Numerical Analysis Qualifier

prepared by
Chuck Gartland & Arden Ruttan
Kent State University
August, 1991

INSTRUCTIONS: Do any 10 of the following 12 problems.

1. (Floating Point Numbers) Consider a machine that uses normalized, 3-digit, base-10 floating-point numbers with exponent range \{-10, -9, \ldots, 9, 10\}.
   (a) What is the smallest positive machine-representable number?
   (b) What is the largest machine-representable number?
   (c) How many distinct positive numbers are representable on this machine?
   (d) What is the machine epsilon for this machine?

2. (Rounding-Error Analysis) Consider the evaluation in floating-point arithmetic of the finite sum
   \[ \sum_{k=0}^{n} x_k = x_0 + x_1 + \cdots + x_n \]
   from left to right. One can establish that
   \[ \text{fl} \left( \sum x_k \right) = \sum x^*_k \]
   where
   \[ x^*_k = x_k (1 + \rho_k)^{n-k+1}, \quad k = 0, \ldots, n, \]
   for some \( |\rho_0|, \ldots, |\rho_n| \leq \text{eps}. \)
   (a) Using this and the inequality
   \[ |(1 + \rho_1)^{\pm1} \cdots (1 + \rho_m)^{\pm1} - 1| \leq \frac{m \text{eps}}{1 - m \text{eps}} \]
   (\( |\rho_1|, \ldots, |\rho_m| \leq \text{eps}, m \text{ eps} < 1 \), derive a bound on the maximum relative error in \( \text{fl}(\sum x_k) \).
   (b) What conclusions can you draw from this for series with positive terms versus alternating series?
   (c) What conclusions can you draw from this for absolutely nonincreasing (\( |x_0| \geq |x_1| \geq \cdots \geq |x_n| \)) series?

3. (Mathematical and Numerical Conditioning) Suppose that you need to code up a function subprogram to evaluate (with high relative accuracy) the function
   \[ f(x) := \frac{x}{e^x - 1}, \]
   for any real \( x \).
(a) What is the \textit{mathematical condition number} of this function? For what ranges of $x$ is this evaluation well conditioned, ill conditioned?
(b) Give a mathematically equivalent formulation that is better suited for numerical computation when $x$ is small. Explain why.
(c) Sketch the algorithm that you would code up for your program.

4. (Interpolation)

(a) State the “classical problem of polynomial interpolation.”
(b) Write the general form of the interpolating polynomial with respect to the \textit{Lagrange} basis and with respect to the \textit{Newton} basis; define all component parts and notation.
(c) Give the error formula for polynomial interpolation and a bound for it that is valid when the data comes from a “sufficiently smooth” function.
(d) What are the \textit{Chebyshev abscissae}? What special property do they satisfy, and what is their relationship to the polynomial interpolation problem?

5. (Piecewise Polynomial Interpolation) A function $H_\Delta$ is in the class $\mathcal{H}_\Delta$ of “real Hermite cubic piecewise polynomials” (relative to a given partition $\Delta : a = x_0 < x_1 < \cdots < x_n = b$) if it satisfies

(a) $H_\Delta \in C^1[a, b]$, 
(b) $H_\Delta|_{[x_i, x_{i+1}]} \in \Pi_3$, $i = 0, \ldots, n - 1$,

that is, $H_\Delta$ is a continuously differentiable function that coincides with a polynomial of degree at most 3 in each subinterval.

(a) How do such functions differ from \textit{cubic splines}?
(b) The classical interpolation problem for $\mathcal{H}_\Delta$ is “given $(x_i, f_i, f'_i)$, $i = 0, \ldots, n$, find $H_\Delta \in \mathcal{H}_\Delta$ such that

$$H_\Delta(x_i) = f_i \quad \text{and} \quad H'_\Delta(x_i) = f'_i, \quad i = 0, \ldots, n.$$  

