Preliminary Examination in Numerical Analysis, August 1999

Direction: There are three categories of problems. Answer one and only one problem from each category.

Category A

A1. In the following, A is a real symmetric positive definite $n \times n$ matrix, and lower case letters denote real vectors of length n.

1. Show that the problem of solving $Ax = b$ is equivalent to minimizing the functional $F(u)$ where
$$F(u) = \frac{1}{2}u^T Au - u^T b.$$
2. Let p and v be given fixed vectors. (Think of v as an approximation of the solution x, and p is a direction you wish to follow in order to improve the approximation.) Determine the value of α so that $r^T r$ is as small as possible where
$$r = Aw - b$$
and
$$w = v + \alpha p.$$
3. Suppose instead of minimizing $r^T r$, we wish to make $F(w)$ as small as possible where w has the same form as above. Determine the value of α that accomplishes this goal.
4. Compare the 2 strategies in parts (b) and (c). In particular, state which one is related to the conjugate gradient algorithm.

A2. Let $\lambda_1, \dots, \lambda_n$ denote the eigenvalues of an $n \times n$ matrix A.

1. Show that the trace and the determinant of A can be obtained from the following identities.
$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_j \quad \text{and} \quad \det(A) = \prod_{j=1}^{n} \lambda_j.$$
2. Suppose that A is diagonalizable. Assume $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$ and let $x_1, \dots x_n$ be the corresponding eigenvectors. Starting from v_0 with $v_0 \notin \mathrm{span}\{x_2, \dots, x_n\}$, show that the sequence from the power method
$$v_{k+1} := \frac{Av_k}{\|Av_k\|_2}, \quad k = 0, 1, 2, \dots,$$
is well-defined and that the sequence of Rayleigh quotients
$$R_k := \frac{\langle Av_k, v_k \rangle}{\|v_k\|_2}, \quad k = 0, 1, 2, \dots,$$
satisfies the estimate
$$|R_k - \lambda_1| \leq Cr^k, \quad k = 0, 1, 2, \dots,$$
for some constant $C > 0$ and $r := |\lambda_2/\lambda_1|$.

Category B

B1. Consider the problem

$$My'(t) = f(y(t)), \qquad t > 0 \tag{1}$$

with initial condition $y(0) = y_0$, where $y_0 \in R^m$ and M is a real $m \times m$-matrix, possibly singular; f is a smooth function.

1. Define the notion of Singular Value Decomposition (SVD) of M and write it $M = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T$ where each term has to be properly defined (Σ corresponds to the positive singular values).
2. Use the SVD to show that (1) can be written as

$$\begin{cases} x'(t) & = g(x, z), \\ 0 & = h(x, z). \end{cases} \tag{2}$$

   Explicitly define g and h.
3. Consider the regularized system ($\varepsilon > 0$)

$$\begin{cases} x'(t) & = g(x, z), \\ \varepsilon z'(t) & = h(x, z). \end{cases} \tag{3}$$

   Discretize (3) by a linearized Backward Euler method, i.e., apply one step of Newton, or the Chord method, to Backward Euler. Show that the algorithm can be written

$$J \begin{bmatrix} x^{n+1} - x^n \\ z^{n+1} - z^n \end{bmatrix} = \Delta t \begin{bmatrix} G(x^n, z^n) \\ H(x^n, z^n) \end{bmatrix}, \tag{4}$$

   Explicitly define J, G and H.
4. Deduce from this a method for solving (2), and thus (1). Justify the fact that for Δt small enough, the discretized problem is well posed.

B2. Given a function $f \in C[0, 2\pi]$,

1. Show that the best approximation to f in the L^2 norm with respect to the space of trigonometric polynomials, i.e., partial sums of the form

$$(\mathcal{F}_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx + \sum_{k=1}^{n} b_k \sin kx, \quad x \in [0, 2\pi]$$

   is given by the truncated Fourier series of f with the Fourier coefficients

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx dx, \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx dx. \end{aligned}$$

2. Briefly explain how this Fourier series is related to the Discrete Fourier Transform (DFT).
3. Briefly outline how the DFT can be implemented as the so called Fast Fourier Transform.

Category C

C1. 1. Define the following terms

   (a) floating point format
   (b) overflow
   (c) machine epsilon
   (d) catastrophic cancelation

2. Consider a toy format with base = 2, precision = 3, smallest value of the exponent = -1 and largest value of the exponent = 1. Sketch the position of all the corresponding floating point numbers on the real axis. Illustrate all the concepts of the previous question in the present format.

3. In terms of precision, is it better to use a small base, .e.g, 2, or a large one (IBM used 16 on some machines)? Comment.

C2. Concerning the numerical integration,

1. What is meant by a quadrature?

2. What is the fundamental difference between the Newton-Cotes quadrature and the Gaussian quadrature?

3. Explain how the Gaussian quadrature can be generated.

4. Consider the Fredholm integral equation of the second kind, i.e.,

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)dy = f(x), \quad x \in [a,b]. \tag{5}$$

   Suppose $\varphi(x)$ is represented by a set of discrete data $\varphi_n := [\varphi(x_1), \ldots, \varphi(x_n)]^T$ at quadrature points, i.e., abscissas, $x_1, \ldots, x_n \in [a,b]$. Describe how the solution to the integral equation (5) is approximated by the solution of a linear system

$$\varphi_n - A_n\varphi_n = f_n. \tag{6}$$

   Describe clearly what the coefficient matrix A_n is.