Applied Numerical Linear Algebra

Round-off analysis in polynomial evaluation

In this section we will discuss stability of polynomial evaluation by Horner's rule. Let the polynomial is given by

$$p(x) = \sum_{i=0}^{d} c_i x^i,$$

where c_i are coefficients of the polynomial, d is its degree.
To compute roots of this polynomial we can use Horner's rule described in Chapter 1. This rule can be programmed as the following iterative algorithm for every mesh point $x_j \in [x_{left}, x_{right}], j \in 1, 2, ...N$, where N is the total number of the discretization points:

Horner's rule for polynomial evaluation

- Step 0. Initialize $p_d = c_d$. Set counter $i = d - 1$.
- Step 1. Compute $p_i = x_j \cdot p_{i+1} + c_i$
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.

To compute bounds in the polynomial evaluation we insert term with error $1 + (\sigma_{1,2)_i}$ for every floating point iteration in Algorithm 2 to obtain following algorithm:

Error bound in polynomial evaluation

- Step 0. Set counter $i = d - 1$ and initialize $p_d = c_d$.
- Step 1. Compute
  $p_i = (x_j \cdot p_{i+1}(1 + (\sigma_1)_i) + c_i)(1 + (\sigma_2)_i), \quad |(\sigma_1)_i|, |(\sigma_2)_i| \leq \varepsilon$
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.

In the algorithm 3 the number ε is the machine epsilon and we define it as the maximum relative representation error $0.5 \cdot \beta^{1-p}$ which is measured in a floating point arithmetic with the base β and with precision $p > 0$. Now the following values of machine epsilon apply to standard floating point formats:

Expanding expression for p_i in the algorithm 3 we can get

$$p_0 = \sum_{i=0}^{d-1} \left((1 + (\sigma_2)_i) \prod_{k=0}^{i-1} (1 + (\sigma_1)_k)(1 + (\sigma_2)_k) \right) c_i x^i \tag{1}$$

$$+ \left(\prod_{k=0}^{d-1} (1 + (\sigma_1)_k)(1 + (\sigma_2)_k) \right) c_d x^d$$

Next, we will write upper and lower bounds for products of $\sigma := \sigma_{1,2}$ provided that $k\varepsilon < 1$:

$$(1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \leq (1 + \varepsilon)^k \leq 1 + k\varepsilon + O(\varepsilon^2),$$

$$1 - k\varepsilon \leq (1 - \varepsilon)^k \leq (1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \tag{2}$$

Applying estimate above we can get the following inequality

$$1 - k\varepsilon \leq (1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \leq 1 + k\varepsilon. \tag{3}$$

Using the estimate (3) we can rewrite (1) as

$$p_0 \approx \sum_{i=0}^{d} (1 + \tilde{\sigma}_i) c_i x^i = \sum_{i=0}^{d} \tilde{c}_i x^i \tag{4}$$

with approximate coefficients $\tilde{c}_i = (1 + \tilde{\sigma}_i) c_i$ such that $|\tilde{\sigma}_i| \leq 2k\varepsilon \leq 2d\varepsilon$.

Now we can write formula for the computing error in the polynomial:

$$|p_0 - p(x)| = \left| \sum_{i=0}^{d} (1 + \tilde{\sigma}_i) c_i x^i - \sum_{i=0}^{d} c_i x^i \right| \tag{5}$$

$$= \left| \sum_{i=0}^{d} \tilde{\sigma}_i c_i x^i \right| \leq 2 \sum_{i=0}^{d} d\varepsilon |c_i x^i| \leq 2d\varepsilon \sum_{i=0}^{d} |c_i x^i|.$$

If we will choose $\tilde{\sigma}_i = \varepsilon \cdot \mathrm{sign}(\mathrm{c_i x^i})$ then the error bound above can be attained within the factor $2d$. In this case we can take

$$bp_{rel} = \frac{\sum_{i=0}^{d} |c_i x^i|}{|\sum_{i=0}^{d} c_i x^i|} \tag{6}$$

as the relative condition number for the case of polynomial evaluation. This condition number can be computed at every point x_j for $[p - bp_{rel}, p + bp_{rel}]$. In the following algorithm we use (5) to compute lower bound in polynomial evaluation.

Computation of the error bp in the polynomial evaluation

- Step 0. Set counter $i = d - 1$ and initialize $p_d = c_d$, $bp_d = |c_d|$.
- Step 1. Compute $p_i = x_j \cdot p_{i+1} + c_i$, $bp_i = |x_j| \cdot bp_i + |c_i|$.
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.
- Step 3. Set $bp = 2 \cdot d \cdot \varepsilon \cdot bp_i$ as error bound at the point $|x_j|$.

