SECTION C

NUMERICAL METHODS 1 (3.09) taught by David Ham

Candidates being examined in "Numerical Methods 1" should answer at least one question from Section C.

C1. (i) Consider the number 5.25

(a) Write the number in the form:

$$s1.m \times 2^{e-b} \tag{1}$$

Use a floating point format with 3 exponent bits and 4 mantissa bits, and a bias of 3. All the numbers must be written in binary.
Answer: 5.25 is 101.01_2 or $1.0101_2 \times 2^2$. We need $e - 3 = 1$ so $e = 5 = 101_2$. The sign is positive so in the prescribed form it's

$$01.0101_2 \times 2^{101_2 - 11_2}$$

1 mark for sign, 2 each for exponent and mantissa and 1 for correctly applying the bias.

(6 marks)

(b) Convert this number into the bit pattern which would actually be stored, according to the layout provided on the formula sheet.
Answer: 01010101
If the candidate makes a mistake on the first part but the bit pattern is consistent with their answer, they get the marks.

(2 marks)

(ii) (a) The Bisection method is given by the algorithm:

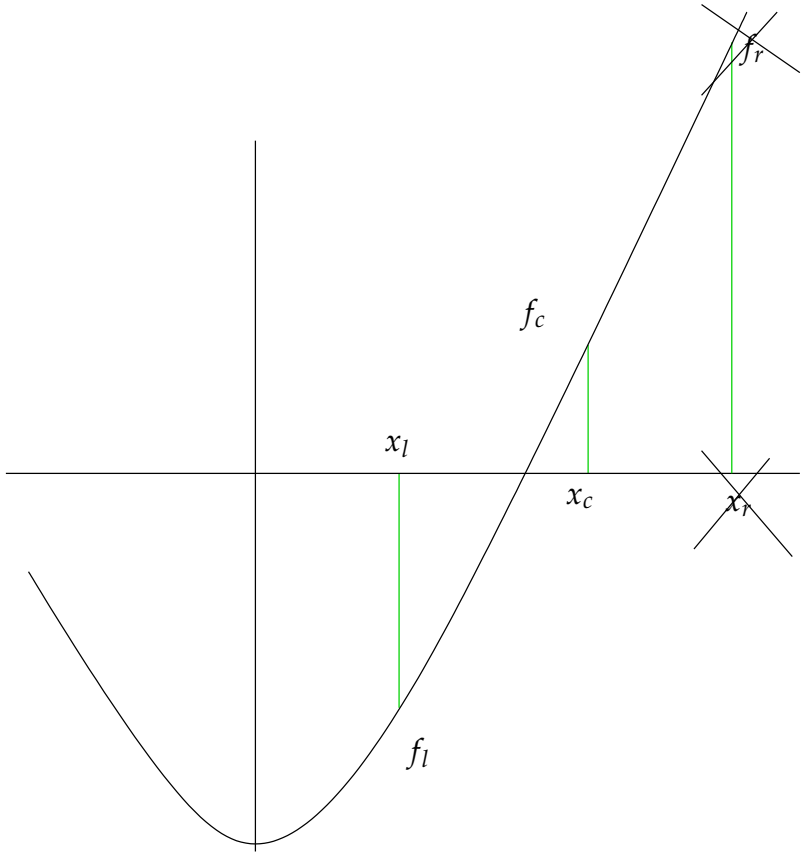
$f_l \leftarrow f(x_l)$
repeat
  $x_c \leftarrow \frac{x_l + x_r}{2}$
  $f_c \leftarrow f(x_c)$
  if $f_c f_l > 0$ **then**
    $x_l \leftarrow x_c$
    $f_l \leftarrow f_c$
  else
    $x_r \leftarrow x_c$
until $|f(x_c)| < \epsilon$

Produce a sketch of one iteration of this algorithm for the function $f(x) = x^2 - 1$ with a starting interval $x_l = 0.5, x_r = 2$. Show all the relevant points and lines and indicate which end of the interval will be removed.

(3 marks)

Answer:

SECTION CONTINUED ON NEXT PAGE



(b) The Newton-Raphson iteration is given by the algorithm:

