$ can be written as

$$s(2^k) = m x_{\mathrm{m}}^{2^k} + \sum_{i=1}^{n-m} x_i^{2^k} \ .$$

Denoting $\Sigma_1 = \sum_{i=1}^{n-m} x_i^{2^k}$, $\Sigma_2 = \sum_{i=1}^{n-m} x_i^{2^{k+1}}$, the basic inequality theorem $f_k^{\mathrm{up}}(x,m) \geq x_{\mathrm{m}}$ translates into the equivalent one

$$\sqrt{\frac{n-m}{m}\left[nm x_{\mathrm{m}}^{2^{k+1}} + n\Sigma_2 - m^2 x_{\mathrm{m}}^{2^k} - 2m x_{\mathrm{m}}^{2^k}\Sigma_1 - (\Sigma_1)^2\right]} \geq (n-m)x_{\mathrm{m}}^{2^k} - \Sigma_1.$$

Squaring and collecting similar terms gives

$$\frac{n-m}{m}\left[m(n-m)x_{\mathrm{m}}^{2^{k+1}} - \Sigma_1^2 - 2m x_{\mathrm{m}}^{2^k}\Sigma_1 + n\Sigma_2\right] \geq (n-m)^2 x_{\mathrm{m}}^{2^{k+1}} + \Sigma_1^2 - 2(n-m)x_{\mathrm{m}}^{2^k}\Sigma_1,$$

and finally $(n-m)\Sigma_2 \geq \Sigma_1^2$. According to (2.1) we have the inequality

$$(n-m)\sum_{i=1}^{n-m}\left(x_i^k - \frac{\sum_{i=1}^{n-m} x_i^k}{n-m}\right)^2 \geq 0$$

fulfilled $\forall x_i$ real and $\forall k = 1, 2, \ldots$. □

Theorem 2.2. *For k a natural number,*

$$f_k^{\mathrm{low}}(x,m) = \left[\frac{s(2^k)}{n} + \sqrt{\frac{n-(m+1)}{m+1}\left(\frac{s(2^{k+1})}{n} - \frac{s^2(2^k)}{n^2}\right)}\right]^{2^{-k}}$$

is an increasing sequence such that

$$f_1^{\mathrm{up}}(x,m) \leq f_2^{\mathrm{up}}(x,m) \leq \cdots \leq f_k^{\mathrm{up}}(x,m) \leq \cdots \leq \max_i |x_i|$$

of lower bound estimations for the modulus of the largest value in the sample $x = \{x_i\}_{i=1}^n$. Here $m, m < n$ is a multiplicity of $\max_i |x_i|$.

Proof. Under the same suppositions as above, the main inequality $f_k^{\mathrm{up}}(x,m) \leq \max_i |x_i|$ can be written as

$$\sqrt{p\left[m(n-m)x_{\mathrm{m}}^{2^{k+1}} + n\Sigma_2 - 2m x_{\mathrm{m}}^{2^k}\Sigma_1 - (\Sigma_1)^2\right]} \leq (n-m)x_{\mathrm{m}}^{2^k} - \Sigma_1 \ .$$

Further simplification leads us to

$$p(m(n-m)x_{\mathrm{m}}^{2^{k+1}} - 2mx_{\mathrm{m}}^{2^k}\Sigma_1 + n\Sigma_2 - (\Sigma_1)^2) \\ \leq (n-m)^2 x_{\mathrm{m}}^{2^{k+1}} - 2(n-m)x_{\mathrm{m}}^{2^k}\Sigma_1 + (\Sigma_1)^2,$$

$$(n-m)x_{\mathrm{m}}^{2^{k+1}}[n-m-pm] + 2x_{\mathrm{m}}^{2^k}\Sigma_1[pm-n+m] + (\Sigma_1)^2(1+p) \geq pn\Sigma_2,$$

$$(n-m(1+p))\left[(n-m)x_{\mathrm{m}}^{2^{k+1}} - 2x_{\mathrm{m}}^{2^k}\Sigma_1 + \frac{(\Sigma_1)^2}{n-m} - \frac{(\Sigma_1)^2}{n-m}\right] + (1+p)(\Sigma_1)^2 \geq pn\Sigma_2,$$

and

$$\frac{n-m(1+p)}{n-m}\left((n-m)x_{\mathrm{m}}^{2^k} - \Sigma_1\right)^2 + \left(1+p-\frac{n-m(1+p)}{n-m}\right)(\Sigma_1)^2 \geq pn\Sigma_2 \,.$$

Simplifying the second factor we have

$$\frac{n-m(1+p)}{n-m}\left((n-m)x_{\mathrm{m}}^{2^k} - \Sigma_1\right)^2 + \frac{pn}{n-m}(\Sigma_1)^2 - pn\Sigma_2 \geq 0.$$

Analyzing the first summand we see that

$$\frac{n-m(1+p)}{n-m} \geq 0 \;\Leftrightarrow\; p \leq \frac{n-m}{m} \,.$$

is a necessary, but not sufficient positivity condition.