Figures 1-a), 2-a) show behavior of the computed solution using Horner's rule (algorithm 2) for the evaluation of roots of polynomial $p(x) = (x-9)^9 = x^9 - 81x^8 + 2916x^7 - 61236x^6 + 826686x^5 - 7440174x^4 + 44641044x^3 - 172186884x^2 + 387420489x^1 - 387420489$. Figures 1-b), 2-b) show computed upper and lower bounds for the polynomial $p(x) = (x-9)^9$ using algorithm 6 on different input intervals for x. We have performed all our computations taking in algorithm 6 $\varepsilon = 0.5 \cdot \beta^{1-p}$. Using these figures we observe that changing the input interval for x slightly can change computed roots drastically.

Number of the correct decimal digits computed by algorithm 6 is given by the formula $e_n = -ln\left|\frac{bp}{p}\right|$. This number is plotted on Figures 1-d), 2-d) in blue color for different input intervals for x. In red color we present the computed relative error by the formula $e_{comp} = -ln\left|\frac{p-(x-9)^9}{p}\right|$. We observe that our estimated lower bound in blue color is quite good to the computed relative error e_{comp}. Figures 1-c), 2-c) show computed estimated relative errors $e = \left|\frac{bp}{p}\right|$ on different input intervals for x which correspond to Figures 1-d), 2-d). Analyzing Figures 1-c),d), 2-c), d) we can conclude that we get difficulties when we want to compute $p(x)$ with a high relative accuracy when $p(x)$ is close to zero. This is because any small changes in ε gives infinite relative error given by $\frac{\varepsilon}{p(x)} = \frac{\varepsilon}{0}$ what means that our relative condition number $2d \cdot \varepsilon \cdot bp$ is infinite. Thus, the problem of finding of relative condition number of the polynomial is ill-posed.

Definition

Let $p(x) = \sum_{i=0}^{d} a_i x^i$ and $q(x) = \sum_{i=0}^{d} b_i x^i$ are two polynomials. Then the relative distance $dist(p, q)$ from $p(x)$ to $q(x)$ is defined as the smallest value of $dist(p, q)$ such that

$$|a_i - b_i| \leq dist(p, q) \cdot |a_i|, \ i \leq 1 \leq d.$$

If $a_i \neq 0, \ i \leq 1 \leq d$ the equation above can be rewritten as

$$\max_{0 \leq i \leq d} \frac{|a_i - b_i|}{|a_i|} = dist(p, q), \ i \leq 1 \leq d.$$

The next theorem says that the distance from p to q (the distance to the nearest ill-posed problem) is the reciprocal of the condition number of $p(x)$.

Theorem

Let polynomial $p(x) = \sum_{i=0}^{d} c_i x^i$ *is not identically zero and* $q(x) = 0$ *is another polynomial whose condition number at* x *is infinite. Then*

$$\min\{dist(p, q) : q(x) = 0\} = \frac{|\sum_{i=0}^{d} c_i x^i|}{\sum_{i=0}^{d} |c_i x^i|}. \tag{7}$$

Proof.

To prove this theorem let us write $q(x) = \sum_{i=0}^{d} b_i x^i = \sum_{i=0}^{d} (1+\varepsilon_i) c_i x^i$ such that $dist(p,q) = \max_i |\varepsilon_i|$. Then $q(x) = 0$ implies that

$$p(x) = |q(x) - p(x)| = |\sum_{i=0}^{d} \varepsilon_i c_i x^i| \leq \sum_{i=0}^{d} |\varepsilon_i c_i x^i| \leq \max_i |\varepsilon_i| \sum_{i=0}^{d} |c_i x^i|.$$

Thus,

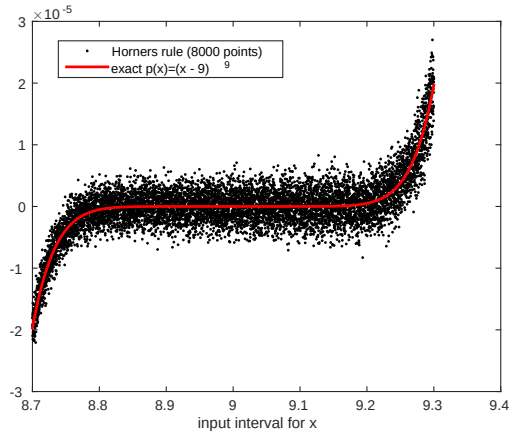
$$dist(p,q) = \max_i |\varepsilon_i| \geq \frac{|p(x)|}{\sum_{i=0}^{d} |c_i x^i|}.$$