repeat

$f_x \leftarrow f(x)$

$x \leftarrow x - \frac{f_x}{f'(x)}$

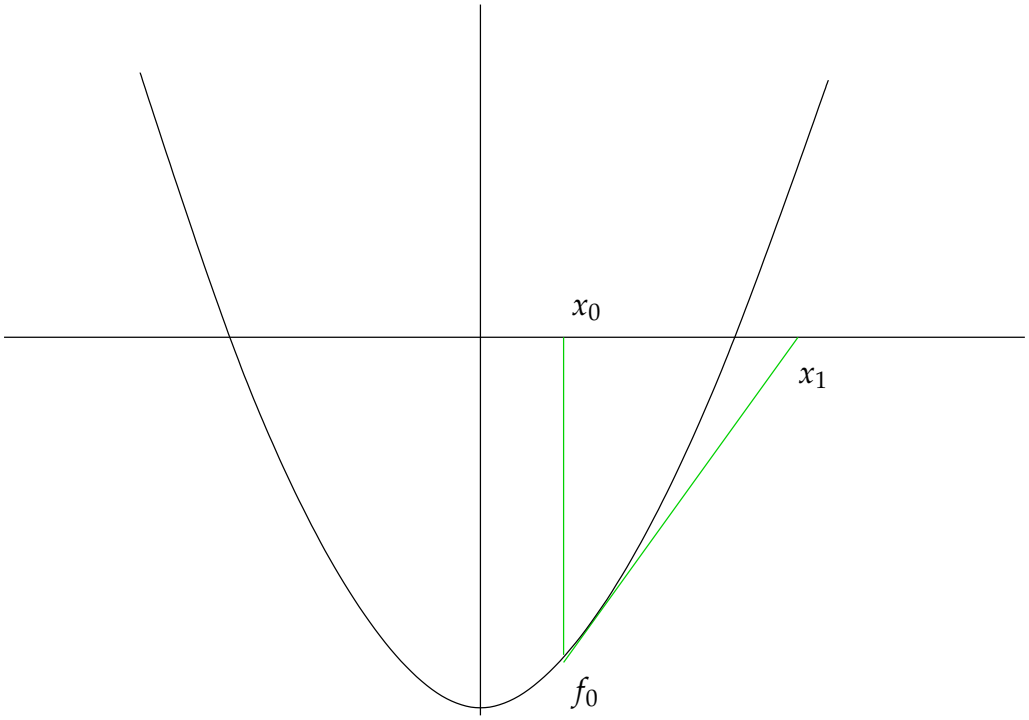
until $(|f_x| < \epsilon)$ or maximum iterations exceeded.

Produce a sketch of one iteration of this algorithm for the function $f(x) = x^2 - 1$ with a starting position $x_0 = 0.5$. Show all the relevant points and lines.

(3 marks)

Answer:

SECTION CONTINUED ON NEXT PAGE



(c) What is the key advantage of Newton-Raphson iteration over the bisection method?

(1 mark)

Answer: Much more rapid convergence (Q-order 2 in most cases rather than 1 for the bisection method).

(d) What additional information is required to use Newton-Raphson iteration rather than the bisection method?

(1 mark)

Answer: The function derivative.

(e) What advantage does the bisection method have over Newton-Raphson iteration?

(1 mark)

Answer: Guaranteed convergence as long as the root lies in the initial interval.

(iii) Suppose that A is a rectangular $n \times m$ matrix with $n > m$, and that the columns of A are all orthogonal to each other, and each column has a modulus of 1 ($|\mathrm{A}_{:,i}| = \sqrt{A_{:,i} \cdot \mathrm{A}_{:,i}} = 1$ for $0 \leq i \leq m-1$).

(a) prove that $\mathrm{A}^{\mathrm{T}}\mathrm{A} = \mathrm{I}$ where I is the $m \times m$ identity matrix.

(4 marks)

Answer: Let $\mathrm{B} = \mathrm{A}^{\mathrm{T}}\mathrm{A}$. Then:

$$\mathrm{B}[i,j] = \mathrm{A}[:,i] \cdot \mathrm{A}[:,j]$$

Where $i \neq j$, $\mathrm{B}[i,j] = 0$ since different columns of A are orthogonal to each other. However, where $i = j$, $\mathrm{B}[i,j] = 1$ since each column of A has modulus 1. We can conclude that B has 1 on the diagonal and 0 elsewhere and is equal to the identity matrix.

(b) prove therefore that the least squares solution to the overdetermined matrix equation:

$$\mathrm{A}\mathbf{x} = \mathbf{b}$$

SECTION CONTINUED ON NEXT PAGE

is given by $\mathbf{x} = \mathrm{A}^{\mathrm{T}}\mathbf{b}$

(4 marks)

Answer: The least squares problem may be solved by multiplying by the transpose of A:

$$\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x} = \mathrm{A}^{\mathrm{T}}\mathbf{b}$$

Now using the previous result, we have:

$$\mathrm{I}\mathbf{x} = \mathrm{A}^{\mathrm{T}}\mathbf{b}$$
$$\mathbf{x} = \mathrm{A}^{\mathrm{T}}\mathbf{b}$$

SECTION CONTINUED ON NEXT PAGE

C2. (i) Convert the numbers in the following problems into 4 bit two's compliment signed binary integers before performing the calculation and converting back into base 10.

(a) $-5+3$

(4 marks)

Answer: 5 is 0101_2 so taking its binary compliment and adding 1 gives: 1011_2 for -5.

$$\begin{aligned} & 1011_2 \\ +& 0011_2 \\ =& \ 1110_2 \end{aligned}$$

Taking the binary compliment of the answer and adding 1 gives $0010_2 = 2$ so the answer is equal to -2. *2 marks for conversions, 2 for the sum*

(b) -1×4

(4 marks)

Answer: 1 is 0001_2 so taking its binary compliment and adding 1 gives: 1111_2 for -1.

$$\begin{aligned} & 1111_2 \\ \times& 0100_2 \\ =& \ \cancel{11}1100_2 \\ =& \ 1100_2 \end{aligned}$$

Taking the binary compliment of the answer and adding 1 gives $0100_2 = 4$ so the answer is equal to -4 *2 marks for conversions, 2 for the product*

(ii) For each of the following Python functions, write a mathematical expression using only matrices and vectors which performs the same operation. In each case, state whether each input and output is a matrix or vector.

