L. Vandenberghe
EE103

11/2/00

Midterm Exam

DO NOT OPEN UNTIL EVERYONE IS READY TO START

- You have time until 3:20.
- Only **this booklet** should be on your desk. You do not need a calculator.
- Write your answers neatly and concisely in the space provided after each question. You can use the blank pages on the left as scratch paper. Provide enough detail to convince us that you derived, not guessed, your answers.

Your name: [redacted]

Your student ID#: [redacted]

Your left neighbor's name: [redacted]

Your right neighbor's name: [redacted]

Problem 1	18 /25
Problem 2	15 /15
Problem 3	21 /30
Problem 4	30 /30
Total	84/100

Important formulas and definitions

Lecture 2. Accuracy of numerical algorithms

- Floating-point numbers with base 2

$$\pm(.d_1 d_2 \ldots d_n)_2 \cdot 2^e = \pm(d_1 2^{-1} + d_2 2^{-2} + \cdots + d_n 2^{-n}) \cdot 2^e$$

  with $d_1 = 1$, $d_i \in \{0, 1\}$
- Machine precision: $\epsilon_M = 2^{-n}$
- IEEE double precision arithmetic: $-1021 \leq e \leq 1024$, $n = 53$, $\epsilon_M \approx 1.11 \cdot 10^{-16}$

Lecture 3. Vectors and matrices

- Geometric interpretation of inner product: $x^T y = \|x\|\|y\| \cos \angle(x, y)$
- Number of flops for basic matrix and vector operations:
  - inner product $x^T y$ where $x, y \in \mathbf{R}^n$: $2n$ flops
  - vector addition $x + y$, scalar multiplication αx where $x, y \in \mathbf{R}^n$, $\alpha \in \mathbf{R}$: n flops
  - matrix-vector multiplication Ax where $A \in \mathbf{R}^{m \times n}$: $2mn$ flops
  - matrix-matrix multiplication AB where $A \in \mathbf{R}^{m \times p}$, $B \in \mathbf{R}^{p \times n}$: $2mnp$ flops

Lecture 5. The solution of a set of linear equations

- Definition of matrix norm: $\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$
- Properties of the matrix norm:

$$\|\alpha A\| = |\alpha|\|A\| \text{ for } \alpha \in \mathbf{R}$$
$$\|A\| \geq 0 \text{ for all } A; \; \|A\| = 0 \text{ iff } A = 0$$
$$\|A + B\| \leq \|A\| + \|B\|$$
$$\|Ax\| \leq \|A\|\|x\| \text{ for all } x \in \mathbf{R}^n$$
$$\|AB\| \leq \|A\|\|B\|$$
$$1/\|A^{-1}\| = \min_{x \neq 0}(\|Ax\|/\|x\|) \text{ if } A \text{ is square and nonsingular}$$
$$\|A\|\|A^{-1}\| \geq 1 \text{ if } A \text{ is square and nonsingular}$$

- Definition of condition number: $\kappa(A) = \|A\|\|A^{-1}\|$
- Error bounds for $Ax = b$, $A(x + \Delta x) = b + \Delta b$:

$$\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|, \qquad \frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|}$$

Lecture 6. Solving sets of linear equations

- cost of solving $Ax = b$ when $A \in \mathbf{R}^{n \times n}$ is upper or lower triangular: n^2 flops
- LU factorization with partial pivoting: $A = PLU$ (P a permutation matrix, L unit lower triangular, U upper triangular). Cost: $2n^3/3$ flops if $A \in \mathbf{R}^{n \times n}$

Problem 1. (25 points)

The following expressions are identical in exact arithmetic:

$$f(x) = \frac{\log(1+x)}{x}, \quad g(x) = \frac{\log(1+x)}{(1+x)-1}.$$

If we evaluate both functions in Matlab (using IEEE double precision arithmetic) at $x = 5 \cdot 10^{-16}$, we obtain

```
>>   log(1+5e-16)/5e-16
ans =
     0.8882

>>   log(1+5e-16)/((1+5e-16)-1)
ans =
     1.0000
```

The second result is much more accurate: $\log(1+x) \approx x$ for small x, so the result should be very close to 1. Explain both results (*i.e.*, the two numerical values 0.8882 and 1.0000).

Remarks.

- You can assume that the machine calculates $\log y$ exactly for any floating-point number y, and then rounds the result to the nearest floating-point number.
- You can use the approximation $\log y \approx y - 1$ for $y \approx 1$.

