Roundoff in Polynomial Evaluation

W. Kahan
Mathematics Dept.
Univ. of Calif.
Berkeley
Nov. 18, 1986

This note discusses two ways to assess the effects of rounding errors that occur during the calculation of a polynomial and its derivative in floating-point arithmetic. One way computes bounds for those effects during the computation of the polynomial and its derivative; the second way compares the rounding errors with hypothetical perturbations in the coefficients of the polynomial. Both ways have their uses, especially when the errors in computed zeros of a polynomial have to be assessed.

Introduction:
Given the coefficients a_j of the polynomial $A(x) = \Sigma_0^N a_j x^{N-j}$, and a numerical value z, we can compute both $p := A(z)$ and the derivative $q := A'(z)$ by means of ...

Horner's recurrence:

```
q := 0 ;  p := a0 ;
for  j = 1 to N  do {  q := z q + p ;
                       p := z p + aj  } .
```

To demonstrate the validity of the recurrence, we need merely assign subscripts to the successive computed values thus:

```
q-1 := 0 ;  p0 := a0 ;  ...  and  p-1 := 0 ;
for  j = 1 to N  do {  qj-1 := z qj-2 + pj-1 ;
                       pj   := z pj-1 + aj  } .
```

Then, by substituting for a_j, we find for all x that

$$A(x) = p_N + (x-z)\left(q_{N-1} + (x-z)\Sigma_0^{N-2} q_j x^{N-2-j} \right) ,$$

whence it soon follows that the final values of p and q are $p_N = A(z)$ and $q_{N-1} = A'(z)$ respectively. But no account has yet been taken of roundoff, to which this note is devoted.

We presume that, in any arithmetic operation " $x := y \hat{o} z$ ", the value actually computed is $x = (y \hat{o} z)(1+\xi)$ where the Greek letter ξ stands for a small rounding error about which we know only that $\varepsilon > |\xi|$; here ε denotes the relative uncertainty due to roundoff in the computer's floating-point arithmetic. We must introduce similar Greek letters to stand for every rounding error committed during Horner's recurrence, after which we find that the computed values p and q actually satisfy the following

perturbed recurrence:

```
p-1 = q-1 = 0 ;  p0 = a0 ;  ... and  π0 = κ0 = 0 ;
for  j = 1 to N  {  qj-1 = ( z qj-2(1+ψj-2) + pj-1 )/(1+κj-1) ;
                    pj   = ( z pj-1(1+ζj-1) + aj )/(1+πj)  } ;
                ... and  ψN-1 = ζN = 0  .
```

At this point we may either compute an upper bound for the effect of the perturbations upon the recurrence, or we may treat those perturbations as if they were equivalent to perturbations in the coefficients instead. The first approach is technically more intricate but philosophically simpler; let's try it first.

Computed Bounds upon Roundoff's Effect:
From the perturbed recurrence, substitute for a_j in the definition of $A(x)$ to get, for all x,

$$A(x) = p_N + \Sigma_{j=0}^{N-1}(\pi_j x - \zeta_j z)p_j x^{N-1-j} + (x-z)\left(q_{N-1} + \Sigma_{j=0}^{N-2}(\kappa_j x - \psi_j z)q_j x^{N-2-j} + (x-z)\Sigma_{j=0}^{N-2} q_j x^{N-2-j} \right) ,$$

whence it soon follows that

$$A(z) = p_N + \Sigma_{j=0}^{N-1}(\pi_j - \zeta_j)p_j z^{N-j} \quad \text{and}$$

$$A'(z) = q_{N-1} + \Sigma_{j=0}^{N-2}\left((\kappa_j - \psi_j)q_j + ((N-j)\pi_j - (N-1-j)\zeta_j)p_j \right) z^{N-1-j} .$$

Since no rounding error appears more than once in each of these formulas, the nonzero Greek letters can be replaced by $\pm\varepsilon$ to get best-possible bounds for the accumulated effect of roundoff:

$$|A(z) - p_N|/\varepsilon < |p_N| + 2\Sigma_{j=1}^{N-1}|p_j| r^{N-j} + |p_0| r^N \quad \text{where } r := |z| ;$$

$$|A'(z) - q_{N-1}|/\varepsilon < |q_{N-1}| + 2\Sigma_{j=1}^{N-2}|q_j| r^{N-1-j} + |q_0| r^{N-1} + \Sigma_{j=1}^{N-1}(2N-2j-1)|p_j| r^{N-1-j} + (N-1)|p_0| r^{N-1} .$$