Transferring the second and third summands to the right and reducing by a $(n-m)$ multiplier gives us

$$(n-m(1+p))((n-m)x_{\mathrm{m}}^{2^k} - \Sigma_1)^2 \geq pn((n-m)\Sigma^2 - (\Sigma_1)^2) \,, \tag{2.3}$$

in which the right hand side attains its maximum with a non-zero number $x = x_{\mathrm{m}}^{2^k} - \varepsilon$, where ε is a positive infinitesimal number. Substitution in (2.3) results in the inequality

$$(n-m(1+p))(n-m-1)x_{\mathrm{m}}^{2^k} + \varepsilon)^2 \geq pn(n-m-1)(x_{\mathrm{m}}^{2^k} - \varepsilon)^2 \,.$$

Upon expansion

$$x_{\mathrm{m}}^2(n-m-1)[(n-m(1+p))(n-m-1) - pn] \\ + \varepsilon^2(n-m-pm-pn^2+pnm+pn) \\ + 2x_{\mathrm{m}}\varepsilon(n-m-1)[n-m-pm+pn] \geq 0 \,. \tag{2.4}$$

Simplified factors at elements x_{m}^2, ε^2 and $x_{\mathrm{m}}\varepsilon$ respectively, we have

$$(i) \quad (n-m)(n-m-1)[(n-m-1) - p(m+1)]$$

$$(ii) \quad (n-m)[p(n-1)+1] \quad \text{and}$$

$$(iii) \quad 2(n-m)(n-m-1)[1+p].$$

If there exists a parameter $p \geq 0$ for which all three coefficients are positive, then our proposition is proved. Since the second and third coefficients give inequalities $p \geq -\frac{1}{n-1}$ and $p \geq -1$ respectively, (2.4) is fulfilled iff

$$n-m-1-p(m+1) \geq 0 \Leftrightarrow p \leq \frac{n-(m+1)}{m+1} < \frac{n-m}{m} \,.$$

□

Corollary 2.3. *In the case when $m = n-1$ and $p = \frac{1}{n-1}$, Theorem 2.2 provides an alternative proof of the convergence theorem (Theorem 4) in* [7].

Remark 2.4. According to Theorems 2.1 and 2.2 bounding sequences are monotonically increasing or decreasing, depending on parameter p. So, if we have at least two estimations f_1, f_2, or consequent f_k, f_{k+1} in general, we can calculate differences $\Delta_k(p) = f_{k+1}(x,p) - f_k(x,p)$ for consequent $p(l),\ l = 1, \ldots, n-1$.

The pair of numbers $(l+1, l)$ for which $\Delta_k(p(l)) \times \Delta_k(p(l+1)) < 0$, shows a change of convergence character for (2.2) i.e. indicating the multiplicity of maximum modulus to be $m = l$.

The following example below illustrates this remark. Let $\{x_i\}_{i=1}^{7} = \{5, 2, -1, -5, 4, -3, 5\}$. Here $x_{\mathrm{m}} = 5$ and $m = 3$. In the table we represent approximate values $f_2(p(l)) - f_1(p(l))$ for consequent $l = 1, \ldots, 6$

multiplicity	1	2	3	4	5	6
difference	-0.6	-0.259	-0.0887	0.02636	0.1184	0.2049

Table 2.1: Evaluating multiplicity

Since the difference changes it's sign for the pair $(3, 4)$, then $m = 3$.

Remark 2.5. Parameter $p = p(n, m)$ was introduced for two reasons:

(1) To provide strict lower and upper bounds for the maximum of the absolute value in the sample;
(2) To make this estimations more exact. Namely, for $k \to \infty$ we have

$$C = \lim_{k\to\infty} \frac{f_k(x,p)^{2^k}}{x_{\mathrm{m}}^{2^k}} = \frac{m + \sqrt{pm(n-m)}}{n}\ .$$

Setting $p = \frac{n-m}{m} \Leftrightarrow C = 1$. Thus, using "sample independent" parameter p, $|x_{\mathrm{m}}|$ could be better bounded.

