Scientific Computation Comprehensive Examination

Answer the following questions explaining all steps that lead to a solution. Partial credit will be awarded for presenting a viable solution strategy. No credit will be given to computations presented without motivation. Your goal is to present skill in formulating precise mathematical statements, and demonstrate understanding of theoretical material

1. Determine the best approximant of $f:[0,1]\to\mathbb{R}_+$, $f(x)=x^a$, $a>0$ by a constant c in the L_p norm

$$E_p(c)=\|f-c\|_p=\left(\int_0^1|f(x)-c|^p\,\mathrm{d}x\right)^{1/p},$$

for $p=1,2,\infty$. Determine E_p for each case.
Solution. $p=1$. From $x^a\in[0,1]$ deduce $c\in(0,1)$, let $\xi=c^{1/a}\in(0,1)$, note that f is monotone, and decompose integration domain

$$E_1(c)=\int_0^\xi(c-x^a)\,\mathrm{d}x+\int_\xi^1(x^a-c)\,\mathrm{d}x.$$

Evaluate integrals

$$\int_0^\xi(c-x^a)\,\mathrm{d}x=c\xi-\frac{\xi^{a+1}}{a+1},\int_\xi^1(x^a-c)\,\mathrm{d}x=c(\xi-1)+\frac{1-\xi^{a+1}}{a+1}.$$

Obtain L_1 error

$$E_1(c)=c(2c^{1/a}-1)+\frac{1-c^{(a+1)/a}}{a+1}.$$

Solve the stationarity condition $\partial E_1/\partial c=0$

$$\frac{2a+1}{a}c^{1/a}-1=0\Rightarrow c_*=\left(\frac{a}{2a+1}\right)^a,$$

and smallest L_1 error is $E_1(c_*)$.
$p=2$. The best L_2 approximant is obtained when $f-c$ is orthogonal to c,

$$(f-c,c)=\int_0^1(x^a-c)c\mathrm{d}x=0\Rightarrow c=\int_0^1x^a\,\mathrm{d}x=\frac{1}{a+1}.$$

The smallest L_2 error is

$$E_2^2=\left\|x^a-\frac{1}{a+1}\right\|_2^2=\int_0^1\left(x^a-\frac{1}{a+1}\right)^2=\frac{a^2}{(1+a)^2(2a+1)}\Rightarrow E_2=\frac{a}{a+1}\cdot\frac{1}{\sqrt{1+2a}}.$$

$p\to\infty$. In the inf-norm

$$c=\mathrm{argmin}_{b,x\in[0,1]}|x^a-b|,b\in[0,1].$$

Since x^a-b is monotone, extrema are attained at interval endpoints, $e_0(b)=|b|=b$, $e_1=|1-b|=1-b$, with $c=1/2$ in which case $E_\infty=1/2$.

2. Estimate the number of subintervals required to obtain $I = \int_0^1 \exp(-x^2)\,\mathrm{d}x$ to $k = 6$ correct decimal places through (a) the composite trapezoidal rule, and (b) the composite Simpson rule.
Solution. Assume a partition of $[0,1]$ with nodes $x_i = ih$, with uniform step size $h = 1/m$, $i = 0,\ldots,m$. Over subinterval $[x_{i-1}, x_i]$ of length h, the trapezoid rule is exact for linear functions, hence must have a truncation error

$$e_i = ah^3|f''(\xi_i)|.$$

The value of $f''(x)$ for $f(x) = \exp(-x^2)$, $x \in [0,1]$ is bounded, $|f''| < 2$, hence the overall error bound $e \leqslant me_i = 2a/m^2$. Since $I = \mathcal{O}(1)$, six correct decimal places are obtained if $e = 10^{-6} \Rightarrow m = 1000/\sqrt{2a}$. Simpson's rule is based upon quadratic interpolation of the integrand, but exhibits cancellation of cubic error terms leading to subinterval truncation error

$$e_i = bh^5|f^{(\mathrm{iv})}(\xi_i)|.$$

Using bound $|f^{(\mathrm{iv})}| \leqslant 12$, the overall error is $e \leqslant me_i = 12b/m^4 = 10^{-6} \Rightarrow m = 10^{1.5}/(12b)^{1/4}$.