The right-hand sides of these inequalities are polynomials in r with coefficients derived from $|p_j|$ and $|q_j|$, so they can be computed by recurrence too. To do that, here is an

augmented recurrence:

```
r := |z| ;  q := 0 ;  p := a_0 ;  e := |p| ;  d := -e/r ;
for  j = 1 to N  do {  q := z q + p ;
                       d := r d + e + |q+q| - |p| ;
                       p := z p + a_j ;
                       e := r e + |p+p|  } ;
e := e - |p| ;  d := d - |q| .
```

Now $|A(z) - p|/\varepsilon < e$ and $|A'(z) - q|/\varepsilon < d$ except for over/-underflow and ignorable roundoff incurred during the calculation of e and d. Verifying that the last two inequalities do follow from the previous two is a challenging exercise in algebraic manipulation; that verification will confirm that the two sides of each inequality could approach each other arbitrarily closely in the event, albeit unlikely, that all the rounding errors had magnitudes ε and appropriate signs.

The augmented recurrence is most useful during the computation of a zero of $A(x)$. For instance, if z is approximately a zero of A then the error in z is approximately $-A(z)/A'(z)$, the next step of Newton's iteration. We may infer from the recurrence that $|A(z)| < |p| + e\varepsilon$ and that $|A'(z)| > |q| - d\varepsilon$, whence follows $|-A(z)/A'(z)| < (|p| + e\varepsilon)/(|q| - d\varepsilon)$ provided this is positive; otherwise roundoff so obscures A' that no such error bound for z can be estimated.

Another instance arises during an iteration to compute a zero of $A(x)$. That iteration has to be stopped when z is so close to a zero that roundoff makes further iteration probably futile. A good time to stop is when $|p| < 2e\varepsilon$; this implies that $|A(z)|$ cannot be much bigger than the roundoff that accrued during its computation. (The factor 2 is necessary to ensure that some machine-representable argument z exists so close to a zero of A that such an inequality can be satisfied. Without that factor, there would be some risk that $|p| \geq e\varepsilon$ for all arguments z, even those adjacent to a zero of A. On the other hand, the factor 2 is big enough because $e\varepsilon$ exceeds the change in $A(z)$ that would be caused by changing z to one of its neighbors.)

During iteration to find a zero, the bound $d\varepsilon$ upon the error in $A'(z)$ is not very useful because substantial errors in $A'(z)$

cause the iteration to misbehave in ways that are usually easy to recognize without d. Therefore, during the iteration, d can be omitted from the augmented recurrence, which then simplifies to a short form that runs significantly faster when N is large:

```
r := |z| ;  q := 0 ;  p := a_0 ;  e := |p|/2 ;
for j = 1 to N do {  q := z q + p ;
                     p := z p + a_j ;
                     e := r e + |p| } ;
e := e - |p| + e .
```

After the iteration has terminated and z has been accepted as an approximate zero, running the previous augmented recurrence once provides an *estimate* $(|p|+e\varepsilon)/(|q|-d\varepsilon)$ of a bound upon the error in z. That estimate is *not* a rigorous bound; it was based upon an estimate $-A(z)/A'(z)$ that could *underestimate* the error in z. How badly? At worst by a factor $1/N$ according to

Laguerre's Theorem: The polynomial $A(x)$ of degree N must have a zero ζ satisfying $|\zeta - z + A(z)/A'(z)| \le (N-1)|A(z)/A'(z)|$.

Proof: We know that $A(x) = a_0(x-x_1)(x-x_2)(\ \dots\ )(x-x_{N-1})(x-x_N)$, where $x_1, x_2, \dots, x_{N-1}, x_N$ are the (unknown) zeros, real and complex, of A. Let ζ be whichever of them is closest to z. Since $\zeta - z + A(z)/A'(z) = (1 - \Sigma_j (z-\zeta)/(z-x_j))\ A(z)/A'(z)$, cancel " 1 " and take magnitudes to deduce the theorem.

