Interval Computations
No 4(6), 1992

# ERROR AUTO-CORRECTION IN RATIONAL APPROXIMATION

Grigory L.Litvinov

An effect of error auto-correction for rational approximations to real functions is considered. This effect occurs in efficient methods of rational approximation (e.g., best approximations, Padé approximations, multipoint Padé approximations, linear and nonlinear Padé–Chebyshev approximations) where very significant errors in the coefficients do not affect the accuracy of the approximation. The thing is that the errors in the coefficients of a rational approximant are not distributed in an arbitrary way but form the coefficients of a new approximant to the approximated function. Concrete examples are presented. Standard methods of interval arithmetic do not allow to take into account the error auto-correction effect and, as a result, to estimate the error of the rational approximant accurately.

# АВТОКОРРЕКЦИЯ ПОГРЕШНОСТИ ПРИ РАЦИОНАЛЬНОЙ АППРОКСИМАЦИИ

Г.Л.Литвинов

Рассматривается эффект автокоррекции погрешности при построении рациональных приближений к вещественнозначным функциям. Эффект состоит в том, что для регулярных методов рациональной аппроксимации (включая наилучшие аппроксимации, аппроксимацию Паде, линейную и нелинейную аппроксимацию Паде-Чебышева и др.) очень значительные вычислительные погрешности коэффициентов приближения мало влияют на его точность.

An effect of error auto-correction for rational approximations to real functions is considered. This effect occurs in efficient methods of rational approximation (e.g., best approximations, Padé approximations, multipoint Padé approximations, linear and nonlinear Padé–Chebyshev approximations) where very significant errors in the coefficients do not affect the accuracy of the approximation. The thing is that the errors in the coefficients of a rational approximant are not distributed in an arbitrary way but form the coefficients of a new approximant to the approximated function. Understanding of the error auto-correction mechanism allows to decrease this error by varying the approximation procedure depending on the form of the approximant, see [1].

The author came across the phenomenon of error auto-correction at the end of the seventies while developing nonstandard algorithms for computing elementary functions on small computers, see [2]. It was required to construct rational approximants of the form

$$R(x) = \frac{a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n}{b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m} \qquad (1)$$

to certain functions of one variable $x$ defined on finite segments of the real line. For this purpose a simple version of the well-known linear Padé–Chebyshev method was used: the method allows to determine the family of coefficients $a_i$, $b_j$ of the approximant (1) as the solution of a certain system of linear algebraic equations. These systems turned out to be ill-conditioned, i.e., the problem of determining the coefficients of the approximant is, generally speaking, ill-posed and small perturbations of the approximated function $f(x)$ or calculation errors lead to considerable errors in the values of coefficients. Nevertheless, the method ensures a paradoxically high quality of the obtained approximants [2].

For example, for the function $\cos x$ the approximant of the form (1) on the segment $[-\pi/4, \pi/4]$ obtained by the method mentioned above for $m = 4$, $n = 6$ has the relative error equal to $0.55 \cdot 10^{-13}$, and the best possible relative error is $0.46 \cdot 10^{-13}$. The corresponding system of linear algebraic equations has the condition number of order $10^9$. Thus we risk losing 9 accurate decimal digits in the solution because of calculation errors. Computer experiments show that this is a serious risk. The method mentioned above was implemented as a Fortran program. The calculations were carried out with double precision (16 decimal positions) by means of ICL 4 50 and ES-1045 computers. These computers

are very similar in their architecture, but when passing from one computer to another the system of linear equations and the computational process are perturbed because of calculation errors, including round-off errors. As a result, the coefficients of the approximant mentioned above to the function $\cos x$ experience a perturbation already at the sixth–ninth decimal digits. But the error of the rational approximant itself remains invariant and is $0.4 \cdot 10^{-13}$ for the absolute error and $0.55 \cdot 10^{-13}$ for the relative error. The same thing happens for approximants of the form (1) to the function $\arctan x$ on the segment [-1,1] obtained by the method mentioned above for $m = 8$, $n = 9$ the relative error is $0.5 \cdot 10^{-11}$ and does not change while passing from ICL–4–50 to ES–1045 although the corresponding system of linear equations has the condition number of order $10^{11}$, and the coefficients of the approximant experience a perturbation with relative error of order $10^{-4}$.

Note that the application of standard procedures known in the theory of ill-posed problems results in this case in losses in accuracy. For example, if one applies the regularization method, two thirds of the accurate figures are lost; in addition, the amount of calculations increases rapidly. The thing is that the exact solution of the system of equations in the present case is not the ultimate goal; the aim is to construct an approximant which is precise enough.

Professor Yudell L. Luke kindly drew the author's attention to his papers [3] where the effect of error auto-correction for the classical Padé approximants was revealed and was explained at a heuristic level.