Answer for problem 1.

In the first method, when it calculates log(1+x), as stated it gets ≈ 5e-16, but this is rounded down ≈ 4e-16 the closest floating point number. It then takes ≈4e-16 / 5e-16 which approximately equals .8882.

In the second method, It does the same for the numerator and gets ≈ 4e-16. But for the denominator it first evaluates (1+5e-16) which is rounded down to ≈ 1+4e-16. Then it subtracts 1, so the end result is ≈4e-16 / ≈4e-16 which is approximately 1.

Problem 2. (15 points)

Express the following problem as a set of linear equations. Find a rational function

$$f(t) = \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2}$$

that satisfies the following conditions

$$f(1) = 2.3, \quad f(2) = 4.8, \quad f(3) = 8.9, \quad f(4) = 16.9, \quad f(5) = 41.0,$$

The variables in the problem are the coefficients c_0, c_1, c_2, d_1 and d_2.
Write the equations in matrix-vector form $Ax = b$.

Remarks.

- You can assume that there is a unique solution, and that the denominator $1+d_1t+d_2t^2$ is nonzero at $t = 1, 2, 3, 4, 5$.
- You don't have to solve the set of linear equations you obtain, and you don't have to show that the coefficient matrix A is nonsingular.

Answer for problem 2.

$c_0 + c_1 + c_2 = 2.3 + 2.3d_1 + 2.3d_2$

$f(1) = c_0 + c_1 + c_2 = 2.3 \rightarrow c_0 + c_1 + c_2 - 2.3d_1 - 2.3d_2 = 2.3$

$f(2) = \frac{c_0 + 2c_1 + 4c_2}{1 + 2d_1 + 4d_2} = 4.8$ $\quad c_0 + 2c_1 + 4c_2 = 4.8 + 9.6d_1 + 19.2d_2$

$c_0 + 2c_1 + 4c_2 - 9.6d_1 - 19.2d_2 = 4.8$

$f(3) = \frac{c_0 + 3c_1 + 9c_2}{1 + 3d_1 + 9d_2} = 8.9$ $\quad \ldots = 8.9 + 26.7d_1 + 80.1d_2$

$f(4) = \frac{c_0 + 4c_1 + 16c_2}{1 + 4d_1 + 16d_2} = 16.9$ $\quad c_0 + 4c_1 + 16c_2 = 16.9 + 67.6d_1 + 270.4d_2$

$c_0 + 4c_1 + 16c_2 - 67.6d_1 - 270.4d_2 = 16.9$

$f(5) = \frac{c_0 + 5c_1 + 25c_2}{1 + 5d_1 + 25d_2} = 41.0$ $\quad c_0 + 5c_1 + 25c_2 = 41.0 + 205d_1 + 1025d_2$

$c_0 + 5c_1 + 25c_2 - 205d_1 - 1025d_2 = 41.0$

$$\begin{bmatrix} 1 & 1 & 1 & -2.3 & -2.3 \\ 1 & 2 & 4 & -9.6 & -19.2 \\ 1 & 3 & 9 & -26.7 & -80.1 \\ 1 & 4 & 16 & -67.6 & -270.4 \\ 1 & 5 & 25 & -205 & -1025 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 2.3 \\ 4.8 \\ 8.9 \\ 16.9 \\ 41.0 \end{bmatrix}$$

3

Problem 3. (30 points) 2

$$A = \begin{bmatrix} 0 & 0 & -10^4 & 0 \\ 0 & 0 & 0 & -10 \\ 0 & 10^{-3} & 0 & 0 \\ 10^{-2} & 0 & 0 & 0 \end{bmatrix}.$$

1. *(10 points)* What is the norm of A?
2. *(10 points)* What is the inverse of A?
3. *(5 points)* What is the norm of the inverse of A?
4. *(5 points)* What is the condition number of A?

Explain your answers.

Answer for problem 3.

