Elementary Numerical Analysis (409.310A)

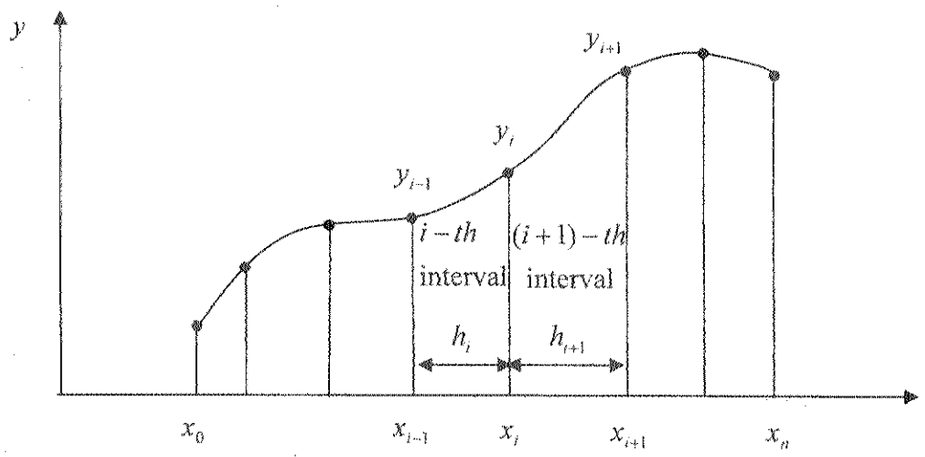
Mid Term Examination

10/26/05

1. Let a be a *positive* real number. You now want to convert this number into a binary floating point number using a finite number of bits for the mantissa and for the exponent. Answer the following: (15)

   1) How would you determine the mantissa first in *decimal* form and the corresponding exponent? The exponent here is of course the exponent of 2 which is to yield 2^n (n can be either positive or negative). Write your logic in words briefly.
   2) Let the mantissa given in *decimal* form be f. Suppose that you converted f into a binary number with m binary digits so that there could be a truncation error. Let the binary mantissa be represented by an array $b(1{:}m)$. $b(1)$ is the first binary number after the decimal point. Now give an expression to obtain the decimal number $\tilde{f}$ represented by b in terms of summation notation.
   3) Given f and $\tilde{f}$, how you would determine the number of significant digits? Answer with an inequality expression.

2. Consider the following setup for the cubic spline fitting. The cubic polynomial for the i-th interval is defined as $p_i(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3$, $x_{i-1} \le x \le x_i$; $h_i = x_i - x_{i-1}$. Answer the following: (20)



   1) The constant term above can be determined easily by the function value given at the left end point of the interval, namely, $a_i = y_{i-1}$, so that there are three unknown coefficients to be determined at each interval. Give i) another equation needed to satisfy the given function value, which should be specified for each interval, ii) and two more equations needed to satisfy the continuity conditions which should be specified for each interior point.
   2) Discuss about the number of unknowns and the number of equations available. What would you do resolve the deficiency?
   3) Obtain the derivative at the left end point using the first *three* data points and relate that with the coefficients of the first interval. Start your derivation with the Lagrange interpolation.
   4) You need to solve a system of linear equations to determine the coefficients. By using the solution vector defined as $u = [b_1, c_1, d_1, \cdots, b_i, c_i, d_i, \cdots, b_n, c_n, d_n]^T$, sketch the structure of the *linear system* (indicate the non-zero entries of the matrix and right hand side vector with x).
   5) In fact, the size of the linear system could be reduced by a factor of 1/3. How is it possible? Answer briefly.

3. The count rate of an radioactivity measurement diminishes exponentially with time, namely, $A(t) = A_0 e^{-\lambda t}$. In order to determine the decay constant (λ) of a radioactive isotope, a series of measurement was performed to yield N pairs of time vs. count rate (t_i, A_i) data. Derive the expression for the decay constant by using the least square method using the N data points. (10)

4. Answer the following as concisely as possible with only essential details. (20)

   1) Explain the minimax property of the Chebyshev polynomial with the two interpretations.
   2) Explain the way to determine the $(m+n+1)$ coefficients of the Pade approximation which is given by $f(x) \approx R_{m,n}(x) = \dfrac{\sum_{i=0}^{n} a_i x^i}{1 + \sum_{i=1}^{m} b_i x^i}$.
   3) Give that the finite difference approximation of the second order derivative using three points (equally spaced interval h) and show that the error of this approximation is in a second order using the Taylor series expansion.

5. Find the L, U factors of the following matrix obtained after Gauss elimination with *partial* pivoting.

$$A = \begin{bmatrix} 2 & 0 & 0 & -1 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ -1 & 1 & 4 & 1 \end{bmatrix}$$

. Note that $PA = LU$ where P is the permutation matrix. (15)

6. Explain the SOR method in the view of extrapolation and then derive the iteration matrix of the SOR method. (10)

7. Answer one of the following two problems. You can get a bonus point if you answer both. (10)

   1) Any real function defined on the interval $[a,b]$ can be approximated by a series of orthogonal functions. Namely,

$$f(x) = \sum_{i=0}^{n} a_i \phi_i(x) \quad \text{where} \quad a_i = \frac{< f, \phi_i >}{< \phi_i, \phi_i >} = \frac{\int_a^b w(x) f(x) \phi_i(x) dx}{\int_a^b w(x) \phi_i(x) \phi_i(x) dx}; \quad < \phi_i, \phi_j >= 0 \ \text{if} \ i \neq j .$$

   Show that this series expansion is equivalent to applying the least square method with these orthogonal functions. (Hint: Define the squared error norm as the integral involving the weight functions and squared error.) (10)

   2) How would you utilize the LU factor of a matrix to find inverse of the matrix? Assume that the LU factor was obtained without any pivoting. (Hint: Think $LUA^{-1} = I$ as n linear systems where $A^{-1} = [x_1, x_2, \cdots, x_n]$ is defined with column vector x_i.)