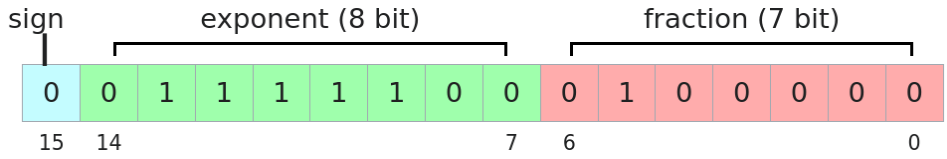
Math 466/666: Sample Theoretical Midterm Version E

1. The `bfloat16` brain floating-point format consists of a sign bit, an 8-bit exponent followed by a 7-bit mantissa for a total of 16 bits.



The sign bit is set for a negative number, the exponent is stored as a biased integer with the offset being 127 and the mantissa is stored with a leading implicit bit similar to classical floating-point formats.

Find the decimal equivalent of 1 10000010 1100100.

negative. exponent fraction

$\frac{1}{4} = .25$
$\frac{1}{8} = .125$
$\frac{1}{16} = .0625$
$\frac{1}{32} = .03125$

128 64 32 16 8 4 2 1
1 0 0 0 0 0 1 0 = 128 + 2 = 130

$e = 130 - 127 = 3$

$1.1100100 = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{32} = 1 + .5 + .25 + .03125 = 1.78125$

$x = -1.78125 \times 2^3 = -14.25$

1.75
.03125
6 5 2 8
1 78 125
× 8
14.25000

2. Why should the subtraction of two nearly equal numbers be avoided?

It leads to loss of precision due to cancellation of significant digits.

Example:
2.518
−2.513
.005 ← 1 sig digit left

started with 4 sig digits

Midterm x2	50 points each
10 Computing Labs	10 points each
Project	20 points
6 Homeworks	5 points each
Final	100 points
	300 points total

Never mind! Only one midterm because of all the labs. 250

3. Find the relative condition number $\kappa_f(x) = |xf'(x)|/|f(x)|$ when $f(x) = \sqrt{x+5}$ and $x = 4$, and explain what it means.

$f'(x) = \frac{1}{2\sqrt{x+5}}$ $\quad \kappa_f(x) = \left| x \frac{1/2\sqrt{x+5}}{\sqrt{x+5}} \right| = \left| \frac{x}{2(x+5)} \right|$

$\kappa_f(4) = \frac{4}{2 \cdot 9} = \frac{2}{9}$.

since $\kappa_f(4) < 1$ then errors input to $f(x)$ when x is near 4 get smaller.

4. Let $A \in \mathbf{R}^{n\times n}$. Define for A to be a strictly diagonally dominant matrix by rows.

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \qquad \text{for all } i = 1, \dots, n.$$

5. State the Jacobi method for solving $Ax = b$. It's a splitting method where $A = M - N$ and M is the diagonal of A. Then $M = \begin{bmatrix} a_{11} & & 0 \\ & a_{22} & \\ 0 & & \ddots \; a_{nn} \end{bmatrix}$ and $N = M - A$. Then set $x^{(0)} = 0$ or some other starting value and iterate $M x^{(k+1)} = N x^{(k)} + b$ to find a sequence $x^{(k)}$. If $x^{(k)}$ converges then it converges to a solution of $Ax = b$.

6. Consider the matrix A and Givens rotation R given by

$$A = \begin{bmatrix} -5 & 2 \\ 12 & -4 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Find the rotation R such that RA is upper triangular. You need not find θ or the resulting upper triagular matrix.

$r_1 = -5 \qquad r_2 = 12 \qquad \rho = \sqrt{r_1^2 + r_2^2} = \sqrt{25 + 144} = \sqrt{169} = 13$

$\cos\theta = \frac{r_1}{\rho} = \frac{-5}{13} \qquad \sin\theta = \frac{r_2}{13} = \frac{12}{13}$

$$R = \begin{bmatrix} -5/13 & 12/13 \\ -12/13 & -5/13 \end{bmatrix}.$$

7. Given a vector norm $\|\cdot\|_p$ and an $m \times n$ matrix A, state the definition of the induced matrix norm $\|A\|_p$.

$$\|A\|_p = \max\{\|Ax\|_p : \|x\|_p = 1\}$$

alternatively...

$$\|A\|_p = \sup\left\{\frac{\|Ax\|_p}{\|x\|_p} : x \in \mathbb{R}^n \text{ and } x \neq 0\right\}$$

8. Suppose $\tilde{x}$ is an approximation of the solution to $Ax = b$. Let $\tilde{b} = A\tilde{x}$ and consider relative residual error $\|\tilde{b} - b\|/\|b\|$ obtained by plugging the approximate solution into the original equation. Show that the relative error in $\tilde{x}$ satisfies

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A)\frac{\|\tilde{b} - b\|}{\|b\|} \qquad \text{where} \qquad \kappa(A) = \|A^{-1}\|\|A\|.$$

Comment why the above bound holds for any vector norm as long as the matrix norms used in the definition $\kappa(A)$ are compatible with that vector norm.