3. Recall that a subset $\mathcal{S} \subseteq \mathcal{T}$ of a topological space $\mathcal{T}$ is *dense* if for any $x \in \mathcal{T}$, either $x \in \mathcal{S}$ or there exists some sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $x = \lim_{n\to\infty} x_n$. Apply the Schur decomposition $\boldsymbol{A} = \boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}$ to show that any matrix $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ can be written as the limit of diagonalizable matrices, i.e., the subset of diagonalizable matrices is dense.
Solution. The eigenvalues of $\boldsymbol{T}$ triangular are its diagonal elements $t_1,\ldots,t_m$, also the eigenvalues of $\boldsymbol{A}$ since they are similar matrices through $\boldsymbol{A} = \boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}$. The matrix $\boldsymbol{T}$ is diagonalizable for distinct eigenvalues. Suppose t_i is a repeated root with algebraic multiplicity n. Define sequences $s_{i+j} = t_i + j/k$, for $j = 0,1,\ldots,n-1$ and construct the matrices

$$\boldsymbol{S}_k = \boldsymbol{T} - \mathrm{diag}(0,\ldots,0,t_i,\ldots,t_{i+n-1},0,\ldots,0) + \mathrm{diag}(0,\ldots,0,s_i,\ldots,s_{i+n-1},0,\ldots,0)$$

The diagonal elements of $\boldsymbol{S}_k$ are distinct, hence $\boldsymbol{S}_k$ is diagonalizable, as is $\boldsymbol{A}_k = \boldsymbol{Q}^*\boldsymbol{S}_k\boldsymbol{Q}$, and $\lim_{k\to\infty}\boldsymbol{A}_k = \boldsymbol{A}$, and the subset of diagonalizable matrices is dense.

4. Consider the initial value problem $y'(t) = -10000\,(y(t) - \cos(t)) - \sin(t)$, $y(0) = 1$.

   a) Find the analytical solution and determine if this is stiff ODE system.
   Solution. Let $a = 10^4$. The homogeneous solution is $y(t) = ce^{-at}$, and variation of parameters, $y(t) = c(t)e^{-at}$, leads to

   $$c' = [a\cos(t) - \sin(t)]e^{at} \Rightarrow c(t) = \cos(t)\,e^{at} + C.$$

   Solution is $y(t) = \cos(t) + Ce^{-at}$, and initial condition $y(0) = 1 + C = 1$ gives $y(t) = \cos(t)$. The ODE does not exhibit disparate time scales and is not stiff in exact arithmetic.

   b) Find the analytical solution for a pertubed initial value to $y(0) = 1 + \varepsilon$, and reconsider whether the system is stiff.

Solution. The solution is $y(t)=\cos(t)+\varepsilon e^{-at}$, exhibits disparate time scales due to the rapid e^{-at} decay, is now stiff, and also indicates case (a) to be stiff in floating point arithmetic.

c) Determine time step constraints for applying: (i) forward Euler, and (ii) backward Euler to this system.
Solution. With $z=-ah$, $h>0$ the step size, forward Euler is stable for $|z+1|\leqslant 1, -1\leqslant 1-ah\leqslant 1 \Rightarrow h\leqslant 2/a=2\times 10^{-4}$. Backward Euler is stable for $|z-1|\geqslant 1$, $1+ah\geqslant 1$, $h\geqslant 0$, i.e., any step size, unconditional stability.

5. Consider Newton's method to build $\{x_n\}_{n\in\mathbb{N}}$, $x_n\to r$, r a root of $f:\mathbb{R}\to\mathbb{R}$, $f(r)=0$.

a) Prove that if r is a multiple root, convergence becomes first order.
Solution. Let $f_n=f(x_n)$, $f_n'=f'(x_n)$, $f_n''=f''(x_n)$, $e_n=x_n-r$, $\delta_n=x_{n+1}-x_n$. Newton's method $(x_{n+1}-x_n)f_n'+f_n=0=f(r)$, and the Taylor series of f are

$$\begin{aligned} f(r) &= f_n+\delta_n f_n' \\ f_{n+1} &= f_n+\delta_n f_n'+\frac{1}{2}\delta_n^2 f_n''+\mathcal{O}(\delta_n^3) \end{aligned}$$

