

CAAM 453 · NUMERICAL ANALYSIS I

Examination 1

Posted (corrected) 22 October 2007.

Due no later than 5pm on Saturday, 27 October 2007.

Instructions:

1. Time limit: **4 uninterrupted hours**.
2. There are four questions worth a total of 100 points.
Please do not look at the questions until you begin the exam.
3. You *may not* use any outside resources, such as books, notes, problem sets, friends, calculators, or MATLAB.
4. Please answer the questions thoroughly and justify all your answers.
Show all your work to maximize partial credit.
5. Print your name on the line below:

6. Time started: _____ Time completed: _____

7. Indicate that this is your own individual effort in compliance with the instructions above and the honor system by writing out in full and signing the traditional pledge on the lines below.

8. Staple this page to the front of your exam.

1. [25 points: (a)=4 points; (b)=4 points; (c)=8 points; (d)=6 points; (e)=3 points]

For this problem, let $x_j = jh$, where $h > 0$ and $j = 0, 1, 2, \dots$, and suppose that f is sufficiently smooth, i.e., has as many continuous derivatives as needed. You may use the shorthand f_j for $f(x_j)$. *Please work out all constants explicitly.*

- (a) Give the Newton form of the polynomial $p_1 \in \mathcal{P}_1$ that interpolates f at x_0 and x_1 .
- (b) Give the Newton form of the polynomial $p_2 \in \mathcal{P}_2$ that interpolates f at x_0 , x_1 , and x_2 .
- (c) One can approximate derivatives of $f(x)$ by derivatives of interpolating polynomials at the same x . Compute formulas for $p'_1(x_0)$, $p'_2(x_0)$, $p'_2(x_1)$, and $p''_2(x_1)$.

(Approximations of this sort are fundamental to the solution of partial differential equation by finite difference methods.)

- (d) In general, approximations obtained from degree- k interpolants have accuracy that behaves like h^k as $h \rightarrow 0$. This suggests that we could continue this procedure to obtain still better approximations to such derivatives from higher degree interpolating polynomials. Is this advisable? Explain.
- (e) Use the Taylor expansion

$$f(x_j) = f(x_0) + jhf'(x_0) + \frac{1}{2}(jh)^2 f''(x_0) + \frac{1}{3!}(jh)^3 f'''(x_0) + \dots$$

and your solution from (c) to confirm that

$$p'_2(x_0) - f'(x_0) = -\frac{h^2}{3} f'''(x_0) + \dots$$

2. [25 points: (a)=5 points; (b)=5 points; (b)=9 points; (c)=6 points]

(a) Confirm that the 2×2 Givens rotation

$$\mathbf{G}(1, 2, \theta) = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

is a unitary matrix.

(b) Calculate the number of arithmetic operations required to multiply $\mathbf{G}(1, 2, \theta)$ against a matrix $\mathbf{B} \in \mathbb{R}^{2 \times p}$.

(c) A matrix \mathbf{A} is *upper Hessenberg* if all entries below the first subdiagonal are zero, i.e., if $a_{jk} = 0$ for $j > k + 1$. For example, a 5×5 upper Hessenberg matrix has the form

$$\mathbf{A} = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix},$$

where \times denotes an arbitrary nonzero entry.

Explain how one can use Givens rotations to *efficiently* compute the QR factorization of an n -by- n upper Hessenberg \mathbf{A} . (You do not need to address the choice of angle, θ .)

(Recall that a Givens rotation $G(j, k, \theta)$ equals the identity matrix, except for the j th and k th rows and columns, which are zero except that the (j, j) , (j, k) , (k, j) , and (k, k) entries follow the model in part (a).)

(d) Roughly estimate the number of arithmetic operations required by the algorithm you proposed in part (b). How does this compare to the operation count of the standard Householder QR algorithm?

3. [24 points: (a)=8 points; (b)=8 points; (c)=8 points]

For this problem, suppose that $\mathbf{A} \in \mathbb{C}^{n \times n}$ is invertible.

