Written Solutions

Lu Vy

September 28, 2020

Problem 1.1

Part A

The existence of a positive root $\rho \in \left(0, \sqrt[n]{\beta}\right)$ is clear from the intermediate value theorem since $p(0) = -\beta < 0 < \beta^{(n-1)/n} = p\left(\sqrt[n]{\beta}\right)$. To show that this positive root is unique, it suffices to note that $p'(x) = nx^{n-1} + (n-1)\,x^{n-2}$ is positive whenever $x > 0$, so $p(x)$ is monotonically increasing on $(0, \infty)$.

Part B

If ρ is a root of $p(x)$, then we require

$$\rho^n + \rho^{n-1} - \beta = 0. \tag{1}$$

Implicit differentiation with respect to β yields

$$n\rho^{n-1}\frac{\mathrm{d}\rho}{\mathrm{d}\beta} + (n-1)\,\rho^{n-2}\frac{\mathrm{d}\rho}{\mathrm{d}\beta} - 1 = 0,$$

from which we attain

$$\frac{\mathrm{d}\rho}{\mathrm{d}\beta} = \frac{1}{n\rho^{n-1} + (n-1)\,\rho^{n-2}}. \tag{2}$$

Substitute this into $c(\beta, n) = \left|\frac{\mathrm{d}\rho}{\mathrm{d}\beta}\right|\left|\frac{\beta}{\rho}\right|$:

$$\begin{aligned}
c(\beta, n) &= \left|\frac{\beta}{n\rho^n + (n-1)\,\rho^{n-1}}\right| \\
&= \left|\frac{\beta}{n\rho^n + (n-1)\,(\beta - \rho^n)}\right| && \text{by (1)} \\
&= \left|\frac{\beta}{n\beta - \beta + \rho^n}\right| \\
&= \left|\frac{1}{n - 1 + \frac{1}{\beta}\rho^n}\right| = \frac{1}{n - 1 + \frac{1}{\beta}\rho^n} && \rho > 0.
\end{aligned}$$

Part C

$$0 < \rho < \sqrt[n]{\beta} \implies \frac{1}{n - 1 + \frac{1}{\beta}} < c(\beta, n) < \frac{1}{n-1}$$

Part D

We have that $0 < c(\beta, n) < 1$ for every selection of $\beta > 0$ and $n \geq 2$, so ρ is always well conditioned. Furthermore, if $n \to \infty$, then $c \to 0$. This means that ρ is less sensitive to perturbations in β for large n.

Problem 1.2

Part A

When $\beta = 2$, a "1" must lead the mantissa of any normalized floating point number. We may save a bit if we omit this "1" in the code-word representation (provided that the number represented is normalized). The floating point number is always assumed to be normalized, unless the exponent indicates otherwise. This is the case with the IEEE single and double precision standards.

Part B

$$1 + 5 + 3\,(12) = 42$$

One bit is devoted to the sign. There are $16 - (-15) + 1 = 32$ possible exponents, which may be represented by five bits. Finally, three bits are required for each of the twelve digits of the mantissa.

Part C

If there are 7 digits of accuracy to begin with, we may lose 5 and still retain 2.

Part D

There is no risk of cancellation with μ (it is the sum of positive terms), so μ is less sensitive to perturbations. For a more rigorous argument, let us compare relative condition numbers (using the L^1 norm). The inner product $\mu = \sum_{i=1}^{n} \xi_i^2$ is a map from $\mathbb{R}^n \mapsto \mathbb{R}$, so

$$\kappa_\mu = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\mathbf{x}} \right\|_1 \frac{\|\mathbf{x}\|_1}{\mu} = \frac{1}{\left|\sum_{i=1}^{n} \xi_i^2\right|} \left\| \begin{bmatrix} 2\xi_1 \\ \vdots \\ 2\xi_n \end{bmatrix} \right\|_1 \left\| \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \right\|_1 = 2\frac{\left(\sum_{i=1}^{n} |\xi_i|\right)^2}{\sum_{i=1}^{n} \xi_i^2}.$$

On the other hand, the inner product $\gamma = \sum_{i=1}^{n} \xi_i \eta_i$ is a map from $\mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$. We may treat this as a map from $\mathbb{R}^{2n} \mapsto \mathbb{R}$ with arguments $(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_n)$ so that we get

$$\kappa_\gamma = \frac{1}{\left|\sum_{i=1}^{n} \xi_i \eta_i\right|} \left\| \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \\ \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \right\|_1 \left\| \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \\ \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \right\|_1 = \left| \frac{\left(\sum_{i=1}^{n} |\eta_i| + |\xi_i|\right)^2}{\sum_{i=1}^{n} \xi_i \eta_i} \right|.$$

The difference is that the terms in the denominator of κ_γ need not all be positive, so they may be chosen so that the denominator is near 0.