Subtract to obtain

$$f_{n+1}-f(r)=\frac{1}{2}\delta_n^2 f_n''+\mathcal{O}(\delta_n^3),$$

and Taylor-expand f_{n+1} around r, using notation $f'=f'(r)$, $f''=f''(r)$,

$$f' e_{n+1}+\frac{1}{2}f'' e_{n+1}^2+\mathcal{O}(e_{n+1}^3)=\frac{1}{2}\delta_n^2 f_n''+\mathcal{O}(\delta_n^3).$$

Since $\delta_n=e_{n+1}-e_n$, and $e_n=\mathcal{O}(\delta_n)$, when $\{x_n\}_{n\in\mathbb{N}}$ is convergent (Cauchy sequence in complete metric space), obtain

$$f' e_{n+1}+\frac{1}{2}f'' e_{n+1}^2=\frac{1}{2}f_n''\cdot[e_{n+1}^2-2e_{n+1}e_n+e_n^2]+\mathcal{O}(e_n^3).$$

For $f\in C^2$, $f_n''=f''+\mathcal{O}(e_n)$, hence

$$f' e_{n+1}=-f'' e_{n+1}e_n+\frac{1}{2}f'' e_n^2+\mathcal{O}(e_n^3)\Rightarrow[f'+f''e_n]e_{n+1}=\frac{1}{2}f'' e_n^2+\mathcal{O}(e_n^3)\Rightarrow$$

$$e_{n+1}=\frac{1}{2}\frac{f''}{f'+f'' e_n}e_n^2+\mathcal{O}(e_n^3). \tag{1}$$

When $f''=f''(r)=0$ (local linear behavior), convergence is cubic. When $f'\neq 0$, $f''\neq 0$, convergence is quadratic. When $f'=0$, $f''\neq 0$

$$e_{n+1}=\frac{1}{2}e_n+\mathcal{O}(e_n^2), \tag{2}$$

i.e., convergence is linear.

b) Use Aitken extrapolation $a_n=x_n-(\Delta x_n)^2/\Delta^2 x_n$ to recover second-order convergence, with $\Delta x_n=x_{n+1}-x_n$. Write out this scheme as a pseudo code.

Solution. Use above notation to write $\delta_n = \Delta x_n$, $a_n = x_n - \delta_n^2/(\delta_{n+1} - \delta_n)$,

$$b_n = a_n - r = e_n - \frac{(e_{n+1} - e_n)^2}{e_{n+2} - 2e_{n+1} + e_n} = \frac{e_n\, e_{n+2} - e_{n+1}^2}{e_{n+2} - 2e_{n+1} + e_n}. \tag{3}$$

Use the $f'(r) = 0$ estimate (2) in (3)

$$b_n = \frac{\frac{1}{2} e_n\, e_{n+1} - e_{n+1}^2}{\frac{1}{2} e_{n+1}} + \mathcal{O}(e_n^2) = e_n - 2e_{n+1} + \mathcal{O}(e_n^2) = \mathcal{O}(e_n^2)$$

to recover second-order convergence.

$x_{n+1} = x_n - f(x_n)/f'(x_n)$; $\delta_n = x_{n+1} - x_n$; $a_n = 0$
repeat
  $f_{n+1} = f(x_{n+1})$; $f'_{n+1} = f'(x_{n+1})$; $a_{n-1} = a_n$
  $x_{n+2} = x_{n+1} - f_{n+1}/f'_{n+1}$; $\delta_{n+1} = x_{n+2} - x_{n+1}$; $a_n = x_n - \delta_n^2/(\delta_{n+1} - \delta_n)$
  $x_{n+1} = x_{n+2}$; $x_n = x_{n+1}$; $f_{n+1} = f(x_{n+1})$; $f'_{n+1} = f'(x_{n+1})$; $\delta_n = \delta_{n+1}$
until $|a_n - a_{n-1}| < \epsilon_{\text{mach}}$

c) Determine the order of the Aitken-extrapolated scheme if r is a simple root.
Solution. Let $c = \frac{1}{2} f''/f' \cong \frac{1}{2} f''/(f' + f'' e_n)$, and replace (1) into (3) to obtain

$$b_n \cong \frac{c e_n\, e_{n+1}^2 - e_{n+1}^2}{c e_{n+1}^2 - 2e_{n+1} + e_n} \cong \frac{c^2 e_n^3 - c^2 e_n^4}{c^2 e_n^4 - 2c e_n^2 + e_n} \cong c^2 e_n^2,$$

hence second-order convergence (no increased order of convergence w.r.t. Newton method).

