Laura Leal

1) $x \in \mathbb{R}$

$x = \pm 0.d_1 d_2 d_3 d_4 \dots \times 2^m$

$rd(x) = \pm 0.e_1 e_2 e_3 e_4 \dots e_p \times 2^m$

where $e_p = d_p$ if $d_{p+1} = 0$ or $e_p = d_p + 1$ if $d_{p+1} = 1$, and the rest of the $e_1, \dots, e_{p-1}$ are the d_i appropriately adjusted (ie, $d_p = 1$ and $d_{p+1} = 1$ then $e_p = 0$ and $e_{p-1} = d_{p-1} + 1$, etc).

Without loss of generality, assume $x > 0$ with rounding. We have 2 cases to consider:

i) $d_{p+1} = 0$:

$\Rightarrow e_i = d_i$, for $i = 1, \dots, p$

$\Rightarrow x - rd(x) = 0.00\dots0d_{p+1}\dots \times 2^m = d_{p+1}.d_{p+2}\dots \times 2^{m-p-1}$

Since $d_{p+1} < 1$, $x - rd(x) \leq 2^{m-p-1}$

ii) $d_{p+1} = 1$:

Take $e_i = d_i$, for $i = 1, \dots, p-1$ and $e_p = d_p + 1$

$\Rightarrow x - rd(x) = -0.00\dots01(2-d_{p+2})(2-d_{p+3})\dots \times 2^m = -1.(\beta - d_{p+2})\dots \times 2^{m-p-1}$

since $d_{p+1} \geq 1$, $x - rd(x) \geq -1 \times 2^{m-p-1} = -2^{m-p-1}$

Hence, from (i) and (ii), $|x - rd(x)| \leq 2^{m-p-1}$

Now, if $x \neq 0$, then $|x| \geq 2^{m-1}$ $\Rightarrow \frac{1}{|x|} \leq 2^{1-m}$.

$\Rightarrow \frac{|x - rd(x)|}{|x|} \leq 2^{m-p-1} \cdot 2^{1-m} = 2^{-p}$

3) a) i) $fl(x) = x$ (assumption: x is a machine number)

$$fl(x \cdot x) = x^2(1+\varepsilon_x)$$

$$fl(x^2(1+\varepsilon_x)\cdot x) = x^3(1+\varepsilon_x)(1+\varepsilon_x) = x^3(1+2\varepsilon_x)$$

$$\vdots$$

$$fl(x^{n-1}(1+\varepsilon_x)\cdot x) = x^n(1+(n-2)\varepsilon_x)(1+\varepsilon_x) = x^n(1+(n-1)\varepsilon_x)$$

$$\Rightarrow \boxed{\varepsilon_{x^n} = (n-1)\varepsilon_x} \quad (i)$$

assumptions

ii) We need $fl\left(e^{fl\left(\overset{=n}{fl(n)}\cdot \overset{=\ln(x)(1+\varepsilon_{ln})}{fl(\ln(fl(x)))}\right)}\right) =$

$$= fl\left(e^{(n\ln x)(1+\varepsilon_{ln})(1+\varepsilon_{rd})}\right) =$$

$$= e^{n\ln x(1+\varepsilon_{ln})(1+\varepsilon_{rd})}(1+\varepsilon_{exp})$$

We want this to be equal to $e^{n\ln x}(1+\varepsilon_{final})$, in order to find ε_{final}.

$$\Rightarrow e^{n\ln x(1+\varepsilon_{ln}+\varepsilon_{rd})}(1+\varepsilon_{exp}) = e^{n\ln x}(1+\varepsilon_{final})$$

$$\Rightarrow e^{n\ln x}\cdot e^{n\ln x(\varepsilon_{ln}+\varepsilon_{rd})}(1+\varepsilon_{exp}) = e^{n\ln x}(1+\varepsilon_{final})$$

For small values of $n\ln x(\varepsilon_{ln}+\varepsilon_{rd})$, we have $e^{n\ln x(\varepsilon_{ln}+\varepsilon_{rd})} = 1+n\ln x(\varepsilon_{ln}+\varepsilon_{rd})$

Hence, $\boxed{\varepsilon_{final} = n\ln x(\varepsilon_{ln}+\varepsilon_{rd}) + \varepsilon_{exp}}$ (ii)

iii) Comparing the two methods, $\varepsilon_{x^n} > \varepsilon_{final}$ when:

$$(n-1)\varepsilon_x > n\ln x(\varepsilon_{ln}+\varepsilon_{rd}) + \varepsilon_{exp}$$

$$\Rightarrow \boxed{\ln x < \frac{1}{(\varepsilon_{ln}+\varepsilon_{rd})n}\left[(n-1)\varepsilon_x - \varepsilon_{exp}\right]}$$

← if this condition is true, then exponentiating via repeated multiplication is more accurate than log-exp method.

