# Stochastic Localization of Instability and Deterministic Enhancement of Accuracy for Iterative Algorithms

Philippe Langlois

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Stochastic Localization of Instability and Deterministic Enhancement of Accuracy for Iterative Algorithms*

Philippe LANGLOIS

**No 3966**

May 2000

THÈME 2


ISSN 0249-6399 ISRN INRIA/RR--3966--FR+ENG

# Stochastic Localization of Instability and Deterministic Enhancement of Accuracy for Iterative Algorithms

Philippe LANGLOIS *

Thème 2 — Génie logiciel
et calcul symbolique
Projet Arénaire

Rapport de recherche n ° 3966 — May 2000 — 10 pages

**Abstract:** Finite precision computations may affect the stability of iterative algorithms and the accuracy of computed solutions. Automatic approaches are proposed to control these effects as for example, the CESTAC and the CENA methods. We focus here on a complementary use of these two methods to localize unstable behavior of the algorithm, improve its stability and the accuracy of the solutions. We present computational experiments on ill-conditioned polynomial roots approximated with Newton's iteration.

**Key-words:** Finite Precision, Stability, Accuracy, Newton's Iteration, Polynomial Multiple Roots, CESTAC Method, CADNA Library, Automatic Correction, CENA Method

*(Résumé : tsvp)*

* Email: `Philippe.Langlois@ens-lyon.fr`. This manuscript is also available as a research report of the Laboratoire de l'Informatique du Parallélisme, Ecole Nornale Supérieure de Lyon, France, `http://www.ens-lyon.fr/LIP`.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

# Localisation stochastique de l'instabilité et amélioration déterministe de la précision pour des algorithmes itératifs

**Résumé :** Les calculs en précision finie peuvent affecter la stabilité des algorithmes itératifs et la précision des solutions calculées. Des approches automatiques sont proposées pour contrôler ces effets acomme par exemple, les méthodes CESTAC et CENA. Nous nous intéressons ici à une application complémentaire de ces deux méthodes pour localiser le comportement instable de l'algorithme, améliorer sa stabilité et la précision des solutions. Nous présentons des expérimentations numériques sur le calcul par la méthode de Newton de racines polynomiales mal conditionnées.

**Mots-clé :** Précision finie, stabilité, précision, Itération de Newton, Racine polynomiale multiple, méthode CESTAC, bibliothèque CADNA, correction automatique, méthode CENA.

# 1 Introduction

Finite precision corrupts fundamental properties of numerical algorithms that hold in exact arithmetic. Algorithm stability and/or computed solution accuracy may be affected in floating point computation. A well-known example is the computation with Newton's iteration of ill-conditioned roots of polynomial, *i.e.*, multiple or nearness roots [14].

Rounding errors analysis is principally performed "by hand" [5] but automatic approaches use the computer to provide more general and easy-to-use applications. In this paper, we focus on two automatic approaches : the CESTAC and the CENA methods. The CESTAC method, developed by J. Vignes and his colleagues since 1974, is a forward stochastic method [13]. It provides a dynamic estimation of the number of significant digits of computed results that allows instability localization and accuracy control. The aim of the CENA method is to improve the accuracy of the computed result [6]. Applying this accuracy improvement to sensitive intermediate variables may also stabilize algorithms suffering from rounding errors.

The two methods are complementary and the following scheme is natural to improve the numerical quality of an algorithm. The CESTAC method is first used to localize an unstable behavior of the algorithms and provides a reliable but inaccurate computed result. Then, the algorithm runs with the CENA method starting with the previously computed result. Hence, an accurate final result is computed without the computing overhead of the correcting method that is useless while the computation remains stable. In this paper, we propose to experiment this approach when ill-conditioned roots of polynomial are computed with Newton's iteration.

We briefly review the two methods in Section 2 and the characteristics of floating point computation of polynomial ill-conditioned roots in Section 3. In Section 4, we illustrate the efficiency of this approach and highlights finite precision effects with experimental results.