This vector produces the largest value of Ax

1) $\|A\|$ max $\frac{\|Ax\|}{\|x\|}$ $x = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$ $Ax = \begin{bmatrix} 10^4 \\ 10 \\ 10^{-3} \\ 10^{-2} \end{bmatrix}$

$\|A\| = \frac{\sqrt{10^{4\,2} + 10^2 + 10^{-3\,2} + 10^{-2\,2}}}{\sqrt{1+1+1+1}} = \frac{\sqrt{10^8 + 10^2 + \ldots}}{\sqrt{4}} = \frac{1}{2}\sqrt{\ldots}$

2) $A^{-1}A = I$

$A^{-1} = \begin{bmatrix} 0 & 0 & 0 & 10^2 \\ 0 & 0 & 10^3 & 0 \\ -10^{-4} & 0 & 0 & 0 \\ 0 & -10^{-1} & 0 & 0 \end{bmatrix}$

3) $\|A\|$ max $\frac{\|Ax\|}{\|x\|}$ $x = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$ $Ax = \begin{bmatrix} 10^2 \\ 10^3 \\ 10^{-4} \\ 10^{-1} \end{bmatrix}$

$\|A\| = \frac{\sqrt{10^{2\,2} + 10^{3\,2} + \ldots}}{\sqrt{\ldots}} = \frac{\sqrt{10^4 + 10^6 + 10^{-8} + 10^{-2}}}{\sqrt{4}} = \frac{1}{2}\sqrt{10^4 + 10^6 + 10^{\ldots}}$

Answer for problem 3 (continued).

$$\frac{1}{4}\left(10^{12} + 10^{14} \ldots 1 + 10^{6} + 10^{6} \ldots\right)$$

Problem 4. (30 points)

Assume $A \in \mathbf{R}^{n \times n}$ is a nonsingular matrix. Consider the matrix $M \in \mathbf{R}^{2n \times 2n}$ defined as

$$M = \begin{bmatrix} A & A + A^{-T} \\ A & A \end{bmatrix}. \tag{1}$$

(A^{-T} stands for the inverse of the transpose of A, or equivalenty, the transpose of the inverse of A.)

1. *(10 points)* Show that the inverse of M is given by

$$M^{-1} = \begin{bmatrix} -A^T & A^{-1} + A^T \\ A^T & -A^T \end{bmatrix}. \tag{2}$$

2. *(10 points)* Compare the cost (number of flops for large n) of the following two methods for solving a set of linear equations $Mx = b$, given A and b.

   Method 1. Calculate A^{-1}, build the matrix M as defined in equation (1), and solve $Mx = b$ using Gaussian elimination with partial pivoting (GEPP). (This method would correspond to the Matlab code `x = [A  A+inv(A)'; A A]\b;`)

   Method 2. Calculate A^{-1}, build the matrix M^{-1} as defined in equation (2), and form the matrix vector product $x = M^{-1}b$. (This method would correspond to the Matlab code `x = [-A' inv(A)+A'; A'  -A']*b;`)

   You can assume that A is a dense matrix.

3. *(10 points)* Can you improve the fastest of the two methods described in part 2? (You can state your answer in the form of an improved version of the Matlab code given in part 2, but it is not necessary, as long as you make the steps in your method very clear.)

Answer for problem 4.

1) $M M^{-1} = I$

$$\begin{bmatrix} A & A+A^{-T} \\ A & A \end{bmatrix} \begin{bmatrix} -A^T & A^{-1}+A^T \\ A^T & -A^T \end{bmatrix} = \begin{bmatrix} -AA^T + AA^T + A^{-T}A^T & AA^{-1} + AA^T - AA^T - A^{-T}A^T \\ -AA^T + AA^T & AA^{-1} + AA^T - AA^T \end{bmatrix}$$

$$= \begin{bmatrix} (AA^{-1})^T & I - (AA^{-1})^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

Answer for problem 4 (continued).

Method 1

calculate A^{-1} $\frac{8n^3}{3}$ flops

$A + A^{-1}$ n^2 flops

$M \backslash b$ $\frac{2(2n)^3}{3}$ = $\frac{16n^3}{3}$

$8n^3$

Method 2

calculate A^{-1} $\frac{8n^3}{3}$

$A^{-1} + A^T$ n^2

$M^{-1} b$ $8n^2$ flops

$\approx \frac{8n^3}{3}$ flops

3) $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -A^T & A^{-1}+A^T \\ A^T & -A^T \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

$x \in R^{2n}$ $x_i \in R^n$
$b \in R^{2n}$ $b_i \in R^n$

$x_1 = -A^T b_1 + A^{-1} b_2 + A^T b_2$

$x_2 = A^T b_1 - A^T b_2$

$\bar{x} = x_1 + x_2 = A^{-1} b_2$

$A\bar{x} = b_2$

$\bar{x} = A \backslash b_2$

(1) $\bar{x} = A \backslash b_2$
(2) $x_2 = A^T b_1 - A^T b_2$
(3) $x_1 = \bar{x} - x_2$

Great!