3) b) i)

$$x^{a(1+\varepsilon_a)} = x^a x^{a\varepsilon_a}$$

$$= x^a e^{a\varepsilon_a \ln x}$$

$$\approx x^a (1 + \underbrace{a\varepsilon_a \ln x}_{=\varepsilon})$$

$$\boxed{\varepsilon = a\varepsilon_a \ln x}$$

⤳ problems : $a \ln x$ really big when
i) x really small ($\to 0$), a large ($\to \infty$)
ii) x really large ($\to \infty$), a large ($\to \infty$)

ii) $$(x(1+\varepsilon_x))^a = x^a(1+\varepsilon_x)^a$$

$$\approx x^a(1 + \underbrace{a\varepsilon_x}_{\varepsilon})$$

$$\boxed{\varepsilon = a\varepsilon_x}$$

⤳ would face an overflow problem when a is too large

(ii) case is the worst...

4) a) $$K = \left|\frac{x f'(x)}{f(x)}\right| = \left|\frac{x \cdot e^{-x}}{1-e^{-x}}\right| = \left|\frac{x}{e^{x}-1}\right|$$

We want to show that $K \leq 1$ for $x \in [0,1]$. Notice that for $g(x) = \frac{x}{e^x - 1}$,

$g'(x) = \frac{e^x(1-x)-1}{(e^x-1)^2} > 0$, since $e^x(1-x) > 1$ (since $e^x > 1$ and $1-x > 1$ for $x \in [0,1]$). Hence, we can check K for $x=0$ and $x=1$.

$K|_{x=0} = 0$ and $K|_{x=1} = \left|\frac{1}{e-1}\right| \leq 1$, since $e-1 > 1$

Therefore, $\boxed{(\text{cond} f)(x) := K = \left|\frac{x}{e^x-1}\right| \leq 1}$

b) Let's calculate the error in the algorithm:

$fl(1.0 - fl(e^{-x})) = fl(1.0 - e^{-x}(1+\varepsilon_{exp}))$

Remember that for sums and subtractions, we have

$\varepsilon_{xy} = \frac{x}{x+y}|\varepsilon_x| + \frac{y}{x+y}|\varepsilon_y|$, which, in our case, is:

$\varepsilon_{xy} = \frac{e^{-x}}{1-e^{-x}}\varepsilon_{exp} = \frac{1}{e^x-1}\varepsilon_{exp}$.

$$\Rightarrow fl(1 - fl(e^{-x})) = fl\left((1-e^{-x})\left(1 + \frac{1}{e^x-1}\varepsilon_{exp}\right)\right)$$

$$= (1-e^{-x})\left(1+\frac{1}{e^x-1}\varepsilon_{exp}\right)(1+\varepsilon_{rd})$$

$$\Rightarrow f_A(x) = (1-e^{-x})\left(1 + \frac{1}{e^x-1}\varepsilon_{exp} + \varepsilon_{rd}\right)$$

$$= 1 - e^{-x} + (1-e^{-x})\varepsilon_{rd} + e^{-x}\varepsilon_{exp}$$

Set $f(x_A) = f_A(x) \Rightarrow \cancel{1} - e^{-x_A} = \cancel{1} - e^{-x} + (1-e^{-x})\varepsilon_{rd} + e^{-x}\varepsilon_{exp}$

$$\Rightarrow e^{x - x_A} = 1 - (e^x - 1)\varepsilon_{rd} + \varepsilon_{exp}$$

$$\Rightarrow x - x_A = \ln(1 + (1-e^x)\varepsilon_{rd} + \varepsilon_{exp})$$

$$\Rightarrow x - x_A \approx (1-e^x)\varepsilon_{rd} + \varepsilon_{exp}$$

continues →

4b - cont)

$$\Rightarrow \quad |x - x_A| \leq eps + eps - e^x eps$$
$$\leq eps(2 - e^x)$$

$$\Rightarrow \quad \left| \frac{x - x_A}{x} \right| \leq eps \left(\frac{2 - e^x}{x} \right)$$

$$(\text{cond}\, A)(x) = \sup \left[\frac{|x - x_A|}{|x|} \cdot \frac{1}{eps} \right] = \frac{2 - e^x}{x}$$

$\Rightarrow (\text{cond}\, A)(x) > 1$ when $x \in [0, 1]$

since $\underbrace{x}_{<1} + \underbrace{e^x}_{<1} < 1 + 1 = 2$

c) (see code for plots)