# 2 Stochastic and Deterministic Methods

We limit our presentation of these methods to the main and hereafter useful aspects. References are proposed in the bibliography. Excepted when indicated, computed values wear a hat in the sequel.

## 2.1 The CESTAC Method

The CESTAC method simulates stochastic arithmetic to provide an estimation of the number of significant digits in the computed result. This estimation allows the definition of the specific value $\underline{0}$ describing a zero value or a non-significant computed value. Underlying hypothesis concerns the probabilistic distribution of the elementary rounding errors and the predominance of their first order propagation. The CADNA library implements the CESTAC method and consists in computing with different rounding modes several samples of intermediate and final values. Therefore, branching tests and validity hypothesis are

dynamically controlled. Numerous examples of instability localization and accuracy control illustrate the practical efficiency of the method, see [1] for example and entries.

### 2.2 The CENA Method

The CENA method relies on a first order correction performed with automatic differentiation and the computation of elementary absolute rounding errors. It is a forward deterministic approach. Given a computed $\widehat{x}$, the method yields a corrected $\overline{x}$ defined as

$$\overline{x} = \widehat{x} + \widehat{\Delta}_L(x).$$

The correcting factor $\widehat{\Delta}_L(x)$ is the computed linearization of the global error in $\widehat{x}$ with respect to the elementary errors, *i.e.*, rounding errors introduced in the intermediate computations of $\widehat{x}$. As the corrected factor $\widehat{\Delta}_L$ and the correction also suffer from rounding errors, a validity bound $B_{E_C}$ of the corrected result is computed. Therefore, the exact result $x$ satisfies $x \in [\overline{x} - B_{E_C}, \overline{x} + B_{E_C}]$ assuming that the first order approximation of the global error is valid. This hypothesis is satisfied for linear algorithms we defined in [7]; polynomial evaluation by Horner's scheme is one of them. We apply the CENA method to improve the accuracy of the final result and/or sensitive intermediate variables to stabilize the algorithm. A complete description of the method is given by [6].

## 3 Computing Ill-Conditioned Roots of Polynomial

Finite precision effects on the computation of ill-conditioned roots of polynomial have been numerously analyzed in the past, *e.g.*, see [8] for entries. We present here the main useful properties for our purpose.

### 3.1 Ill-Condition and Attainable Accuracy

Ill-conditioned roots are sensitive to small perturbation of the polynomial coefficients. Such coefficient perturbation may be data errors, representation errors, *e.g.*, decimal to binary conversion, and/or may come from the backward effect of rounding errors in the algorithm provided by the backward analysis. In both cases, the computation may yield inaccurate roots and/or a wrong number of computed roots and hence with arbitrary order.

The following property highlights the sensitivity of polynomial roots [11]. Let $x^*$ be a root of multiplicity $m$ of polynomial $p(x) = a_n x^n + \cdots + a_0$. A relative error $\mathbf{u}$ of $a_i$ perturbs the root $x^*$ to $x^*(\mathbf{u})$ such that the forward error satisfies to a first order of approximation

$$x^*(\mathbf{u}) - x^* = \mathbf{u}^{1/m} \left[ -\frac{m!\, a_i x^{*i}}{p^{(m)}(x^*)} \right]^{1/m}. \tag{1}$$

Hence, multiple roots are always ill-conditioned and the forward error is $O(\mathbf{u}^{1/m})$. Single roots are also ill-conditioned when $\left| a_i x^{*i} / p'(x^*) \right| \gg \mathbf{u}$.

Let us remark that relation (1) yields the attainable accuracy, *i.e.*, the best forward error bound we can expect when the root finding method evaluates $p(x)$, *e.g.*, Newton's method. Let $\delta(p,x)$ be the absolute rounding error in $p(x)$ evaluation, $\widehat{p}(x) = p(x) + \delta(p,x)$, and $\delta$ such that $|\delta(p,x)| \leq |a_{i_0} x^{*i_0}| \leq \delta$. The best computed approximation $\widehat{x}^*$ of $x^*$ satisfies $\widehat{p}(\widehat{x}^*) = 0$. From relation (1), the attainable accuracy is

$$|\widehat{x}^* - x^*| \leq \left( \frac{m!\,\delta}{|p^{(m)}(x^*)|} \right)^{1/m}. \tag{2}$$

As in [4], we remark that relation (2) is a method-independent error estimate. The accuracy of the polynomial evaluation controls the accuracy of the computed root. We discuss the practical use of this attainable accuracy in the next paragraph.