Therefore the error in an approximate zero z cannot exceed

$$N\,(|p|+e\varepsilon)/(|q|-d\varepsilon) .$$

This error bound is rigorous but, like Laguerre's theorem, it is usually pessimistic by a factor near N which may be annoying if N is very big. A rigorous error bound that is usually far less pessimistic costs slightly more computation and much more thought.

The simplest thoughts arise when z approximates a real zero of a real polynomial $A(x)$. The sign of $A(z)$ is the same as that of $p+e\varepsilon$ and $p-e\varepsilon$ when they have the same signs; otherwise the sign of $A(z)$ is obscured by roundoff, as it usually would be when z is the best available approximation to a zero of $A(x)$. Now let ξ be any approximate bound for the error in z; for instance, try $\xi := (|p|+e\varepsilon)/|q|$. We can check whether ξ truly bounds the error in z by running the short form of the augmented recurrence twice to see whether the signs of $A(z-\xi)$ and $A(z+\xi)$ are opposite, taking roundoff into account. Usually those signs do differ, and then we know for sure that $A(x)$ vanishes between $z-\xi$ and $z+\xi$. Otherwise accept what Laguerre's theorem tells us.

When complex zeros of a polynomial are being computed, error bounds better than are provided by Laguerre's theorem require a lot of thought. Because complex variables lie beyond the syllabus of this course, only a rough outline of those thoughts will be sketched here. They begin with the divided difference

$$\begin{aligned} \Delta f(\{x, y\}) &:= (f(x) - f(y))/(x - y) \quad \text{if } x \ne y , \\ &:= f'(y) \quad \text{if } x = y . \end{aligned}$$

It resembles the derivative in many ways besides their similar definitions; they also figure in similar estimates for the zeros of a function. For instance, ...

Lemma: Let z be fixed at the center of some region $|x-z| \le \delta$ throughout which $f(x)$ is continuously differentiable, and suppose also that $|\Delta f(\{x, z\})| > |f(z)|/\delta$ at every x therein. Then $f(x)$ must vanish *at least* once inside that region. (It may be an interval on the real axis or a disk in the complex plane. And if also $|\Delta f(\{x, y\})| > 0$ at all x and y in the region, $f(x)$ vanishes *just* once therein.)

Proof: Construct the map $\Phi(x) := x - f(x)/\Delta f(\{x, z\})$. This map is inspired by Newton's and the Secant iterations. Since the divisor cannot vanish, Φ must be continuous throughout the region. Moreover, $\Phi(x)-z = -f(z)/\Delta f(\{x, z\})$, which implies $|\Phi(x)-z| < \delta$ throughout the region, which means that Φ is a continuous map of this closed bounded convex region into itself. By Brouwer's fixed-point theorem, Φ must have at least one fixed point $x = \Phi(x)$ in the region; that is where $f(x) = 0$.

The derivative and the divided difference of our polynomial $A(x)$ share another property; they can be computed without any division from a revised version of Horner's recurrence as follows:

$$Q := 0 \ ; \quad p := a_0 \ ;$$
$$\text{for } j = 1 \text{ to } N \text{ do } \{ \ Q := y\,Q + p \ ;$$
$$p := z\,p + a_j \ \} \ .$$

The final values of $p = A(z)$ and $Q = \Delta A(\{y, z\})$ would be correct but for rounding errors which shall be ignored here to simplify the exposition. How does Q vary as y runs about some tiny region containing a zero of $A(x)$ close to z ? If $y = z$, $Q = q = Q'(z)$ as computed by Horner's recurrence. Otherwise, using the subscripted values p_j we introduced to explain that recurrence, we can infer that

$$Q - q := (y-z)\,\Sigma_{0 \le j \le N-2} \, p_j S_{N-1-j} \quad \text{where}$$
$$S_k := (z^k-y^k)/(z-y) = \Sigma_{0 \le j \le k-1} \, z^j \, y^{k-1-j} \ .$$