In [1], using theoretical arguments and the results of computer experiments, the error auto-correction mechanism is considered for quite a general situation. Let $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ and $\{\psi_0, \psi_1, \ldots, \psi_m\}$ be collections consisting of linearly independent functions of the argument $x$ belonging some (possibly multidimensional) set $X$. Consider the problem of constructing an approximant of the form

$$R(x) = \frac{a_0\varphi_0 + a_1\varphi_1 + \cdots + a_n\varphi_n}{b_0\psi_0 + b_1\psi_1 + \cdots + b_m\psi_m} \qquad (2$$

to a given function $f(x)$ defined on $X$. If $X$ coincides with a real line segment $[A, B]$, $\varphi_k = x^k$ and $\psi_k = x^k$ for all $k$, then the expression (2 turns out to be a rational function of the form (1). It is clear that expression (2) also gives a rational function in the case when we tak

Chebyshev polynomials $T_k$ or, for example, Legendre, Laguerre, Hermite, etc. polynomials as $\varphi_k$ and $\psi_k$.

Fix an abstract construction method (problem) for an approximant of the form (2) to the function $f(x)$. Quite often this problem is ill-posed. Let the coefficients $a_i$, $b_j$ give an exact or an approximate solution of this problem, and let the $\tilde{a}_i$, $\tilde{b}_j$ give another approximate solution obtained in the same way. Denote by $\Delta a_i$, $\Delta b_j$ the absolute errors of the coefficients, i.e., $\Delta a_i = \tilde{a}_i - a_i$, $\Delta b_j = \tilde{b}_j - b_j$; these errors arise due to perturbations of the approximated function $f(x)$ or due to calculation errors. Set $P(x) = \sum_{i=0}^{n} a_i\varphi_i$, $Q(x) = \sum_{j=0}^{m} b_j\psi_j$, $\Delta P(x) = \sum_{i=0}^{n} \Delta a_i\varphi_i$, $\Delta Q(x) = \sum_{j=0}^{m} \Delta b_j\psi_j$, $\tilde{P}(x) = P + \Delta P$, $\tilde{Q}(x) = Q + \Delta Q$.

It is easy to verify that the following exact equality is valid:

$$\frac{P+\Delta P}{Q+\Delta Q} - \frac{P}{Q} = \frac{\Delta Q}{\tilde{Q}}\left(\frac{\Delta P}{\Delta Q} - \frac{P}{Q}\right). \tag{3}$$

As mentioned above, the fact that the problem of calculating coefficients is ill-posed can nevertheless be accompanied by high accuracy of the approximants obtained. This means that the approximants $P/Q$ and $\tilde{P}/\tilde{Q}$ are close to the approximated function and, therefore, are close to each other, although the coefficients of these approximants differ greatly. In this case the relative error $\Delta Q/\tilde{Q} = \Delta Q/(Q+\Delta Q)$ of the denominator considerably exceeds in absolute value the left-hand side of equality (3). This is possible only in the case when the difference $\Delta P/\Delta Q - P/Q$ is small, i.e., the function $\Delta P/\Delta Q$ is close to $P/Q$, and, hence, to the approximated function. For "efficient" methods the function $\Delta P/\Delta Q$ provides indeed a good approximation for the approximated function, and, thus, $P/Q$ and $\tilde{P}/\tilde{Q}$ differ from each other by a product of small quantities in the right-hand side of (3). The thing is that the errors $\Delta a_i$, $\Delta b_j$ are not arbitrary, but are connected by certain relations.

Let an abstract construction method for the approximant of the form (2) be linear in the sense that the coefficients of the approximant can be determined from a homogeneous system of linear algebraic equations. The homogeneity condition is connected with the fact that, when multiplying the numerator and the denominator of fraction (2) by the same nonzero number, the approximant (2) does not change. Denote by $y$ the vector

whose components are the coefficients $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_m$. Assume that the coefficients can be obtained from the homogeneous system of equations

$$Hy = 0, \tag{4}$$

where $H$ is a matrix of dimension $(m+n+2) \times (m+n+1)$. The vector $\tilde{y}$ is an approximate solution of system (4) if the quantity $\|H\tilde{y}\|$ is small. If $y$ and $\tilde{y}$ are approximate solutions of system (4), then the vector $\Delta y = \tilde{y} - y$ is also an approximate solution of this system since $\|H\Delta y\| = \|H\tilde{y} - Hy\| \leqslant \|H\tilde{y}\| + \|Hy\|$. Thus it is natural to assume that the function $\Delta P/\Delta Q$ corresponding to the solution $\Delta y$ is an approximant to $f(x)$.

It is clear that the above reasoning is not rigorous; for each specific construction method for approximations it is necessary to carry out some additional analysis. The presence of the error auto-correction mechanism described above is also verified by a numerical experiment. The effect of error auto-correction reveals itself for certain nonlinear construction methods for rational approximations as well [1].

It can be easily understood that the standard methods of interval arithmetic (see, for example [4]) do not allow to take into account the error auto-correction effect and, as a result, to estimate the error of the rational approximant accurately.

## References

1. Litvinov, G.L. *Approximate construction of rational approximations and the effect of error autocorrection. Applications.* Russian J.of math.phys., **1** (3) (1993).

2. Litvinov, G.L. e.a. *Mathematical algorithms and programs for small computers.* Finansy i Statistika, Moscow, 1981 (in Russian).

3. Luke, Y.L. *Computations of coefficients in the polynomials of Padé approximations by solving systems of linear equations*, J. comp. and appl. math. **6** (3) (1980), pp. 213–218.

4. Alefeld, G. and Herzberger, J. *Introduction to interval computations.* Academic Press, New York, 1983.