## 3.2 Newton's Iteration, Finite Precision and Stopping Criteria

Assuming $x(0)$ is close enough to the root $x^*$, we compute the Newton iteration

$$x(k+1) = x(k) - \frac{p[x(k)]}{p'[x(k)]}. \tag{3}$$

In exact arithmetic, iteration (3) converges quadratically to a single root but only geometrically with ratio $m/(m-1)$ to a root of multiplicity $m$ [10]. In finite precision, the convergence of iteration (3) may suffer from the following limitations (computed quantities have lost the hat hereafter).

- $p'[x(k)] = 0$ but $p[x(k)] \neq 0$,
- $x(k+1) = x(k)$ but $p[x(k)] \neq 0$.
- $p[x(k)] = p'[x(k)] = 0$ but $x(k)$ is not an accurate approximation of $x^*$,

Hence, an efficient stopping criterion of iteration (3) has to avoid such cases. The previously discussed attainable accuracy should be useful to define such a criterion. Alas, relation (2) is not computable as long as the multiplicity of the root remains unknown, and that is often the case.

The following criteria are well-known for terminating iteration (3).

**RE** (Relative Evolution) : $|x(k+1) - x(k)| \leq \mathbf{u}'|x(k)|$

**AR** (Absolute Residual) : $|p[x(k+1)]| \leq \sigma'$

The *a priori* relative bound $\mathbf{u}' = O(\mathbf{u})$ limits the stagnation of the iterate but is generally inefficient for ill-conditioned roots. Choosing $\sigma' = O(\sigma)$ where $\sigma$ is the smallest non-zero positive floating point number ($\sigma \ll \mathbf{u}$) would be possible if the computation of the residual $p[x(k+1)]$ is error free. Considering relation (2), $\sigma'$ should dynamically control the inaccuracy of the polynomial evaluation, choosing for example $\sigma' = \mathbf{u}\ \left(\sum_{i=0}^{n} |a_i\, x(k+1)^i|\right)$ [9].

With stochastic arithmetic, it is well-known that alternative criteria are available [13].

**SAE** (Stochastic Absolute Evolution) : $|x(k+1) - x(k)| = \underline{0}$

**SAR** (Stochastic Absolute Residual) : $p[x(k+1)] = \underline{0}$

These criteria use the specific value $\underline{0}$. We recall it describes a zero value or a non-significant computed value [12, 3]. When $x(k+1)$ satisfies the SAE test, computing $p[x(k+1)]$ is necessary to decide if $x(k+1)$ is an acceptable approximation of the root.

# 4 Experimental Results

Let us consider the computation of the double root $x^* = 3/7$ of polynomial

$$p(x) = 1.47x^3 + 1.19x^2 - 1.83x + 0.45;$$

the other root is $y^* = -5/3$ [1]. Numerical experiments are performed using IEEE-754 binary floating point arithmetic on a Sun Ultra5 workstation (SunOS 5.7, Solaris 1.3), FORTRAN 90 (WorkShop Compilers 5.0, FORTRAN 90 2.0) and CADNA library v2.2. We limit the following presentation to IEEE-754 single precision results, *i.e.*, $\mathbf{u_s} \approx 5.96 \times 10^{-8}$, excepted in the next paragraph where are also proposed double precision results, *i.e.*, $\mathbf{u_d} \approx 1.11 \times 10^{-16}$.