If we can find some $s > \max\{ |z|, |y| \}$, then $|S_k| < k\,s^{k-1}$ and

$$|Q-q| < |y-z|\,\Sigma_{0 \le j \le N-2} \, (N-1-j)|p_j|s^{N-2-j} \ ,$$
$$= |y-z|\,R(s)$$

where $R(s)$ in the last inequality is a polynomial in s that can be computed by another augmented recurrence similar to the one that computes d above. To determine s , we choose any known error bound $\xi := N\,(|p|+e\varepsilon)/(|q|-d\varepsilon)$, say, and assume that $|y-z| < \xi$; then $s := |z| + \xi$. Then we run another augmented recurrence to evaluate $R(s)$, and deduce that for all $|y-z| < \xi$

$$|\Delta A(\{y,z\})| = |q + (Q-q)| \ge |q| - |Q-q| > |q| - \xi\,R(s)$$

to within a small computable allowance for roundoff. If the last expression exceeds $(|p|+e\varepsilon)/\xi$, as it usually does, then the Lemma tells us that $A(x)$ vanishes somewhere in the region

$$|x-z| < (|p|+e\varepsilon)/(|q| - \xi\,R(s)) \ ,$$

which is usually much tinier than ξ .

A number of details have been omitted from the foregoing account because they lie beyond the scope of this course. The important conclusions to be drawn from what has been presented so far are that we can compute a zero of a polynomial as accurately as we like if we carry enough precision, and that we can prove our result correct taking roundoff into account with complete rigor. Although it is possible to compute better error bounds that come ever closer to the limits of uncertainty imposed by roundoff, such bounds are hardly ever worth their cost because other sources of uncertainty so often predominate over roundoff.

Interpreting Roundoff as Perturbations in Data:
Let us return to the perturbed recurrence and express p_j in terms of the rounding errors (the Greek letters) and the given coefficients a_j. Instead of computing $A(z) := \sum_{0}^{N} a_j z^{N-j}$, we can prove by induction that actually $p_N = A(z) := \sum_{0}^{N} a_j z^{N-j}$, where

$$\begin{aligned} a_j &= a_j(1+\zeta_j)(1+\zeta_{j+1})(\ldots)(1+\zeta_{N-1})/((1+\pi_j)(1+\pi_{j+1})(\ldots)(1+\pi_N)) \\ &= a_j(1\pm\varepsilon)^{\pm(2N-2j+1)} \quad \text{if } j \neq 0, \text{ otherwise } a_0(1\pm\varepsilon)^{\pm 2N}. \end{aligned}$$

In other words, the computed value p, obtained instead of the desired value $A(z)$, is exactly what would have been computed without roundoff if each coefficient a_j had first been perturbed to a nearly indistinguishable number A_j. This is the sense in which roundoff committed during the computation of $A(z)$ is no worse than a few rounding errors per coefficient committed before that computation. If the given coefficients are uncorrelatedly uncertain by as much as $2N$ units in their last significant digits, then that uncertainty in $A(z)$ will dominate whatever uncertainty subsequently accrues to p because of roundoff. This view of the rounding errors is called a "*Backward Error-Analysis*".

Let us reconsider the computation of a zero of $A(x)$ from this point of view. We shall accept z as a purported approximation to a zero of A when p, the computed value of $A(z)$, is deemed negligible; but that will actually mean that the perturbed polynomial $p = A(z)$ is negligible, so z will actually lie close to a zero of the perturbed polynomial A. Because the zeros of a polynomial are known to be continuous functions of its coefficients, and because the coefficients of A and A are so nearly indistinguishable, one might hope their zeros are almost the same too. However, some closely neighboring polynomials A and A have surprisingly different zeros. For example, let

$$\begin{aligned} A(x) &:= x^{12} - 12x^{11} + 66x^{10} - 220x^9 + 465x^8 - 792x^7 + 924x^6 - 792x^5 + 495x^4 - 220x^3 + 66x^2 - 12x + 1 \\ &= (x - 1)^{12}, \end{aligned}$$

and let $A(x) := A(x) - x^6/10^6$. In other words, the perturbed polynomial $A(x)$ is obtained from $A(x)$ by replacing $924x^6$ by $923.999999x^6$, a change in the ninth significant decimal. Then, although all twelve zeros of A are at $z = 1$, two of the zeros of A are at $z = 0.729843788$ and $z = 1.370156212$. (These are the zeros of $(x-1)^2 - x/10$, which you should confirm as a divisor of $A(x)$.) A similar but slightly more extreme example is constructed from the same $A(x) := (x-1)^{12}$, but now perturb it to $A(x) := A(x) - A(-x)/5_{10}9$, changing each coefficient in its tenth sig. dec. or beyond. For instance, $12x$ gets changed into $12.0000000024x$, and $924x^6$ into $923.9999998152x^6$. You should be able easily to compute the two real zeros $z = 1.36828744$ and $z = 0.73084059$ of $A(x)$.