Now we are ready to establish corresponding convergence theorems for the sequence (2.2).

3. CONVERGENCE ANALYSIS

Theorem 3.1. *Let $\varepsilon_k = x_{\mathrm{m}} - f_k(x,p)$ – be the estimation residual and $\delta = \frac{|x_2|}{x_{\mathrm{m}}} < 1$; $x_2 : |x_2| < x_{\mathrm{m}}$ be the second greatest, by absolute value, number in the sample of multiplicity l, then the asymptotic convergence speed of the sequence*

$$f_k(x,p) = \left[\frac{s(2^k)}{n} + \sqrt{p\left(\frac{s(2^{k+1})}{n} - \frac{s^2(2^k)}{n^2}\right)}\right]^{2^{-k}}$$

is

$$\lim_{k\to\infty} \frac{\varepsilon_{k+1}}{\varepsilon_k} = \lim_{k\to\infty} \frac{1}{1 + \left(\frac{m+\sqrt{pm(n-m)}}{n}\right)^{2^{-k-1}}} = \frac{1}{2}\ ,\ p \neq \frac{n-m}{m} \tag{3.1}$$

$$\lim_{k\to\infty} \frac{\varepsilon_{k+1}}{\varepsilon_k} = \frac{1}{2} \lim_{k\to\infty} \delta^{2^{k+1}}\ ,\ p = \frac{n-m}{m}. \tag{3.2}$$

Proof. Transforming $f_k(x,p)$ we have

$$\begin{aligned}\frac{f_k^{2^k}}{x_{\mathrm{m}}^{2^k}} &= \frac{nf_k^{2^k}}{nx_{\mathrm{m}}^{2^k}}\\ &= \frac{s(2^k)+\sqrt{p\left(ns(2^{k+1})-s^2(2^k)\right)}}{nx_{\mathrm{m}}^{2^k}}\\ &= \frac{m+\sum\limits_{i=1}^{n-m}\left(\frac{x_i}{x_{\mathrm{m}}}\right)^{2^k}+\sqrt{p\left(mn+n\sum\limits_{i=1}^{n-m}\left(\frac{x_i}{x_{\mathrm{m}}}\right)^{2^{k+1}}-\left[m+\sum\limits_{i=1}^{n-m}\left(\frac{x_i}{x_{\mathrm{m}}}\right)^{2^k}\right]^2\right)}}{n}.\end{aligned}$$

For sufficiently large k we have

$$F(k,p) = m+l\delta^{2^k}+\sqrt{p\left[(n-m)m+l(n-l)\delta^{2^{k+1}}-2lm\delta^{2^k}\right]}, \tag{3.3}$$

$$\left(\frac{f_k(x,p)}{x_{\mathrm{m}}}\right)^{2^k} = \frac{F(k,p)}{n},$$

with

$$\lim_{k\to\infty}F(k,p)=m+\sqrt{pm(n-m)}.$$

For $p=\frac{n-m}{m}$, expression (3.3) becomes

$$F(k) = m+l\delta^{2^k}+\sqrt{(n-m)^2+\frac{l(n-l)(n-m)}{m}\delta^{2^{k+1}}-2l(n-m)\delta^{2^k}} \tag{3.4}$$

$$\lim_{k\to\infty}F(k) = n\,.$$

Analysis of its residuals ratio gives, in general,

$$\lim_{k\to\infty}\frac{\varepsilon_{k+1}}{\varepsilon_k}=\lim_{k\to\infty}\frac{1-\left(\frac{F(k+1,p)}{n}\right)^{2^{-k-1}}}{1-\left(\frac{F(k,p)}{n}\right)^{2^{-k}}}\,.$$

Considering both parameter cases, we obtain

(1) $p\neq\frac{n-m}{m}$. Denoting $C=\frac{m+\sqrt{pm(n-m)}}{n}\neq 1$, we have

$$\lim_{k\to\infty}\frac{1-C^{2^{-k-1}}}{1-C^{2^{-k}}}=\lim_{k\to\infty}\frac{1-C^{2^{-k-1}}}{(1-C^{2^{-k-1}})(1+C^{2^{-k-1}})}=\lim_{k\to\infty}\frac{1}{1+C^{2^{-k-1}}}=\frac{1}{2}.$$