## 4.1 The IEEE-754 Computation

We compute Newton iteration (3) in IEEE-754 single and double precisions using Horner's scheme for the polynomial evaluations $\widehat{p}$ and $\widehat{p'}$. The two stopping criteria RE and AR are implemented respectively with $\mathbf{u}' = \mathbf{u_s}$ or $\mathbf{u_d}$ and $\sigma' = 1.0 \times 10^{-38}$. Starting with $x(0) = 0.5$, both computations yield $x(k)$ satisfying the stopping criterion AR (see next table). Other acceptable choices of $x(0)$ give similar results.

| Precision | Initial Value $x(0)$ | $k$ | $x(k)$ | $p[x(k)]$ | $p'[x(k)]$ | $\lvert x^* - x(k)\rvert / \lvert x^*\rvert$ |
|---|---|---|---|---|---|---|
| $\mathbf{u_s}$ | 0.5 | 16 | 0.4285 6118 | 0.0 | -6.31E-5 | 2.39E-5 |
| $\mathbf{u_d}$ | 0.5 | 25 | 0.4285 7143 3545 3048 | 0.0 | 3.06E-8 | 1.67E-9 |

The absolute error bound for polynomial $p$ evaluated by Horner's scheme in a root neighborhood satisfies $\delta \approx \mathbf{u}'$. Hence comparing with $x^* = 0.4285\,7142\,86$, the computed root $x(k)$ agrees the predicted accuracy given by relation (2).

## 4.2 Localizing the instability with the CESTAC Method

We use the CADNA library to compute iteration (3) with criteria SAE and SAR. Starting again with $x(0) = 0.5$, both SAE and SAR criteria are satisfied for the following value.

| Precision | Initial Value $x(0)$ | $k$ | $x(k)$ | $p[x(k)]$ | $p'[x(k)]$ |
|---|---|---|---|---|---|
| $\mathbf{u_s}$ | 0.5 | 10 | **0.4287** | $\underline{0}$ | 0.96E-3 |

The CESTAC method reduces the number of iteration and provides 3 guaranteed decimal digits (CADNA only displays significant digits and does not guarantee the last displayed decimal digit [2]). To analyze the forward error, we remark that the guaranteed digits of the CADNA result are exact comparing to $x^*$. The number of significant digits agrees the predicted accuracy.

## 4.3 Improving the Accuracy with the CENA Method

The corrected iteration (3) with the CENA method is

$$x(k+1) = x(k) - \frac{\overline{p}[x(k)]}{\overline{p'}[x(k)]}. \tag{4}$$

where $\overline{p} = \widehat{p} + \widehat{\Delta}_L(p)$ and $\overline{p'} = \widehat{p'} + \widehat{\Delta}_L(p')$ are the corrected evaluations of polynomials $p$ and $p'$. Iteration (4) is controlled with the stopping criteria RE and AR.

Using CADNA result, an optimal initial value $x(0)$ satisfies $x(0)_- \leq x(0) \leq x(0)_+$ with the IEEE-754 single precision values $x(0)_- = 0.4280\,0000$ and $x(0)_+ = 0.4289\,9999$. We compute iteration (4) using the two initial values $x(0)_-$ and $x(0)_+$; the results are the following.

| Precision | Initial Value $x(0)$ | $k$ | $x(k)$ | $p[x(k)]$ | $\lvert x(k) - x(k-1)\rvert / \lvert x(k)\rvert$ | $B_{E_C}$ |
|---|---|---|---|---|---|---|
| $\mathbf{u_s}$ | $x(0)_-$ | 7 | 0.4284 9594 | 3.70E-12 | 0.0 | 2.60E-14 |
| $\mathbf{u_s}$ | $x(0)_+$ | 6 | 0.4286 4692 | 4.83E-12 | 0.0 | 9.19E-15 |

Both iterations terminate satisfying the RE criterion. So no more accurate approximation of the root could be computed in IEEE-single precision starting with chosen $x(0)$. The two computed approximations $\overline{x}^*_- = 0.4284\,9594$ and $\overline{x}^*_+ = 0.4286\,4692$ are different after the third digit. As the $B_{E_C}$ bound value validates the 8 figures of both approximations, polynomial $\widehat{p}$ has two separate roots in the neighborhood of $x^*$ approximated by $\overline{x}^*_-$ and $\overline{x}^*_+$. Other choices of $x(0)$, $x(0)_- \leq x(0) \leq x(0)_+$, yield similar convergence to the same two finite precision limits $\overline{x}^*_-$ and $\overline{x}^*_+$.