The foregoing two examples may give the false impression that the zeros of a polynomial can be hypersensitive to tiny perturbations in its coefficients only if the zeros are repeated. Actually the truth is slightly but crucially different from that. Zeros can be hypersensitive to tiny perturbations in coefficients only if such tiny perturbations could cause the zeros in question to change their multiplicities. An explanation and proof of this assertion would be too complicated to include in this note; instead, the assertion will be illustrated by an example. This example is similar to one discovered in the late 1950's by James H. Wilkinson and used for the same purpose. Let

$$A(x) := x^{12} - 78x^{11} + 2717x^{10} - 55770x^9 + 749463x^8 - 6926634x^7 + \\ + 44990231x^6 - 206070150x^5 + 657206836x^4 - 1414014888x^3 + \\ + 1931559552x^2 - 1486442880x + 479001600$$
$$= (x-1)(x-2)(x-3)(\ldots)(x-10)(x-11)(x-12)$$
$$= \Gamma(x)/\Gamma(x-12) = (x-1)!/(x-13)! \ .$$

Its zeros, the consecutive integers from 1 to 12, do not seem especially close together, but in fact an extremely tiny change to its coefficients can change the zeros enough to make two of them coalesce. Specifically, $A(x) := A(x) - \lambda A(-x)$ has a double zero at $z = 8.4835138$ when $\lambda = 5.600278_{10}\text{-}10$. Although the zeros of A are hypersensitive to roundoff, these numbers z and λ can be calculated easily on a programmable calculator by means interesting enough to merit inclusion in this note.

The double zero z of $A(x)$ must satisfy $A(z) = A'(z) = 0$; that means $A(z) - \lambda A(-z) = A'(z) + \lambda A'(-z) = 0$. Eliminating λ produces an equation, $A'(z)/A(z) + A'(-z)/A(-z) = 0$, that identifies z as one of the 22 finite zeros of

$$A'(x)/A(x) + A'(-x)/A(-x) = \Sigma_{j=1}^{12} \left(1/(x-j) - 1/(x+j) \right) .$$

Every such zero lies between two consecutive nonzero integers between -12 and 12, so it is easy to compute accurately by Newton's or Secant iteration. Each such zero z determines a corresponding $\lambda = A(z)/A(-z) = \Gamma(z)\,\Gamma(z+1)/(\Gamma(z-12)\,\Gamma(z+13))$. The smallest λ and its z are the ones exhibited above.

The foregoing examples warn us that some polynomials are so hypersensitive to roundoff in their coefficients that their zeros cannot be determined without carrying extravagant precision that may be unjustified if the coefficients are intrinsically uncertain by as little as a few rounding errors. In such cases, we should try to find out where the polynomial came from; it may have come from a problem that determined its roots quite accurately until it was transformed into an explicit polynomial equation. For example $A(x) = (1-\lambda)x^{12} - 78(1+\lambda)x^{11} + 2717(1-\lambda)x^{10} - \ldots + 479001600(1-\lambda)$ has the same zeros as the expression

$$\Gamma(x)\,\Gamma(x+1)/\big(\Gamma(x-12)\,\Gamma(x+13)\big) - \lambda \ ,$$

but, for any fixed tiny value of λ, at least about $-\log(|\lambda|)$ more sig. dec. must be carried when the polynomial $A(x)$ is used than when the latter expression is used to calculate those zeros to any preassigned accuracy, as we shall see.