(a)
```
def function_1(a,b,c):
    import numpy

    d=numpy.zeros(numpy.size(b))

    for j in range(len(d)):
        d[j]=numpy.dot(a[j,:],c)-b[j]

    return d
```

(4 marks)

Answer: By counting indices, we can see that `a` is a matrix, and `b` and `d` are vectors. For the dot product to be defined, `c` must also be a vector. The function calculates:

$$\mathbf{d} = \mathrm{A}\mathbf{c} - \mathbf{b}$$

2 marks for types, 2 for the operation.

(b)
```
def function_2(a):
    import numpy

    c=a.shape[1]
```

SECTION CONTINUED ON NEXT PAGE

```
b=numpy.zeros((c,c))

for i in range(c):
    for j in range(c):
        b[i,j]=numpy.dot(a[:,i],a[:,j])

return b
```

(4 marks)

Answer: By counting indices, we can see that `a` and `b` must both be matrices.
The function calculates:

$$B = A^T A$$

2 marks for types, 2 for the operation.

(iii) A central difference approximation to the third derivative of a function $f(x)$ is given by:

$$f'''(x) = \frac{-f(x-2h) + 2f(x-h) - 2f(x+h) + 2f(x+2h)}{2h^3} + \mathcal{O}(h^p)$$

Where h is some small positive step size, and p is the order of convergence. Using Taylor series for the function f, or otherwise, prove that the order of convergence, p, is equal to 2.
You may use without proof any of the properties of $\mathcal{O}$.

(9 marks)

Answer: First write Taylor series centred at x for each of the function evaluations:

$$f(x-2h) = f(x) - 2hf'(x) + \frac{4h^2}{2}f''(x) - \frac{8h^3}{6}f^{(3)}(x) + \frac{16h^4}{24}f^{(4)}(x) + \mathcal{O}(h^5)$$
$$f(x+2h) = f(x) + 2hf'(x) + \frac{4h^2}{2}f''(x) + \frac{8h^3}{6}f^{(3)}(x) + \frac{16h^4}{24}f^{(4)}(x) + \mathcal{O}(h^5)$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f^{(3)}(x) + \frac{h^4}{24}f^{(4)}(x) + \mathcal{O}(h^5)$$
$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f^{(3)}(x) + \frac{h^4}{24}f^{(4)}(x) + \mathcal{O}(h^5)$$

3 marks for Taylor series.
Now, write $-f(x-2h) + 2f(x-h) - 2f(x+h) + f(x+2h)$ by taking the weighted

SECTION CONTINUED ON NEXT PAGE

sum of the power series:

$$\begin{aligned}
-f(x-2h)+2f(x-h)-2f(x+h)+f(x+2h) = \\
&-f(x)+2hf'(x)-\frac{4h^2}{2}f''(x)+\frac{8h^3}{6}f^{(3)}(x)-\frac{16h^4}{24}f^{(4)}(x)\\
&+2f(x)-2hf'(x)+\frac{2h^2}{2}f''(x)-\frac{2h^3}{6}f^{(3)}(x)+\frac{2h^4}{24}f^{(4)}(x)\\
&-2f(x)-2hf'(x)-\frac{2h^2}{2}f''(x)-\frac{2h^3}{6}f^{(3)}(x)-\frac{2h^4}{24}f^{(4)}(x)\\
&+f(x)+2hf'(x)+\frac{4h^2}{2}f''(x)+\frac{8h^3}{6}f^{(3)}(x)+\frac{16h^4}{24}f^{(4)}(x)\\
&+\mathcal{O}(h^5)
\end{aligned}$$

Next gather like terms:

$$\begin{aligned}
-f(x-2h)+2f(x-h)-2f(x+h)+f(x+2h) = \\
&(-1+2-2+1)f(x)\\
&+(2-2-2+2)hf'(x)\\
&+(-2+1-1+2)h^2f''(x)\\
&+(4-1-1+4)\frac{h^3}{3}f^{(3)}(x)\\
&+(-2+1-1+2)\frac{h^4}{3}f^{(4)}(x)+\mathcal{O}(h^5)
\end{aligned}$$

Cancelling all the zero sums gives:

$$-f(x-2h)+2f(x-h)-2f(x+h)+f(x+2h) \;=\; 6\frac{h^3}{3}f^{(3)}(x)+\mathcal{O}(h^5)$$

4 marks for the manipulations.
Finally dividing through by $2h^3$ as in the original formula gives:

$$\frac{-f(x-2h)+2f(x-h)-2f(x+h)+2f(x+2h)}{2h^3} = f^{(3)}(x)+\mathcal{O}(h^2)$$

So $p = 2$.
2 marks for the final division.

END OF SECTION C