## 4.4 The Exact Computed Solution

These experiments illustrate a well-known effect of finite precision when computing multiple polynomial roots. The double root $x^*$, *i.e.*, the exact arithmetic solution, is transformed into two separate single roots $\widehat{x}^*_-$ and $\widehat{x}^*_+$ in finite precision. Nevertheless, the convergence rate and the attainable accuracy are controled by the order 2 of the exact problem root.

IEEE-754 computation suffers from the inaccurate polynomial evaluation $\widehat{p}(x)$ and no guarantee could be given to digits after about half the precision (as relation (2) indicates). The CESTAC method localizes this instability and returns these significant digits. In both cases, the attainable accuracy bound prevents to distinguish these separate roots. Computing corrected $\overline{p}$, the CENA method improves the accuracy of the polynomial evaluation in

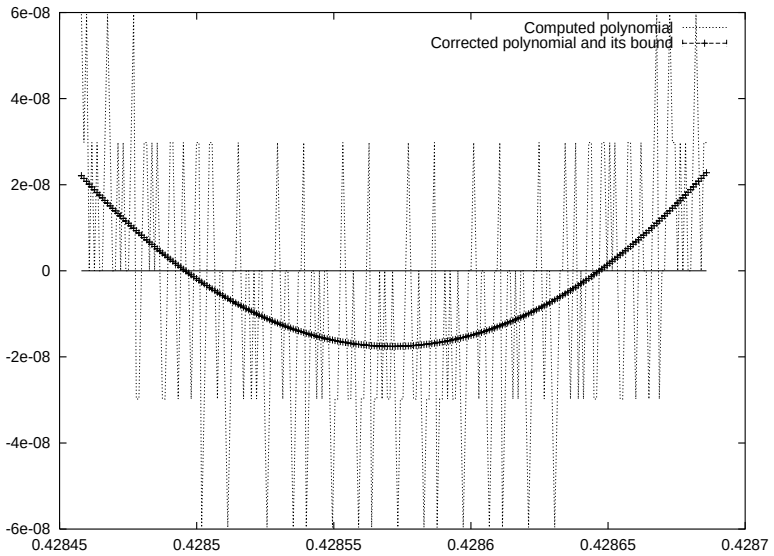


Figure 1: The two separate roots of $\widehat{p}$ appear computing $\overline{p}$.

the root neighborhood. Therefore, the absolute residual does not flush to zero and provides a sequence of iterates $x(k)$ that converges to $\overline{x}^*_-$ or $\overline{x}^*_+$ while the precision $\mathbf{u}$ is sufficient to improve the accuracy of $x(k)$, *i.e.*, while the RE criterion is not satisfied.

It follows that the splitting of the double root into two single roots does not come from the rounding errors of iteration (4). We prove that the decimal to binary conversion of the coefficients of polynomial $p$ is such that $\widehat{p}$ has two separate single roots in exact arithmetic. The following results come from symbolic computation with Maple. We represent computed polynomial $\widehat{p}$ as $\widehat{p}(x) = \widehat{a_3}x^3 + \widehat{a_2}x^2 + \widehat{a_1}x + \widehat{a_0}$, where $\widehat{a_i}$ is the exact (rational) value of IEEE-754 single precision binary value of the $a_i$ coefficient of $p$ ($0 \leq i \leq 3$). Using Sturm sequences, intervals $I_-$ and $I_+$ include the considered exact roots of $\widehat{p}$ with, limiting to the maximum 8 significant digits of IEEE-754 single precision,

$$I_- = [0.4284\,9593; 0.4284\,9594] \qquad \text{and} \qquad I_+ = [0.4286\,4691; 0.4286\,4692].$$