Problem 1.3

Part A

For $n = 3$, it can be verified that the accumulation of error from this results in a sum of the form

$$\begin{aligned}
\sigma = & \sum_{i=1}^{8} \xi_i \left(1+\epsilon_{i,1}\right)\left(1+\epsilon_{i,2}\right)\left(1+\epsilon_{i,3}\right) && \left|\epsilon_{i,j}\right| \leq u \\
= & \sum_{i=1}^{8} \xi_i \left(1+\eta_i\right) && \left|\eta_i\right| \lesssim 3u \\
= & \left(\xi_1 + \cdots + \xi_8\right) + \left(\xi_1\eta_1 + \cdots + \xi_8\eta_8\right).
\end{aligned}$$

The third form is most convenient for evaluating the absolute forward error, which is $|\xi_1\eta_1 + \cdots + \xi_8\eta_8|$. Generalizing, the absolute forward error is

$$\left|\xi_1\eta_1 + \cdots + \xi_{2^k}\eta_{2^k}\right|, \quad |\eta| \lesssim ku.$$

Part B

From the observation that

$$\sigma = \sum_{i=1}^{8} \xi_i \left(1+\eta_i\right),$$

we see that the computed solution actually solves the problem given by the data

$$\left(\xi_1\left(1+\eta_1\right), \ldots, \xi_8\left(1+\eta_8\right)\right)^T.$$

The absolute backward error is therefore

$$\left\| \begin{bmatrix} \xi_1\left(1+\eta_1\right) \\ \vdots \\ \xi_8\left(1+\eta_8\right) \end{bmatrix} - \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_8 \end{bmatrix} \right\|_1 = \left\| \begin{bmatrix} \xi_1\eta_1 \\ \vdots \\ \xi_8\eta_8 \end{bmatrix} \right\|_1 = \sum_{i=1}^{8} |\xi_i\eta_i|.$$

For general n, the absolute backward error by the L_1 norm is

$$\sum_{i=1}^{n} |\xi_i\eta_i|.$$

Part C

An upper bound for the absolute forward error is given by

$$\begin{aligned}
\left|\xi_1\eta_1 + \cdots + \xi_{2^k}\eta_{2^k}\right| & \leq |\xi_1||\eta_1| + \cdots + |\xi_{2^k}||\eta_{2^k}| \\
& \lesssim ku|\xi_1| + \cdots + ku|\xi_{2^k}|
\end{aligned}$$

$$
\begin{aligned}
&= \left(|\xi_1| + \cdots + |\xi_{2^k}|\right) ku \\
&= \kappa_{\text{abs}} \log_2 nu.
\end{aligned}
$$

Under the L_1 norm, the upper bound for the absolute backward error is the same:

$$
\begin{aligned}
|\xi_1\eta_1| + \cdots + |\xi_n\eta_n| &= |\xi_1||\eta_1| + \cdots + |\xi_n||\eta_n| \\
&\lesssim ku|\xi_1| + \cdots + ku|\xi_{2^k}| \\
&= \left(|\xi_1| + \cdots + |\xi_{2^k}|\right) ku \\
&= \kappa_{\text{abs}} \log_2 nu.
\end{aligned}
$$

We have that $p(n) \approx \log_2 n$ for the binary fan-in tree, which grows slower than $p(n) \approx n$ of sequential summation. We have shown that the algorithm is not only weakly stable (because it can be bounded by $\kappa_{\text{abs}} p(n) u$ for some $p(n) \in \mathcal{O}(\log_2 n)$), but also backward stable because it solves a nearby problem (the backward error is small).