(2) for $p=\frac{n-m}{m}$. We analyze the influence of fast vanishing numbers, such that

$$\left(\frac{F(k)}{n}\right)^{2^{-k}}=\left(\frac{F(k)-n+n}{n}\right)^{2^{-k}}\approx 1+\frac{1}{2^k}\frac{F(k)-n}{n}\,. \tag{3.5}$$

Now (3.4) may be expressed in the form

$$c+\sqrt{b^2+a}\approx c+b+\frac{a}{2b}\,,\ 0\le a\ll 1$$

and an estimation for $F(k)-n$ is

$$-(n-m)+l\delta^{2^k}+(n-m)\left(1+\frac{l(n-l)}{2m(n-m)}\delta^{2^{k+1}}-\frac{l}{n-m}\delta^{2^k}\right)\,,$$

so that finally

$$F(k) - n \approx \frac{l(n-l)}{2m}\delta^{2^{k+1}}.$$

Substituting this approximation in (3.5) we obtain

$$\lim_{k\to\infty}\frac{\varepsilon_{k+1}}{\varepsilon_k} = \lim_{k\to\infty}\frac{1-1-\frac{l(n-l)}{mn}\frac{\delta^{2^{k+2}}}{2^{k+2}}}{1-1-\frac{l(n-l)}{mn}\frac{\delta^{2^{k+1}}}{2^{k+1}}} = \lim_{k\to\infty}\frac{2^{k+1}}{2^{k+2}}\frac{\delta^{2^{k+2}}}{\delta^{2^{k+1}}} = \frac{1}{2}\lim_{k\to\infty}\delta^{2^{k+1}}\ .$$

□

We illustrate Theorem 3.1 and Remark 3.2 by the same test sample, consider

$$\{x_i\}_{i=1}^{7} = \{5, 2, -1, -5, 4, -3, 5\},\ x_{\mathrm{m}} = 5,\ m = 3\ , \delta = 0.8\ .$$

The column pairs in Table 3.1 represent the numerically evaluated convergence ratio and the difference modulus between numerical and theoretical estimations.

Iter.	$p = 6$		$p = 1/6$	
	ratio	difference	ratio	difference
6	0.498144	$0.7826774e-4$	0.501926	$0.1253511e-3$
7	0.499033	$0.6200297e-7$	0.500901	$0.9950568e-7$
8	0.499516	$0.386122e-13$	0.500450	$0.629444e-13$
9	0.499758	$0.406412e-15$	0.500225	$0.247138e-15$
10	0.499879	$0.335205e-15$	0.5001125	$0.406203e-15$

Table 3.1: Asymptotical convergence rates: "non-optimal" case

Remark 3.2. Considering the content of Table 3.1, one can be confident in the convergence character described by the $\frac{m+\sqrt{pm(n-m)}}{n}$ summand. When p corresponds to multiplicity less than real one, this constant is greater than 1 and the numerical estimates (see column $p = 6$) converges to $\frac{1}{2}$ from below.

Vice versa, for p corresponding to greater multiplicities, this summand is less than 1 and the numerical estimates (see column $p = 1/6$) converges to $\frac{1}{2}$ from above.

According to computer FPU, limitations estimations for $k \geq 10$ are not reliable and outline the coincidence between theoretical and numerical results.

Table 3.2 presents the numerical estimations of residual ε and "error"–based calculated δ for the optimal parameter value $4/3$.

Iter.	ε_k	δ
2	$-2.4795358e-2$	0.8107121
3	$-2.1582945e-3$	0.8037043
4	$-3.0960442e-5$	0.8009546
5	$-1.2261571e-8$	0.7999936
6	$-3.84762e-15$	0.7999976

Table 3.2: Asymptotical convergence rates: optimal $p = 4/3$

Here we represent three examples of applying (2.2) in practice.

3.1. **Estimation of the Matrix Spectral Radius.** One can apply the introduced sequence in matrix analysis for bounding matrix spectral radius. According to the spectral property of a matrix trace operator we have

$$\sum_{i=1}^{n} \lambda_i^k = \operatorname{tr}\left\{A^k\right\},\ \forall k \geq 1\ ,$$

(see [2]) where λ_i denote eigenvalues of any matrix A. Hence, replacing $\sum_{i=1}^{n} x_i^k$ by $\operatorname{tr}\left\{A^k\right\}$ we obtain the required sequence. But these estimations are valid (compare with results in [7]) only for matrices with real spectrum, for example, a symmetric matrix.