Prove that this problem has a unique solution.
(c) Supposing that the data come from a 4-times continuously differentiable function, i.e., $f_i = f(x_i)$ and $f'_i = f'(x_i)$, $i = 0, \ldots, n$, $( f \in C^4[a, b] )$, derive the error bound

$$\| f - H_\Delta(f) \|_\infty \leq \frac{1}{384} \| f^{(4)} \|_\infty \| \Delta \|^4, \quad \| \Delta \| := \max_{j=0,\ldots,n-1} |x_{j+1} - x_j|.$$ 

6. (Quadrature Rules)
(a) Given the basic Trapezoid Rule with error formula
\[ \int_a^b f(x) \, dx = \frac{(b-a)}{2} [f(a) + f(b)] - \frac{(b-a)^3}{12} f''(\xi), \]
derive the associated composite Trapezoid Rule with respect to a not-necessarily-uniform partition
\[ a = x_0 < x_1 < \cdots < x_n = b, \quad h_{i+1} := x_{i+1} - x_i. \]

(b) Assuming \( f \in C^2[a, b] \), derive the following bound on the error:
\[ \left| \int_a^b f - T_\Delta(f) \right| \leq \frac{1}{12} h^2 \| f'' \|_{\infty} (b - a), \quad h := \max_{i=1, \ldots, n} h_i. \]

7. (Orthogonal Polynomials) Let the polynomials \( \{P_n(x)\}, n = 0, 1, 2, \ldots \), be orthogonal over \([a, b]\) with respect to the non-negative weight function \( w(x) \). Prove that the roots \( x^n_j, j = 1, 2, \ldots, n \) of \( P_n(x) = 0 \), \( n = 1, 2, \ldots \), are simple and lie in \([a, b]\).

8. (Linear Systems)

(a) What is the difference between direct methods and iterative methods to solve general nonsingular linear algebraic systems of equations? Give examples of each, and describe situations in which one type of method might be preferred over the other.

(b) Derive the forward substitution algorithm to solve a nonsingular lower-triangular linear system. Exactly how many additions, subtractions, multiplications, and divisions does it require?

9. (Matrix Norms and Condition Numbers)

(a) Let \( \| \cdot \| \) be a norm on \( \mathbb{C}^n \). Define the matrix norm \( \| \cdot \|^* \) induced by \( \| \cdot \| \).

(b) Prove that any such induced matrix norm (or operator norm) is necessarily submultiplicative.

(c) Give a formula for the matrix norm induced by the vector infinity norm (max norm).

(d) Define the condition number of a matrix. What is its significance? Give a formula for the condition number with respect to the matrix 2-norm in terms of singular values.

10. (Nonlinear Equations) Let \( g(x) \) map the interval \([a, b]\) continuously into itself, and suppose that \( |g(x) - g(y)| \leq \gamma |x - y| \), for all \( x, y \in [a, b] \) and for some constant \( \gamma \) satisfying \( 0 \leq \gamma < 1 \). Show that \( g(x) \) has a unique fixed point \( \hat{x} \) in \([a, b]\) and that the fixed-point iteration \( x_{k+1} = g(x_k), k = 1, 2, \ldots \), converges to \( \hat{x} \) for any initial \( x_0 \in [a, b] \). If, in addition, \( g(x) \) is differentiable on \([a, b]\) and \( x_k \neq \hat{x}, k = 0, 1, 2, \ldots \), determine
\[ \lim_{k \to \infty} \frac{x_{k+1} - \hat{x}}{x_k - \hat{x}}. \]

Justify your answer.
11. (Least Squares Problems)

(a) Discuss how orthogonalization can be used to solve a full-rank, linear, least-squares problem (LS).

(b) Discuss the conditioning of LS.

(c) Discuss how the normal equations can be used to solve LS, and compare their conditioning with the orthogonalization approach.

12. (Matrix Normal Forms)

(a) Schur’s Theorem guarantees that any complex matrix is unitarily similar to an upper-triangular matrix. Use this to prove the Spectral Theorem for Normal Matrices: “A matrix is normal if and only if it is unitarily similar to a diagonal matrix.”

(b) It is true that for any square matrix $T$,

$$\lim_{k \to \infty} T^k = 0 \iff \rho(T) < 1.$$  

Here $\rho(\cdot)$ denote the spectral radius. Prove this in the case where $T$ is assumed to be normal.