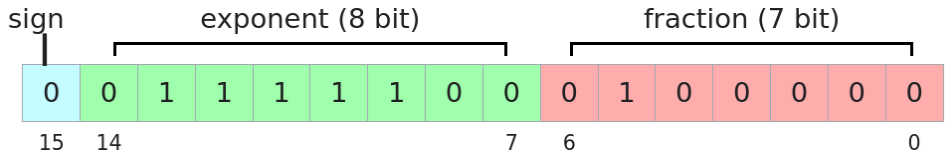
Math 466/666: Sample Theoretical Midterm Version E

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`.

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

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.

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

5. State the Jacobi method for solving $Ax = b$.

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

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

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

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

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.

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

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

Let

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

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

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

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

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

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

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

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.