1.8 bounding the error of expressions

modelling expressions with simple bivariate functions

- let a set of integers $i_1, \ldots, i_n$ and $j_1, \ldots, j_n$ satisfy
  - either $i_k = j_k = 0$
  - or $j_k < i_k < k$
- let $f_1, \ldots, f_n$ be bivariate real functions defined on compact domains
  - the functions f_k are either arithmetic binary operations or univariate functions
- let $u_0 = 0$ and u_k be defined by the system of equations

$$u_k = f_k(u_{i_k}, u_{j_k}), \quad k = 1, \ldots, n$$

evaluation of the expression

- these equations are thus solved (i.e. all u_k computed) by substitution

$$
\begin{aligned}
u_1 &= f_1(u_0, u_0) = f_1(0, 0) \\
u_2 &= f_2(u_{i_2}, u_0) = f_2(u_{i_2}, 0), \quad i_2 \in \{0, 1\} \\
u_3 &= f_3(u_{i_3}, u_{j_3}), \quad i_3 \in \{0, 1, 2\},\ j_3 \in \{0, \ldots, i_3\} \\
&\ldots \\
u_n &= f_n(u_{i_n}, u_{j_n}), \quad i_n \in \{0, \ldots, n-1\},\ j_3 \in \{0, \ldots, i_n\}
\end{aligned}
$$

- with this we have modeled the evaluation of numerical expressions where u_n is the value of the expression and the other u_k intermediate results

example $\left(-p+\sqrt{p^2-4\,q}\right)/2$

$$
\begin{aligned}
u_1 &= p \\
u_2 &= q \\
u_3 &= u_1^2 \\
u_4 &= u_3 - 4\,u_2 \\
u_5 &= \sqrt{u_4} \\
u_6 &= (-u_1 + u_5)\,/2
\end{aligned}
$$

the same with rounding errors at every step

- now let v_k be the numerical versions of u_k defined by

$$v_k = (1 + \delta_k)\, f_k(v_{i_k}, v_{j_k}), \quad k = 1, \ldots, n$$

  and $v_0 = 0$
- as usual $|\delta_k| \leq \epsilon$
- the relative error of v_k, i.e., $(v_k - u_k)/u_k$ is denoted by θ_k so that

$$v_k = (1 + \theta_k)u_k$$

example with rounding errors

$$
\begin{aligned}
v_1 &= (1+\delta_1)p \\
v_2 &= (1+\delta_2)q \\
v_3 &= (1+\delta_3)v_1^2 \\
v_4 &= (1+\delta_4)(v_3 - 4\,v_2) \\
v_5 &= (1+\delta_5)\sqrt{v_4} \\
v_6 &= (1+\delta_6)\left(-v_1 + v_5\right)/2
\end{aligned}
$$

total error at every step – for multiplication and division

- recall: $f_k(x_i, x_j)$ is either an arithmetic binary operation (like sum) of x_i and x_f *or* a unary operation $f_k(x_i)$
- the simplest cases are multiplication and division
- for multiplication $f_k(v_i, v_j) = (1+\theta_i)(1+\theta_j)u_i u_j$ and so

$$v_k = (1+\delta_k)(1+\theta_i)(1+\theta_j)\, u_k$$

  - multiplication:

$$\theta_k = (1+\delta_k)(1+\theta_i)(1+\theta_j) - 1 \approx \theta_i + \theta_j + \delta_k$$

  - division:

$$\theta_k = (1+\delta_k)(1+\theta_i)/(1+\theta_j) - 1 \approx \theta_i - \theta_j + \delta_k$$

total error at every step – for addition and subtraction

- for addition $f_k(v_i, v_j) = (1+\theta_i)u_i + (1+\theta_j)u_j$ and so

$$v_k = (1+\delta_k)\left((1+\theta_i)\frac{u_i}{u_i+u_j} + (1+\theta_j)\frac{u_j}{u_i+u_j}\right)(u_i+u_j)$$

  - addition:

    $$\theta_k = (1+\delta_k)\left(1+\zeta_k\theta_i + (1-\zeta_k)\theta_j\right) - 1 \approx \zeta_k\theta_i + (1-\zeta_k)\theta_j + \delta_k$$

    where $\zeta_k = u_i/(u_i+u_j)$
- convex combination if u_i and u_j have equal sign
- if different sign, error can be very large despite the fact that some times $\delta_k = 0$ in this case
- similar for subtraction

total error at every step – for univariate function

- $f_k(v_i) = f_k((1+\theta_i)u_i)$ and so

$$\begin{aligned} v_k &= (1+\delta_k)\, f_k((1+\theta_i)u_i \\ &= (1+\delta_k)\, \left(1 + \frac{f_k((1+\theta_i)u_i) - f_k(u_i)}{f_k(u_i)}\right) u_k \\ &= (1+\delta_k)\,(1+\zeta_k\theta_i)u_k \end{aligned}$$

  where $\zeta_k = \frac{f_k((1+\theta_i)u_i) - f_k(u_i)}{\theta_i f_k(u_i)}$ and

$$|\zeta_k| \le \frac{L_k|u_i|}{|f_k(u_i)|}$$

  if L_k is Lipschitz constant of f_k

- relative error of v_k is then

$$\theta_k = (1+\delta_k)(1+\zeta_k\theta_i) - 1 \approx \zeta_k\theta_i + \delta_k$$

relative errors for example

$$
\begin{aligned}
\theta_1 &= \delta_1 \\
\theta_2 &= \delta_2 \\
\theta_3 &= (1+\delta_3)(1+\theta_1)^2 - 1 \\
\theta_4 &= (1+\delta_4)(1+\zeta_4\theta_3 - (1-\zeta_4)\theta_2) - 1 \\
\theta_5 &= (1+\delta_5)(1+\zeta_5\theta_4) - 1 \\
\theta_6 &= (1+\delta_6)(1-\zeta_6\theta_1 + (1-\zeta_6)\theta_5) - 1
\end{aligned}
$$