We come back to previous floating point computations noting that both results computed with the presented approach satisfies $\overline{x}^*_- \in I_-$ and $\overline{x}^*_+ \in I_+$. Figure (1) represents computed $\widehat{p}$, corrected $\overline{p}$ and $B_{E_C}$ bounds in $x^*$ neighborhood. It exhibits that $\overline{p}$ has two single separate roots in intervals $\widehat{I}_-$ and $\widehat{I}_+$ such that $\widehat{I}_- \subset I_-$ and $\widehat{I}_+ \subset I_+$. Floating point experiments agree the results of the symbolic computation.

# 5 Conclusion

Finite precision computation is not reliable when the problem to solve is ill-conditioned. In this case, floating point results are hard to interpret and may even yield wrong conclusions. Automatic approaches are efficient tools that complement theoretical analysis. Finite precision effects are often subtle and merged with numerous and difficult other aspects. Let us cite the effects of condition, the underlined presence of singularities, data errors, truncation errors, algorithm stability, properties of floating point arithmetic, elementary functions faithfulness, library evolutions, compilers options, woolly specifications of programming languages, ... Therefore, automatic approaches can not answer universally. It do not replace the knowledge of numerical software expert but help him to confirm his intuition.

**Acknowledgments:** The author thanks Jean-Marie Chesneaux for the CADNA library and his help during its use, Marc Daumas and Claire Finot for valuable discussions and Maple use.

# References

[1] Jean-Marie Chesneaux. *Stochastic Arithmetic and CADNA software.* Habilitation à Diriger des Recherches Thesis, Université Pierre et Marie Curie, Paris, France, November 1995. (In French).

[2] Jean-Marie Chesneaux, Stéphane Guilain, and Jean Vignes. La bibliothèque CADNA : présentation et utilisation. Manual, Université P. et M. Curie, Paris, November 1996. Available at URL = `http://www-anp.lip6.fr/cadna/`, (in French).

[3] Jean-Marie Chesneaux and Jean Vignes. Les fondements de l'arithmétique stochastique. *C. R. Acad. Sci. Paris Sér. I Math.*, 315(13):1435–1440, 1992.

[4] Germund Dahlquist and Åke Björck. *Numerical Methods.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1974. Translated by Ned Anderson.

[5] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

[6] Philippe Langlois. Automatic linear correction of rounding errors. Numerical Analysis Report 355, Manchester Centre for Computational Mathematics, Manchester, England, November 1999. Also available as INRIA Research Report RR–3828, Dec. 1999. Submitted to BIT.

[7] Philippe Langlois and Fabrice Nativel. When automatic linear correction of rounding errors is exact. *C.R. Acad. Sci. Paris, Série 1*, 328:543–548, 1999. Erratum in 328:829, 1999.

[8] John Michael McNamee. A bibliography on roots of polynomials. *J. Comput. Appl. Math.*, 47(3):391–394, 1993. (Supplementary WWW bibliography at URL = `http://www.elsevier.nl/locate/cam`.).

[9] G. Peters and J. H. Wilkinson. Practical problems arising in the solution of polynomial equations. *J. Inst. Maths Applics*, 8:16–35, 1971.

[10] Louis B. Rall. Convergence of the Newton process to multiple solutions. *Numer. Math.*, 9:25–37, 1966.

[11] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis.* Springer-Verlag, New York, second edition, 1993.

[12] Jean Vignes. Zéro mathématique et zéro informatique. *La vie des Sciences, C.R. Acad. Sci. Paris*, 4(1):1–13, 1987. (In French).

[13] Jean Vignes. A stochastic arithmetic for reliable scientific computation. *Math. and Comp. in Sim.*, 35:233–261, 1993.

[14] James H. Wilkinson. *Rounding Errors in Algebraic Processes.* Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. Also published by Prentice-Hall, Englewood Cliffs, NJ, USA. Reprinted by Dover, New York, 1994.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399