6. a) Find the eigenvalues and eigenvectors of a circulant matrix (right-rotated rows)

$$\boldsymbol{C} = \begin{bmatrix} c_1 & c_2 & \dots & c_{n-1} & c_n \\ c_n & c_1 & \dots & c_{n-2} & c_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_3 & c_4 & \dots & c_1 & c_2 \\ c_2 & c_3 & \dots & c_n & c_1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Solution. Construction of $\boldsymbol{C}$ suggests use permutation matrices. Recall that $\boldsymbol{I} = [\, \boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \dots \;\; \boldsymbol{e}_n \,]$ is the identity permutation, a circular permutation by one index position is carried by $\boldsymbol{P} = [\, \boldsymbol{e}_2 \;\; \boldsymbol{e}_3 \;\; \dots \;\; \boldsymbol{e}_n \;\; \boldsymbol{e}_1 \,]$, and permutation by k index positions is carried out by $\boldsymbol{P}^k = [\, \boldsymbol{e}_{1+k} \;\; \boldsymbol{e}_{2+k} \;\; \dots \;\; \boldsymbol{e}_n \;\; \boldsymbol{e}_1 \;\; \dots \;\; \boldsymbol{e}_k \,]$. Rewrite the circulant matrix

$$\boldsymbol{C} = c_1 \boldsymbol{P}^0 + c_n \boldsymbol{P} + c_{n-1} \boldsymbol{P}^2 + \dots + c_2 \boldsymbol{P}^{n-1},$$

which can be re-expressed as

$$\boldsymbol{C} = a_0 \boldsymbol{P}^0 + a_1 \boldsymbol{P} + \dots + a_{n-1} \boldsymbol{P}^{n-1} = p(\boldsymbol{P}),$$

using $a_0 = c_1, a_1 = c_n, a_2 = c_{n-1}, \ldots, a_{n-1} = c_2$. Let $(\mu_i, \boldsymbol{y}_i)$ be the i^{th} eigenvalue, eigenvector pair of $\boldsymbol{P}$. Then

$$\boldsymbol{C}\boldsymbol{y}_i = p(\mu_i)\, \boldsymbol{y}_i$$

so $(\lambda_i = p(\mu_i), \boldsymbol{y}_i)$ are the eigenvalue, vectors of $\boldsymbol{C}$. The eigenvalues of $\boldsymbol{P}$ are solutions of

$$\det(\lambda \boldsymbol{I} - \boldsymbol{P}) = \lambda^n - 1,$$

the roots of unity, $\lambda_k = \exp(2\pi i k/n)$.

b) Write an efficient algorithm to compute $\boldsymbol{v} = \boldsymbol{C}\boldsymbol{u}$, $\boldsymbol{u} \in \mathbb{R}^n$, with respect to both storage and arithmetic operations.
Solution. Use $\boldsymbol{C} = c_1\boldsymbol{P}^0 + c_n\boldsymbol{P} + c_{n-1}\boldsymbol{P}^2 + \cdots + c_2\boldsymbol{P}^{n-1}$, to implement the matrix-vector product as repeated permutations

$$\boldsymbol{v} = a_0\boldsymbol{u} + a_1\boldsymbol{P}\boldsymbol{u} + \cdots + a_{n-1}\boldsymbol{P}^{n-1}\boldsymbol{u}$$

$\boldsymbol{v} = a_0\boldsymbol{u}$
for $k = 1$ to $n-1$
  $\boldsymbol{u} = \text{permute}(\boldsymbol{u})$
  $\boldsymbol{v} = \boldsymbol{v} + a_k\boldsymbol{u}$