Mid term - CPSC 302

1. In a floating point system:

5
β — the base of the system
t — the precision of the system → the number of digits in the mantissa
L — a lower bound on the exponent
U — an upper bound on the exponent

3
$$\underbrace{d_0 . d_1 d_2 d_3 \cdots d_{t-1}} \times \beta^{e}$$
The mantissa is this; exponent, where $L \le e \le U$.

2 of
4

- If $a \simeq b$, then $a-b$ will have a large relative error (cancellation error)
- If $a \ll 1$, then b/a will have a large abs & rel. error
- If $a \gg b$, then $a+b$ will have a large absolute error

a) x and $\sqrt{x^2-1}$ will be very close to each other for large x, so the second formula
$$-\ln(x+\sqrt{x^2-1})$$
4
is the most suitable, as this avoids cancellation error

b) The formula chosen above still squares x, and so may be liable to overflow. This can be avoided by taking, for example,
$$-\ln(x + x\sqrt{1 - 1/x^2})$$
4
we may still get underflow, but this is far less damaging

2\. i) $|x_{k+1} - x| \le L|x_k - x|$, where $|L| < 1$

4 ii) $|x_{k+1} - x| \le Q|x_k - x|^2$ where Q is a const.

Newton's method converges quadratically for a f^{1} $f \in C^2[a,b]$

4 to a root at ξ if $f'(\xi) \neq 0$ and if we start

Newton's method sufficiently close to the root.

$$x_{k+1} = g(x_k)$$

$$|x_{k+1} - \xi| = |g(x_k) - g(\xi)|$$

4 $$= |g'(c)(x_k - \xi)| \quad \text{for some } c \in (a,b)$$

$$\le \max_{c \in (a,b)} |g'(c)| \, |x_k - \xi|$$

$$= L|x_k - \xi|$$

for Newton's method,

3 $$g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}$$

$f(x) = 25x^2 - 10x + 1$

This has a double root at $x = 1/5$.

hence we cannot get quadratic convergence, as

$f'(1/5) = 50(1/5) - 10 = 0$.

Note that, for this function

$$g(x) = x - \frac{25x^2 - 10x + 1}{50x - 10}$$

$$= \frac{50x^2 - \cancel{10x} - 25x^2 + \cancel{10x} - 1}{50x - 10}$$

3 $$= \frac{25x^2 - 1}{10(5x - 1)} = \frac{(\cancel{5x-1})(5x+1)}{10(\cancel{5x-1})}$$

$$= \frac{1}{2}x + \frac{1}{10}$$

$\Rightarrow$ $g'(x) = \frac{1}{2}$, a constant for all x

4 applying the theorem above

$|x_{k+1} - x| \le \frac{1}{2}|x_k - x|$

hence convergence is linear

2 The bisection method will not converge, as there are not points a, b s.t. $f(a)f(b) < 0$.

3.

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 3 \\ 4 & 3 & 3 & 8 \\ 8 & 7 & 9 & 24 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 2 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 3 & 5 & 12 \end{array}\right] \begin{array}{l} R_2 - 2R_1 \\ R_3 - 4R_1 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc|c} 2 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 2 & 6 \end{array}\right] \quad R_3 - 3R_2$$

$$\Rightarrow \quad z = \frac{6}{2} = 3$$

5

$$y = \frac{2-3}{1} = -1$$

$$x = \frac{3 - 3 + 1}{2} = \frac{1}{2}$$

2 2

$$\therefore \begin{bmatrix} 2 & 1 & 1 \\ 4 & 3 & 3 \\ 8 & 7 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

& $\det(A) = \det(L)\det(U)$

$= 1.1.1 \times 2.1.2$

$= 4$

1