Ill-conditioning comes from subtracting e^{-x} from 1. For small values of x, e^x is very close to 1, hence the error is magnified.

d) From class, we have $2^{-b} \leq 1 - \frac{y}{x} \leq 2^{-a}$. We will only use the first inequality, with $b = 1, 2, 3, 4$.

$$2^{-b} \leq 1 - e^{-x} \quad \Rightarrow \quad x \geq -\ln(1 - 2^{-b})$$

solving for x (see code), we get: $x \approx \begin{array}{cccc} b=1 & b=2 & b=3 & b=4 \\ \end{array}$ $[0.85, 0.68, 0.62, 0.59]$

e) Now, we need the other side of the equation: $1 - \frac{y}{x} \leq 2^{-a}$.

$$\Rightarrow \quad 1 - e^{-x} \leq 2^{-a} \quad \Rightarrow \quad a \leq -\frac{1}{\ln 2} \ln(1 - e^{-x})$$

$$\Rightarrow \quad a = \begin{bmatrix} 1, & 1, & 2, & 3 \end{bmatrix}$$
$b=1 \quad b=2 \quad b=3 \quad b=4$

(see code)

4f) For small x, we can approximate

$$1-e^{-x} = \cancel{1} - \left[\cancel{1} - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \dots\right]$$

$$= x\left[1 - \frac{x}{2!} + \frac{x^2}{3!} - \frac{x^3}{4!} + \dots\right]$$

$$= x\left[1 - \frac{x}{2}\left[1 - \frac{x}{3}\left[1 - \frac{x}{4}\left[\dots\right]\right]\right]\right]$$

We can truncate this embedding at some level of accuracy and obtain a reasonable, well-conditioned algorithm for $1-e^{-x}$ for x small, since $1 - \frac{x}{n}\cdot c$ does not have exploding error for x small.

$|c| < \infty$.

6) Very cool problem! (please see code for implementation)

Explanation:

Notice $x^{\frac{1}{2^n}} \approx 1 + \ln\left(x^{\frac{1}{2^n}}\right)$

$= 1 + \frac{1}{2^n} \ln(x)$

$\boxed{\ln x = -2^{-n}}$

Machine number at double precision is $\varepsilon = 2^{-52}$. Hence, we are rounding down to 0 all the excess $\frac{1}{2^n}\ln(x)$ when we take the square root. When we square back the numbers, only those who had the rounded part already 0 (in binary expansion) will go back to what they were before.

7) e) Source: Tsai, Edison "A method for reducing ill-conditioning of polynomial root finding using a change of basis" (2014) University of Honors thesis, paper 109

i) Theorem: Let $p(x) = \sum_{i=0}^{n} c_i x^i$ be a degree n polynomial with coefficients c_i for $0 \leq i \leq n$. If r is a nonzero root of $p(x)$ with multiplicity 1, and $c_j \neq 0$, then the relative condition number of r wrt c_j is $K = \dfrac{|c_j r^{j-1}|}{|p'(r)|}$.

Proof: Let Δc_j be any perturbation of the j^{th} coefficient. Define the polynomial $\hat{p}(x)$ as the result of perturbing the j^{th} coefficient of $p(x)$ by Δc_j, so that $\hat{p}(x) = p(x) + \Delta c_j x^j$, and denote the corresponding root of $\hat{p}(x)$ by $\hat{r}$. Since the coefficients of any polynomial can be given as continuously differentiable function of the roots, it follows from the inverse function theorem that the roots are continuous functions of the coefficients as well. In particular, r may be given as a continuous function $r(c_j)$ of the j^{th} coefficient c_j, with all other coefficients being held constant. Therefore, as $\Delta c_j \to 0$, $\hat{r} \to r$, and we have $r(c_j) = r$, $r(c_j + \Delta c_j) = \hat{r}$. By the definition of condition number, $K = \lim_{\delta \to 0} \sup_{\Delta c_j < \delta} \dfrac{\frac{|r(c_j + \Delta c_j) - r(c_j)|}{|r(c_j)|}}{\frac{|\Delta c_j|}{|c_j|}}$. Consider the limit $\lim_{\Delta c_j \to 0} \dfrac{\frac{|r(c_j+\Delta c_j) - r(c_j)|}{|r(c_j)|}}{\frac{|\Delta c_j|}{|c_j|}}$.