For example, we can choose $\varepsilon_i = \frac{-p(x)}{\sum_{i=0}^{d} |c_i x^i|} \cdot sign(c_i x^i)$. □

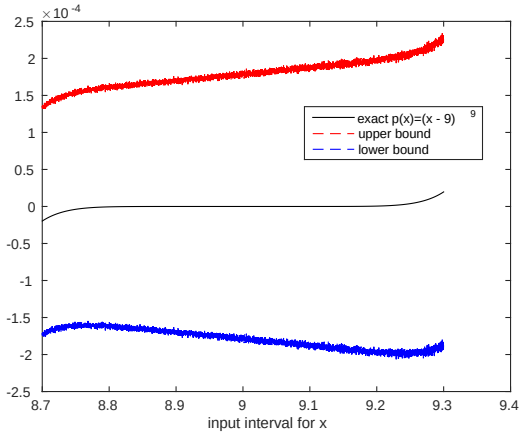
At the end of this subsection we will present the bisection algorithm to find roots of the polynomial $p(x) = 0$. Suppose that the input interval for x where we want to find roots of $p(x) = 0$ is $x \in [x_{left}, x_{right}]$. At every iteration this algorithm divides the input interval in two by computing the midpoint $x_{middle} = (x_{left} + x_{right})/2$ of the input interval as well as the value of the polynomial $p(x_{middle})$ at that point. Value of the polynomial $p(x_{middle})$ we will compute using Horner's rule (algorithm 2). Then if p_{left} and p_{mid} have opposite signs, then the bisection algorithm sets x_{middle} as the new value for x_{right}, and if p_{right} and p_{mid} have opposite signs then the method sets x_{middle} as the new x_{left}. If $p(x_{middle}) = 0$ then x_{middle} may be taken as the root of polynomial and algorithm stops.

Bisection algorithm to find zeros of polynomial $p(x)$

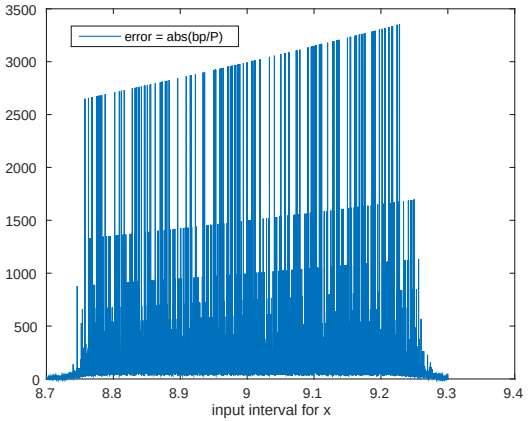
- Step 0. Initialization: set left x_{left} and right x_{right} bounds for input interval for $x \in [x_{left}, x_{right}]$ where we will seek roots of polynomial. Set computational tolerance tol.
- Step 1. Evaluate polynomial $p(x)$ at x_{left} and x_{right} to get $p_{left} = p(x_{left})$ and $p_{right} = p(x_{right})$ using algorithm 2.
  Perform steps 2-3 while $x_{right} - x_{left} > 2 \cdot tol$
- Step 2. Compute point $x_{mid} = \frac{(x_{left} + x_{right})}{2}$ and then $p_{mid} = p(x_{mid})$ using algorithm 2.
- Step 3. Check:
  If $p_{left} \cdot p_{mid} < 0$ then we have a root at the interval $[x_{left}, x_{mid}]$. Assign $x_{right} = x_{mid}$ and $p_{right} = p_{mid}$.
  Else if $p_{right} \cdot p_{mid} < 0$ then we have a root at the interval $[x_{mid}, x_{right}]$. Assign $x_{left} = x_{mid}$ and $p_{left} = p_{mid}$.
  Else we have found a root at x_{mid} and assign $x_{left} = x_{mid}, x_{right} = x_{mid}$.
- Step 4. Compute root as $\frac{(x_{left} + x_{right})}{2}$.