For interested readers we recommend the recent articles [5, 6] and [8] and compare these results to the older ones [7] and [9]. The unique convergence speed estimation $\frac{1}{2}$ of this type was done by Friedland in [1]. He obtained the result

$$\rho(A) = \lim_{k\to\infty} \sqrt[2^k]{\left\|A^{2^k}\right\|_\infty}$$

to be linear. This upper bound estimation rises from matrix norm properties [2].

3.2. **Processing Data Measurements.** Experimental measurements are made by using sets of identical measuring units that are normally independent. Measurements are however, close enough to give detailed information about the device being tested.

Typically these units are equipped with several circuits, registering several observations during the external synchronization cycle. In this case, we are usually given the mean and dispersion of the internally registered sample. Moreover, the measurement scale is usually shifted to output only positive numbers.

According to Theorems 2.1 and 2.2, we can guarantee that $f_1(x, n-1) \leq x_{\mathrm{m}} \leq f_1(x, 1)$.

If we could construct measuring units producing s_4, s_6 (better s_8 than s_6) and consequent $s(2^k)$, then closer bounds for x_{m} can be obtained. According to the afore-mentioned theorems, we need to calculate differences $f_{k+1}(x, l) - f_k(x, l)$ for consequent $l = 1, \ldots, n-1$. The pair of indices $l+1, l$ locating the change of difference sign points to $m = l$. Hence the best currently available estimation will be $x_{\mathrm{m}} \in [f_k(x, m+1), f_k(x, m)]$ for the last made step k.

As we outlined in the introduction, the notion of the maximum or absolute maximum value of the noised measurement sample could be meaningless. In contrast, the set of embedded localization intervals containing this value is of great practical interest. The same principle concerns the next example, which could be treated as a generalization of this principle for large sample size n.

3.3. **Monte Carlo Methods.** Monte Carlo and quasi-Monte Carlo methods are now widely used in different fields of numerical modelling. Monte Carlo methods have their origins in physics and mathematics, and are now used in computer graphics, bioinformatics, geoscience and many other domains. Due to it's probabilistic properties and overall computational complexity, Monte Carlo algorithms are optimized for best computational performance.

Therefore finding a maximum over the used lattice translates into a programming problem. Moreover, since we can not guarantee that any one of the lattice points directly coincides with points of the global maximum (minimum), the sequence of nested intervals becomes the only reasonable approach to such a problem.

Consideration of computer hardware is an important component of the problem. Modern scalar, super-scalar, parallel and distributed computers (often referred to as *number crunchers*) employ branch prediction, prefetching and carrying to increase their performance. The branch direction cannot be predicted in a maximum search. This means that inside the main MCM cycle we must proceed through a branch prediction, which leads to a drastic drop of performance for

all modern CPU's (for example, Intel's Pentium 4 architecture will seriously suffer from the performance drop-down in contrary to AMD's Athlon).

Hence a reasonable compromise may be found in calculating several additional sums for $f(x_i)^2, f(x_i)^4, f(x_i)^8, \ldots$ by consecutive squaring (take into account the increased effectiveness of extended integer/float register files embedded in modern CPU's).

Leaving the main cycle we will have a number of sums

$$s_k = \frac{\sum_{i=1}^{n} f(x_i)^k}{n}, k = 1, 2, 4, \ldots$$

where n is a number of processed points. Consequently we can do the following:

(1) **Model 1.** Implement an aposteriori cycle to revise the behavior of differences $f_{k+1}(x,l) - f_k(x,l)$ for consecutive $l = 1, \ldots, n-1$. We need to determine a pair of numbers $l+1, l$ for which the sequence change a convergence character. Hence for $m = l$ the range estimation is $[f_k(x, m+1), f_k(x, m)]$ for the last available k.

Since the typical number of lattice points can be in the order of millions, the above mentioned approach can be treated as "doubling" the number of lattice points and being directly implemented, Model 1 is just a point of theoretical interest. However, at the end of this section we introduce an evident solution.

(2) **Model 2.** Without an additional cycle we can only estimate the lowest and highest interval bounds:

$$f_k(x, n-1) \le x_{\mathrm{m}} \le f_k(x, 1) \; .$$

This rough solution is reasonable for small n or fast changing functions.