Condition Numbers:
How sensitive can a zero z of some function $f(x)$ be to small perturbations Δf in f? *Condition numbers* answer questions like this in a quantitative way. Before we compute a condition number, we must choose a *norm* $\|\ldots\|$ to measure the size of small perturbations; ideally $\|\Delta f\|$ takes the same value for all perturbations Δf that are about equally likely or about equally (in)significant. There is no reason why $\|\Delta f\|$ should not depend upon z; hence, one such norm for perturbations

$$\Delta A(x) = \Sigma_{0}^{N} \Delta a_j \, x^{N-j}$$

in the polynomial $A(x)$ might be

$$\|\Delta A\| := \Sigma_{0}^{N} |\Delta a_j \, z^{N-j}| \ ,$$

which serves as an upper bound for $|\Delta A(z)|$ (though not for $|\Delta A(x)|$ when $|x| > |z|$). Given a norm for Δf, a *relative* condition number for the infinitesimal changes Δz caused by infinitesimal perturbations Δf in f would be

$$\kappa(z, f, \|\dots\|) := \max_{\Delta f \neq 0} |\Delta z/z|/(\|\Delta f\|/\|f\|)$$

where the maximum is taken over all infinitesimal nonzero Δf and Δz is obtained from a root of $(f+\Delta f)(z+\Delta z) = 0$ closest to the root z of $f(z) = 0$. Consequently, $\Delta z = -\Delta f(z)/f'(z)$ if $f'(z) \neq 0$ and terms of order $(\Delta z)^2$ are ignored, and then

$$\kappa(z, f, \|\dots\|) = \left(\max_{\Delta f \neq 0} |\Delta f(z)|/\|\Delta f\| \right) \|f\|/|z\, f'(z)| .$$

The maximum in parentheses depends upon z and $\|\dots\|$ but not upon f, and turns out to be $\|e(z)^*\|$ where $e(x)^*$ is called "the evaluation functional" because it extracts a value $f(x) = e(x)^* f$ from any function f. For the example $\|\Delta A\|$ above,

$$\|e(z)^*\| = \max_{\Delta A \neq 0} |\Delta A(z)|/\|\Delta A\| = 1 , \text{ so}$$

$$\kappa(z, A, \|\dots\|) = \|A\|/|z\, A'(z)| .$$

This condition number κ roughly bounds the magnification ratio $(|\Delta z|/|z|)/(\|\Delta A\|/\|A\|)$ for relatively tiny perturbations ΔA.

A similar analysis applies to equations $f(x) = \lambda$ when $f(x)$ can be computed correct to nearly as many sig. dec. as are carried by the arithmetic. The appropriate norm then is actually a semi-norm $\|\Delta f\| = |\Delta f(z)|$, and the appropriate relative condition number is

$$\kappa(z, f-\lambda, \|\dots\|) := \max_{\Delta f} (|\Delta z|/|z|)/(\|\Delta f\|/\|f\|) = |\lambda|/|z\, f'(z)| .$$

Let us compare condition numbers for the foregoing examples

$$A(x) := (1-\lambda)x^{12} - 78(1+\lambda)x^{11} + 2717(1-\lambda)x^{10} - \dots + 479001600(1-\lambda)$$

and

$$f(x) := \Gamma(x)\,\Gamma(x+1)/\big(\Gamma(x-12)\,\Gamma(x+13)\big) = A(x)/\big(\Gamma(x+13)/\Gamma(x+1)\big) + \lambda .$$

Since $\|A\|/\big(\Gamma(z+13)/\Gamma(z+1)\big)$ lies between $1\pm\lambda$ we soon find that

$$\kappa(z, A, \|\dots\|)/\kappa(z, f-\lambda, \|\dots\|) = (1\pm\lambda)/\lambda ,$$

which vindicates the earlier claim that about $-\log_{10}(|\lambda|)$ more sig. dec. must be carried to solve $A(z) = 0$ than to solve $f(z) = \lambda$ equally accurately. And when $\lambda = 0$ the condition number of the integer $z = 1, 2, \dots, \text{or } 12$ as a zero of A is

$$\kappa(z, A, \|\dots\|) := (z+12)!/((12-z)!(z!)^2) = 64664600 \text{ when } z = 9 ,$$

which explains a loss of 7 sig. dec. when that zero is computed.

The foregoing condition numbers κ pertain only to simple zeros z; condition numbers for multiple zeros lie beyond this course.

Acknowledgements:
Although this note was prepared for an introductory class in Numerical Analysis, the computable error bounds are based upon researches supported at times by the U. S. Office of Naval Research under contract number N00014-76-C-0013, and by the Air Force Office of Scientific Research under AFOSR-84-0158.