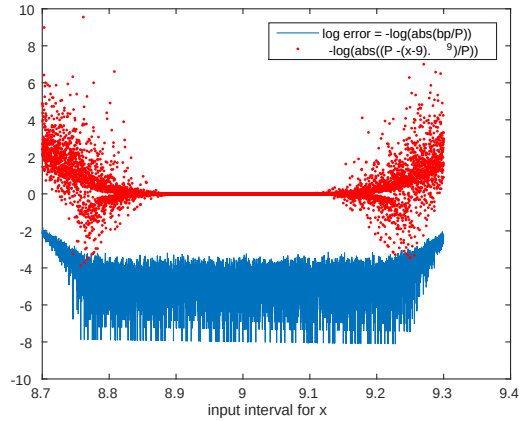
a)



b)

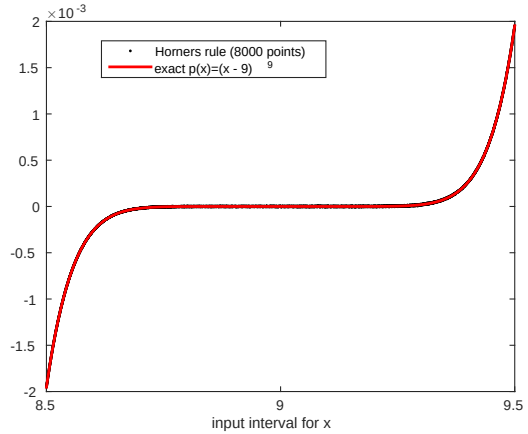


c)

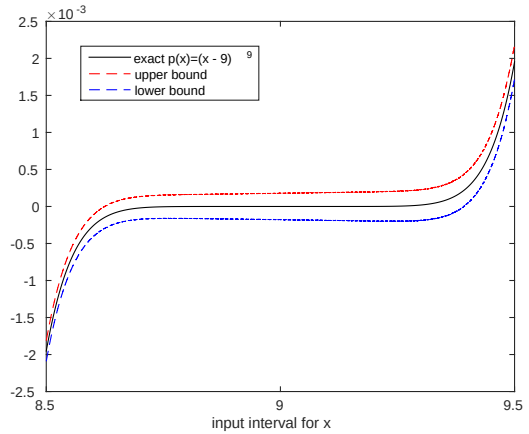


d)

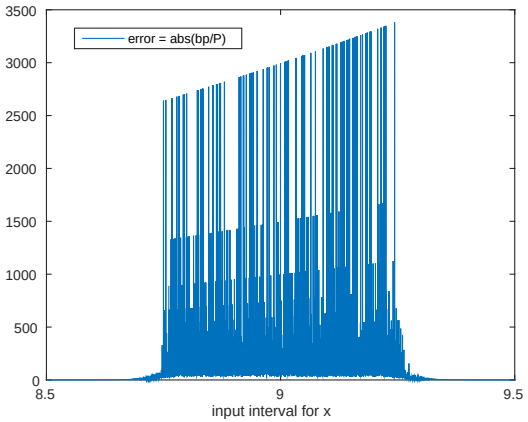
Figure: *a) Evaluation of the polynomial* $p(x) = (x-9)^9$ *by Horner's rule (algorithm 2) compared with the exact one polynomial. b) Computed upper and lower bounds for the polynomial* $p(x) = (x-9)^9$ *using algorithm 6. c) Plot of the graph of the estimated relative error* $e = \left|\frac{bp}{p}\right|$*. d) Plot of the graph of the estimated relative error* $e_{ln} = -ln\left|\frac{bp}{p}\right|$ *(presented in blue color) compared with the computed relative error* $e_{comp} = -ln\left|\frac{p-(x-9)^9}{p}\right|$ *(presented in red color). Input interval for* x *in this example is* $x \in [8.7, 9.3]$.



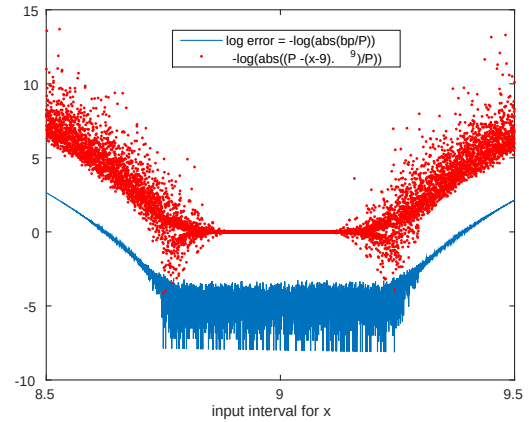
a)



b)

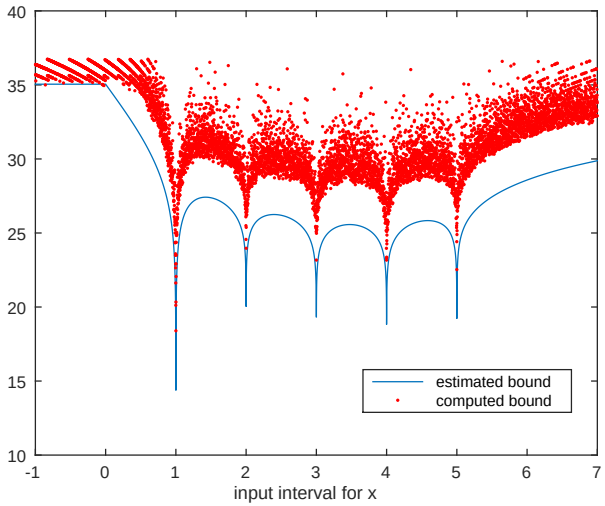


c)