We have

$$\lim_{\Delta c_j \to 0} \frac{\frac{|r(c_j+\Delta c_j) - r(c_j)|}{|r(c_j)|}}{\frac{|\Delta c_j|}{|c_j|}} = \lim_{\Delta c_j \to 0} \frac{\frac{|\hat{r} - r|}{|r|}}{\frac{|\Delta c_j|}{|c_j|}} = \lim_{\Delta c_j \to 0} \left| \frac{c_j r^{j-1}}{\frac{\Delta c_j \hat{r}^j}{\hat{r} - r}} \right| = \lim_{\Delta c_j \to 0} \left| \frac{c_j r^{j-1}}{\frac{p(\hat{r}) - p(r)}{\hat{r} - r}} \right| =$$

$$= \frac{|c_j r^{j-1}|}{\left| \lim_{\Delta c_j \to 0} \frac{p(\hat{r}) - p(r)}{\hat{r} - r} \right|} = \frac{|c_j r^{j-1}|}{|p'(r)|}.$$

Since this limit exists, we must have

$$K = \lim_{\delta \to 0} \sup_{\Delta c_j < \delta} \frac{\frac{|r(c_j+\Delta c_j) - r(c_j)|}{|r(c_j)|}}{\frac{|\Delta c_j|}{|c_j|}} = \lim_{\Delta c_j \to 0} \frac{\frac{|r(c_j+\Delta c_j) - r(c_j)|}{|r(c_j)|}}{\frac{|\Delta c_j|}{|c_j|}} = \frac{|c_j r^{j-1}|}{|p'(r)|}$$

For our problem:

$$\boxed{(\mathrm{cond}\, \Omega_k)(\vec{a}) = \sum_{l=0}^{n-1} \frac{|a_l \Omega_k^{l-1}|}{|p'(\Omega_k)|}}$$

7) e) iii) A clever algorithm could help us here. So far, we have assumed that polynomials are represented as $p(x) = \sum_{k=0}^{n} a_k x^k$, but this doesn't have to be so. We may represent it as $p(x) = \sum_{k=0}^{n} b_k p_k(x)$, where $\{p_0, p_1, \dots, p_n\}$ is a basis for the vector space $\mathbb{P}_n$ of all polynomials of degree $\le n$. Let's get the conditioning number for this new basis.

__Theorem__: Let $\{p_0, p_1, \dots, p_n\}$ be a basis for $\mathbb{P}_n$, and let $p(x) = \sum_{k=0}^{n} b_k p_k(x)$ be a degree n polynomial. If r is a non-zero root of $p(x)$ with multiplicity 1, and $b_j \ne 0$ for some $0 \le j \le n$, then the relative condition number of r wrt b_j is $K = \dfrac{|b_j p_j(r)|}{|r p'(r)|}$.

__Proof:__ Let Δb_j be an arbitrary perturbation of the coefficient b_j, and define $\hat{p}(x)$ to be the result of perturbing the j^{th} coefficient of $p(x)$ by Δb_j. Then $\hat{p}(x) = p(x) + \Delta b_j p_j(x)$. Define $\hat{r}$ to be the corresponding root of the perturbed polynomial. By the same argument as the previous theorem, as $\Delta b_j \to 0$, $\hat{r} \to r$ also. By the definition of condition number, we have $K = \lim_{\Delta \to 0} \sup_{\Delta b_j < \delta} \dfrac{\frac{|\hat{r} - r|}{|r|}}{\frac{|\Delta b_j|}{|b_j|}}$. Consider the limit

$$\lim_{\Delta b_j \to 0} \frac{\frac{|\hat{r}-r|}{|r|}}{\frac{|\Delta b_j|}{|b_j|}} = \lim_{\Delta b_j \to 0} \left| \frac{b_j p_j(r)}{\frac{r \Delta b_j p_j(r)}{\hat{r}-r}} \right| = \lim_{\Delta b_j \to 0} \left| \frac{b_j p_j(r)}{r \left(\frac{p(\hat{r}) - p(r)}{\hat{r} - r} \right)} \right| = \frac{|b_j p_j(r)|}{\left| r \lim_{\Delta b_j \to 0} \frac{p(\hat{r}) - p(r)}{\hat{r} - r} \right|} =$$

$= \dfrac{|b_j p_j(r)|}{|r p'(r)|}$. Since the limit exists, it follows that:

$$K = \lim_{\Delta \to 0} \sup_{\Delta b_j < \delta} \frac{\frac{|\hat{r}-r|}{|r|}}{\frac{|\Delta b_j|}{|b_j|}} = \lim_{\Delta b_j \to 0} \frac{\frac{|\hat{r}-r|}{|r|}}{\frac{|\Delta b_j|}{|b_j|}} = \frac{|b_j p_j(r)|}{|r p'(r)|}.$$

Notice that large values of K come from large values of coefficients b_i and values $p_i(r)$. We can't control the coefficients, thus we want to choose a basis such that $p_i(x)$ is small over the interval $[a, b]$ where the roots are contained. For that, we use Chebyshev polynomials $T_n(x)$ [remember, for all n, $|T_n(x)| \le 1$, for $x \in [-1, 1]$], hence we rescale the interval $[a, b]$ to $[-1, 1]$. Let $t = \dfrac{2(x-a)}{b-a} - 1 \in [-1, 1]$ when $x \in [a, b]$.