Note $\tilde{x} = A^{-1}\tilde{b}$ and $x = A^{-1}b$ so

$$\|\tilde{x} - x\| = \|A^{-1}\tilde{b} - A^{-1}b\| = \|A^{-1}(\tilde{b} - b)\| \leq \|A^{-1}\|\,\|\tilde{b} - b\|.$$

$\|x\| = \|A^{-1}b\|$

$$\|b\| = \|Ax\| \leq \|A\|\,\|x\| \quad \text{or} \quad \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

Now combining the two inequalities

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \|A^{-1}\|\,\|\tilde{b} - b\| \cdot \frac{\|A\|}{\|b\|} = \kappa(A)\frac{\|\tilde{b} - b\|}{\|b\|}$$

The reason it works for any vector norm is because the only thing we used was the matrix norm compatibility. First in $\|A^{-1}(\tilde{b} - b)\| \leq \|A^{-1}\|\,\|\tilde{b} - b\|$ and then in $\|Ax\| \leq \|A\|\,\|x\|$.

9. Given a matrix $A \in \mathbf{R}^{m\times n}$ consider the matrix norms

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |A_{ij}| \quad \text{and} \quad \|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |A_{ij}|$$

Let

$$A = \begin{bmatrix} 8 & -7 & -1 & -2 \\ -4 & -9 & 4 & -8 \\ 4 & -6 & -9 & 1 \end{bmatrix}.$$

18
25
20

16 22 14 11

(i) Find $\|A\|_\infty$

$\|A\|_\infty = \max(18, 25, 20) = 25$

(ii) Find $\|A\|_1$

$\|A\|_1 = \max(16, 22, 14, 11) = 22.$

(iii) In general is it true or false that $\|A\|_\infty = \|A^T\|_1$. If true, explain why. If false, provide a counterexample.

True. Because transpose switches rows for columns, so $\|A\|_\infty$ the maximum absolute sum along a row is the same as $\|A^T\|_1$ the maximum absolute sum of A^T along a column.

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |A_{ij}| = \max_{1\le i\le m} \sum_{j=1}^{n} \left|[A^T]_{ji}\right| = \max_{1\le j\le m} \sum_{i=1}^{n} \left|[A^T]_{ij}\right| = \|A^T\|_1$$

along a row

along column

10. Prove only one of the following:

Theorem. Let f be $n+1$ times continuously differentiable and p the interpolating polynomial of degree n or less passing through the distinct points

$$(x_0, f(x_0)), \quad (x_1, f(x_1)), \quad \ldots, \quad (x_n, f(x_n)).$$

Then

$$f(x) = p(x) + \frac{(x-x_0)(x-x_1)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(c)$$

for some c between $\min(x, x_0, \ldots, x_n)$ and $\max(x, x_0, \ldots, x_n)$.

Theorem. Let $A \in \mathbf{R}^{m\times n}$ and define

$$\|A\|_2 = \max\{\, \|Ax\|_2 : \|x\|_2 = 1 \,\}.$$

Then $\|A\|_2 = \max\{\, \sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^TA \,\}$.

$\|Ax\|_2^2 = Ax \cdot Ax = x \cdot A^TA\, x$ Let $B = A^TA$. Then by the spectral theorem there is an orthonormal basis of eigenvectors of B such that $Bu_i = \lambda_i u_i$ $\sum c_i u_i$ $u_i \cdot u_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases}$

Let $x = \sum c_i u_i$ then

$$\|Ax\|_2^2 = \sum_i c_i u_i \cdot \sum_j c_j B u_j = \sum_{ij} c_i c_j \lambda_j u_i u_j \overset{\text{by orthogonality}}{=} \sum_{i=1}^{n} c_i^2 \lambda_i$$

also

$$\|x\|_2^2 = \sum c_i u_i \cdot \sum c_j u_j = \sum c_i^2 = 1$$

view this as a weighted average of the eigenvalues λ_i

Since AVG $\leq$ Max then $\|Ax\|_2^2 \leq \max\{\lambda_1, \ldots, \lambda_n\}$.

Note that $B = A^TA$ is positive (semi) definite so $\lambda_i \geq 0$ thus we can take square roots so

$$\|A\|_2 \leq \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^TA\}.$$

Finally setting $x = u_i$ where λ_i is the largest eigenvalue shows equality so $\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^TA\}$.

11. [Extra Credit and for Math 666] Suppose $A \in \mathbf{R}^{n \times n}$ is a symmetric positive definite matrix. Show the Gauss-Seidel method for solving $Ax = b$ converges.