Here we represent a small computational example in accordance to model 2. The scalar function

$$f(x) = e^{x+\sin(\pi(x-\frac{1}{2}))}, \; \max_{x\in[0,1]} f \approx 7.389056$$

is evaluated over a GLP (good lattice points, see [4]) set. A number of points taken is $n = 10000001$. To compare the computational time an experiment was repeated for $k = 1, 2$, $k = 1, 2, 4$ and $1, 2, 4, 8$ (calculating f_1, f_2 and f_4).

Table 3.3 below contains calculation times in seconds for the two processors, an *AMD K6-2/500* (weak floating point unit, strong branch prediction) and an *Intel P-II/350* (strong floating point unit and moderate branch prediction efficiency).

The third column of the table, entitled "upper" contains results calculated for $m = 2$ (applicable because test function has only one maximum over $[0, 1]$).

Experiment	Maximum	Upper	AMD	Intel
Direct	7.389049		36.3	42.9
$1, 2$	$[1.617683, 5866.923986]$	4149.015164	38.9	42.6
$1, 2, 4$	$[2.461533, 199.026414]$	167.365892	40.4	42.9
$1, 2, 4, 8$	$[3.730634, 36.416130]$	33.394119	48.3	43.2

Table 3.3: Numerical experiment timings

Computational times prove that for CPU with several floating-point out-of-order execution units we can replace the exhaustive maximum search by computation of two additional sums minimum. In this case we can use Model 1 to compute a bounding range.

The other results, concerning the lowest and highest boundaries, seems to be inapplicable. However, as a careful reader will see, that lattice points are very dense for large $n = 10000001$ and smooth function f. Hence, points close to $x = 1$ will be treated numerically as equal. This

leads to overestimating of upper bound and underestimating the lower one. For example, setting $m_1 = 400000, m_2 = 350000, k = 1, 2, 4, 8$ provides a much better estimation $7.389056 \in [7.349532, 7.469766]$

The simplest way to resolve this problem is by setting $m = m + \Delta$, Δ step have to be about 50000 for our example. This additional cycle will be repeated only 200 times, which is significantly less than a number of lattice points.

The other important remark is evident lattice dependence of multiplicity m. Taking the other GLP sequence for the same m_1, m_2, k provides a wrong localization interval $[7.954751, 8.080737]$. But setting $m_1 = 750000, m_2 = 700000$ gives the correct interval $[7.388547, 7.448697] \ni 7.389056$.

The detailed discussion of this numerical example highlighted the best possible way for locating maximum value (and it's multiplicity) — *binary search*. Applying it instead of fastened linear search will require only $\log_2 n = 20$ checks in our example.

4. CONCLUSIONS

In this paper we presented a non-standard, but powerful approach for solving some auxiliary tasks aimed at bounding the maximum by modulus value in the given sample. This approach is closely linked with current needs of interval analysis and can be neatly applied in several engineering and mathematical tasks, as was illustrated above. Accompanied by a binary search algorithm, this iterative bounding sequence can be successfully applied to MCM and quasi-MCM computational algorithms even for huge lattices and multi-dimensional tasks.

Having a computational shortage in matrix algebra due to a time consumable matrix squaring, this approach still is applicable in other cases.

REFERENCES

[1] S. FRIEDLAND, Revisiting matrix squaring, *Linear Algebra Appl.*, **154/156** (1991), 59–63.

[2] R.A. HORNAND C.R. JOHNSON, *Matrix Analysis* Cambridge Univ. Press, Cambrige 1986.

[3] G.A. KORNAND T.M. KORN, *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill Book Company Inc., New York 1969.

[4] H. NIEDERREITER Lattice rules for multiple integration, in Stochastic Optimization, K.MARTI, ed., *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Berlin 1992, 15–26.

[5] O. ROJO, Futher Bounds for the Smallest Singular Value and the Spectral Condition Number, *Computers Math. Applic.,* **38**(7-8) (1999), 215–228.

[6] H. ROJO, O. ROJO AND R. SOTO, Related Bounds for the Extreme Eigenvalues, *Computers Math. Applic.,* **38**(7-8) (1999), 229–242.

[7] O. ROJO, R. SOTOAND H. ROJO, Bounds of the Spectral Radius and the Largest Singular Value, *Computers Math. Applic.,* **36**(1) (1998), 41–50.

[8] O. ROJO, R. SOTOAND H. ROJO, Bounds for Sums of Eigenvalues and Applications, *Computers Math. Applic.,* **39**(7-8) (2000), 1–15.

[9] H. WOLKOWICZ AND G.P.H. STYAN Bounds for eigenvalues using traces, *Linear Algebra Applic.,* **29** (1980), 471–506.