$\rightarrow$

7e iii - cont) Algorithm:

1) Start with a set of $n+1$ data points (t_i, y_i) for $0 \le i \le n$, and suppose that the roots of the polynomial $p(x)$ (which interpolates these data points) all lie in the interval $[a,b]$.

2) Make the change of variables $t = \frac{2(x-a)}{b-a} - 1$ to obtain $t_i \in [-1,1]$. There exists a unique poly of degree n $\bar{p}(t)$ st $\bar{p}(t_i) = y_i$, $0 \le i \le n$.

3) Express the interpolating polynomial as a linear combination of Chebyshev polynomials

4) find the roots of $\bar{p}(t)$: $\{\bar{r}_1, \bar{r}_2, \dots, \bar{r}_n\}$

5) Make the change of variables $r_i = \frac{b-a}{2}(\bar{r}_i + 1) + a$ to find the roots of the original polynomial $p(x)$.

8) a -

$$y_m = \frac{e - y_{m+1}}{m+1}$$

$$y_{N-1} = \frac{e - y_N}{N}$$

$$y_{N-2} = \frac{1}{N-1}\left[e - \frac{1}{N}(e - y_N)\right]$$

$$= \frac{1}{N(N-1)}\left[Ne - e + y_N\right]$$

$$= \frac{1}{N(N-1)}\left[(N-1)e + y_N\right]$$

$$y_{N-3} = \frac{1}{N-2}(e - y_{N-2})$$

$$= \frac{1}{N-2}\left[e - \frac{1}{N(N-1)}((N-1)e + y_N)\right]$$

$$= \frac{1}{N(N-1)(N-2)}\left[(N(N-1) - N + 1)e + y_N\right]$$

$$y_{N-4} = \frac{1}{N-3}(e - y_{N-3})$$

$$= \frac{1}{N-3}\left[e - \frac{1}{N(N-1)(N-2)}((N(N-1) - N + 1)e + y_N\right]$$

$$= \frac{1}{N(N-1)(N-2)(N-3)}\left[(N(N-1)(N-2) - N(N-1) + N - 1)e + y_N\right]$$

(...)

$$\boxed{y_{N-k} = \left(\prod_{m=0}^{k-1}(N-m)^{-1}\right)\cdot\left[\left[\left(\sum_{l=2}^{k}(-1)^l \prod_{i=0}^{l-2}(N-i)\right) + (-1)^{k+1}\right]e + y_N\right]}$$

$\boxed{\text{renaming to get } y_k = g(y_N):}$ $\Rightarrow$ $$\boxed{y_k = \left(\prod_{m=0}^{N-k-1}(N-m)^{-1}\right)\cdot\left[\left[\sum_{l=2}^{N-k}(-1)^l \prod_{i=0}^{l-2}(N-i)\right] + (-1)^{N-k+1}\right]e + y_N\Big]}$$

continues →

8 - a , cont) We know that $(\text{cond} f)(x) = \left| \dfrac{x f'(x)}{f(x)} \right|$

In the case of this problem, $(\text{cond}\, g_k)(y_N) = \left| \dfrac{y_N \cdot g'(y_N)}{y_k} \right|$

Notice $g'(y_N) = \prod_{m=0}^{N-k-1} (N-m)^{-1}$. Hence, the upper limit of $(\text{cond}\, g_k)(y_N)$ in terms of k and N can be written as

$$(\text{cond}\, g_k)(y_N) = \left| \frac{y_N}{y_k} \cdot \prod_{m=0}^{N-k-1} (N-m)^{-1} \right| \leq \left| \frac{k!}{N!} \right|$$

since $\dfrac{y_N}{y_k} < 1$.

b) We want $\text{rel\_error}_{\text{output}} \leq (\text{cond}\, g_k)(y_N) \cdot \text{rel\_error}_{\text{input}}$

$$\Rightarrow \quad \varepsilon \leq \frac{k!}{N!} \cdot 1$$

$$\Rightarrow \quad N! \leq \frac{k!}{\varepsilon}$$ ⇝ find $\underline{N}$.

c) $N = 32$