- homework: what are the ζ_k, get bounds and obtain a bound for θ_6

stability and growth factor

- we say that the f_k are **stable** for if there exists some $L > 0$ such that for all k one has

$$|f_k(x_1, x_2) - f_k(y_1, y_2)| \leq L \max_i |x_i - y_i|$$

- we assume that for $k > 0$ one has $u_k \neq 0$
- then one can define a *growth factor*

$$\rho = \max\{|u_j|/|u_k| \mid j < k\}$$

a simple global error bound

Proposition Let $\alpha = (1+\epsilon)L\rho$ where L be as defined above, ρ be the growth factor then

$$v_k = (1+\theta_k)u_k$$

where

$$|\theta_k| \leq \left(\frac{\alpha^{k+1}-1}{\alpha-1}\right)\epsilon$$

proof.

- induction
- first one has
$$v_1 = (1+\delta_1)u_1$$
and thus $\theta_1 = \delta_1$ and $|\theta_1| = |\delta_1| \le \epsilon$
- then
$$\begin{aligned} v_{k+1} &= (1+\delta_{k+1}) f_{k+1}(v_{i_{k+1}}, v_{j_{k+1}}) \\ &= (1+\theta_{k+1}) u_{k+1} \end{aligned}$$
where
$$\theta_{k+1} = \delta_{k+1} + (1+\delta_{k+1}) \frac{f_{k+1}(v_{i_{k+1}}, v_{j_{k+1}}) - f_{k+1}(u_{i_{k+1}}, u_{j_{k+1}})}{u_{k+1}}$$

- the (absolute value of the) first term is bounded by ϵ and for the second term one has for some $0 < i \leq k$:

$$(1+\delta_{k+1})\left|\frac{f_{k+1}(v_{i_{k+1}}, v_{j_{k+1}}) - f_{k+1}(u_{i_{k+1}}, u_{j_{k+1}})}{u_{k+1}}\right| \leq (1+\epsilon)L\frac{|v_i - u_i|}{|u_{k+1}|}$$
$$= \frac{(1+\epsilon)L|\theta_i| \cdot |u_i|}{|u_{k+1}|}$$
$$\leq L(1+\epsilon)\frac{\alpha^{i+1}-1}{\alpha-1}\rho\epsilon$$
$$\leq \frac{\alpha^{k+2}-\alpha}{\alpha-1}\epsilon$$

from which one gets

$$|\theta_{k+1}| \leq \frac{\alpha^{k+2}-1}{\alpha-1}\epsilon$$

■

example $\left(-p+\sqrt{p^2-4\,q}\right)/2$

$$
\begin{aligned}
u_1 &= p \\
u_2 &= q \\
u_3 &= u_1^2 \\
u_4 &= u_3 - 4\,u_2 \\
u_5 &= \sqrt{u_4} \\
u_6 &= \left(-u_1 + u_5\right)/2
\end{aligned}
$$

$$\Longleftrightarrow$$

$$\mathbf{U} = \mathbf{F}(\mathbf{U}), \quad \mathbf{U} = \left(u_1, u_2, u_3, u_4, u_5, u_6\right)^T$$

Linearized model

$$\mathbf{V} = \mathbf{F}(\mathbf{V}) + \underbrace{\boldsymbol{\beta}}_{\text{abs. round error}}, \quad \mathbf{V} = (v_1, v_2, v_3, v_4, v_5, v_6)^T$$

- Assume that the f_k are **continuously differentiable** so that

$$\begin{aligned} \mathbf{F}(\mathbf{V}) &\approx \mathbf{F}(\mathbf{U}) + \mathbf{J}(\mathbf{V} - \mathbf{U}) \\ &= \mathbf{U} + \mathbf{J}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = \mathbf{V} - \mathbf{U} \end{aligned}$$

  J: $J_{ki} = \partial f_k / \partial u_i$ is the **Jacobian** and $\boldsymbol{\epsilon}$ is the **absolute error**

$$\mathbf{V} = \mathbf{U} + \mathbf{J}\boldsymbol{\epsilon} + \boldsymbol{\beta} \iff (\mathbf{I} - \mathbf{J})\boldsymbol{\epsilon} = \boldsymbol{\beta} \iff \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{J})^{-1}\boldsymbol{\beta}$$

$$\|\boldsymbol{\epsilon}\| \leq \|(\mathbf{I} - \mathbf{J})^{-1}\| \|\boldsymbol{\beta}\|$$

- For the relative error:

$$\|\boldsymbol{\epsilon}_{\text{rel}}\| \leq \mathbf{M} \|(\mathbf{I} - \mathbf{J})^{-1}\| \epsilon_{\text{machine}}, \quad \mathbf{M} > 0.$$