- (a) Use the singular value decomposition to describe nonzero vectors \mathbf{y} and \mathbf{z} such that

$$\|\mathbf{A}\mathbf{y}\|_2 = \|\mathbf{A}\|_2\|\mathbf{y}\|_2$$

and

$$\|\mathbf{A}^{-1}\mathbf{z}\|_2 = \|\mathbf{A}^{-1}\|_2\|\mathbf{z}\|_2.$$

- (b) Suppose that $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$. Show that the perturbation $\delta\mathbf{x}$ in the solution induced by the perturbation $\delta\mathbf{b}$ in the right hand side can be bounded by

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 \frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}.$$

Hint: think about $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{x}\|_2$.

- (c) Describe how to construct vectors \mathbf{b} and $\delta\mathbf{b}$, each of arbitrary norm, such that equality holds in the bound (b).

4. [26 points: (a)=12 points; (b)=8 points; (c)=4 points; (d) = 2 points]

Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ be a full rank matrix, with $m > n$.

- (a) Consider the standard least squares problem

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{Ax} - \mathbf{b}\|_2.$$

- Write down the normal equations.
- Show that the optimal \mathbf{x} satisfies $\mathbf{Ax} = \mathbf{\Pi b}$, where $\mathbf{\Pi} = \mathbf{A}(\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^*$.
- Write the optimal \mathbf{x} in terms of the singular value decomposition of \mathbf{A} .

In general, $\mathbf{Ax} \neq \mathbf{b}$ for all $\mathbf{x} \in \mathbb{C}^n$. Problem (a) indicates that

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{Ax} - \mathbf{b}\|_2 = \|\mathbf{b} - \mathbf{\Pi b}\|_2 = \min_{\widehat{\mathbf{b}} \in \text{Ran}(\mathbf{A})} \|\mathbf{b} - \widehat{\mathbf{b}}\|_2.$$

Thus, the standard least squares problem seeks the smallest perturbation $\delta \mathbf{b}$ such that there exists some \mathbf{x} for which $\mathbf{Ax} = \mathbf{b} + \delta \mathbf{b}$. Implicitly, we are thus assuming that the matrix \mathbf{A} is exact, but the data \mathbf{b} has some errors.

An alternative approach, called *total least squares*, allows for errors in both \mathbf{A} and \mathbf{b} . Now we look for the smallest $\delta \mathbf{A}$ and $\delta \mathbf{b}$ such that there exists some \mathbf{x} for which $(\mathbf{A} + \delta \mathbf{A})\mathbf{x} = \mathbf{b} + \delta \mathbf{b}$, i.e.,

$$[\mathbf{A} + \delta \mathbf{A} \quad \mathbf{b} + \delta \mathbf{b}] \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \quad (*)$$

This equation implies that the matrix $[\mathbf{A} + \delta \mathbf{A} \quad \mathbf{b} + \delta \mathbf{b}] \in \mathbb{C}^{m \times (n+1)}$ has rank less than $n + 1$. (Recall that $m > n$.)

- (b) Use the singular value decomposition of the matrix $[\mathbf{A} \quad \mathbf{b}]$ to describe how to compute the matrix $[\delta \mathbf{A} \quad \delta \mathbf{b}]$ that makes $[\mathbf{A} + \delta \mathbf{A} \quad \mathbf{b} + \delta \mathbf{b}]$ rank-deficient and minimizes $\|[\delta \mathbf{A} \quad \delta \mathbf{b}]\|_2$.
- (c) Use the optimal $[\delta \mathbf{A} \quad \delta \mathbf{b}]$ in (b) to write a simple formula for the solution \mathbf{x} in (*) in terms of appropriate singular values and/or vectors of $[\mathbf{A} \quad \mathbf{b}]$.
- (d) Explain why $\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$ cannot be smaller than the smallest singular value of $[\mathbf{A} \quad \mathbf{b}]$.