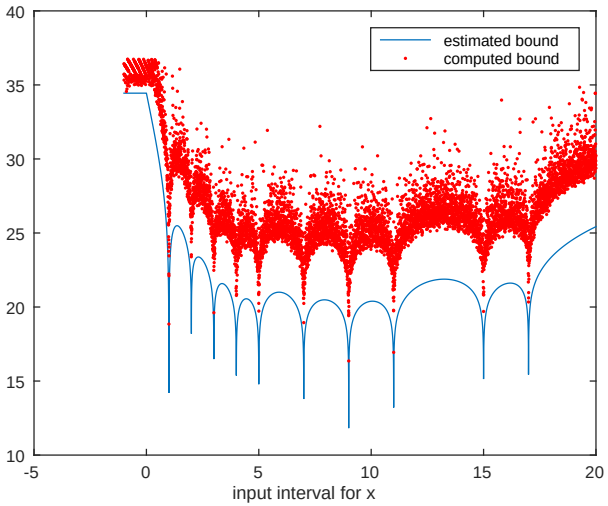


d)

Figure: *a) Evaluation of the polynomial* $p(x) = (x-9)^9$ *by Horner's rule (algorithm 2) compared with the exact one polynomial. b) Computed upper and lower bounds for the polynomial* $p(x) = (x-9)^9$ *using algorithm 6. c) Plot of the graph of the estimated relative error* $e = \left|\frac{bp}{p}\right|$*. d) Plot of the graph of the estimated relative error* $e_{ln} = -ln\left|\frac{bp}{p}\right|$ *(presented in blue color) compared with the computed relative error* $e_{comp} = -ln\left|\frac{p-(x-9)^9}{p}\right|$ *(presented in red color). Input interval for* x *in this example is* $x \in [8.5, 9.5]$.



a)



b)

Figure: *Plot of the graph of the estimated relative error* $e_{ln} = -ln\left|\frac{bp}{p}\right|$ *(presented in blue color) compared with the computed relative error* e_{comp}*: a) for the polynomial* $p(x) = (x-1)^2(x-2)(x-3)(x-4)(x-5)$ *and b) for the polynomial* $p(x) = (x-1)^2(x-2)(x-3)(x-4)(x-5)(x-7)(x-9)(x-11)(x-15)(x-17)$.

Perturbation theory in the solution of linear equations

Let us consider numerical solution of linear system $Ax = b$. Here, A, x, b are exact matrix, vector of solution and the right hand side of the equation, correspondingly. We are interested to find bound δx for the error in the computed solution $\tilde{x}$ which can be computed as $\delta x = \tilde{x} - x$. We assume that matrix A have a small error δA and the right hand side b is given with an error δb. More precisely, we have

$$(A + \delta A)\tilde{x} = b + \delta b. \tag{8}$$

Subtracting $Ax = b$ from (8) we get

$$(A + \delta A)\tilde{x} - Ax = b + \delta b - b.$$

Rearranging terms in the above equation and noting that $\delta x = \tilde{x} - x$ and thus $x = \tilde{x} - \delta x$ we have

$$\begin{aligned}
(A + \delta A)\tilde{x} - Ax - \delta b = (A + \delta A)(x + \delta x) - Ax - \delta b = 0, \\
Ax + A\delta x + \delta Ax + \delta A\delta x - Ax - \delta b = (A + \delta A)\delta x + \delta Ax - \delta b = 0, \\
(A + \delta A)\delta x + \delta A(\tilde{x} - \delta x) - \delta b = 0, \\
A\delta x + \delta A\tilde{x} - \delta b = 0,
\end{aligned} \tag{9}$$

resulting in the equation

$$\delta x = A^{-1}(-\delta A \cdot \tilde{x} + \delta b). \tag{10}$$

Taking norms in the equation above and using triangle inequality leads us to

$$\|\delta x\| \leq \|A^{-1}\|(\|\delta A\| \cdot \|\tilde{x}\| + \|\delta b\|).$$

Dividing this inequality to $\|\tilde{x}\|$ and compensating by $\|A\|/\|A\|$ gives us

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|\tilde{x}\|} \right), \tag{11}$$

where $k(A) = \|A^{-1}\| \cdot \|A\|$ is called *the condition number of the matrix A*.