Numerical Methods

CSE Written Qualifying Exam

Spring 2026

Instructions

- This is a **CLOSED BOOK** exam. No books or notes are allowed.
- No calculators, computers, phones, or internet usage allowed at any time during the exam (except for purposes of electronic proctoring, e.g., Honorlock).
- Answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total.
- Show all your work and write in a readable way. Points will be awarded for correctness as well as clarity.
- Good luck!

1 Problem 1

Let `x[1:n]` be a floating-point array of length `n`. Consider the natural linear-time sequential algorithm to compute its sum:

```
s = 0.0
for i = 1 up to n
    s = s + x[i]
end for
return s
```

(a) Let x be the vector of input values and $s(x) = \sum_{i=1}^{n} x_i$ be the true sum in exact arithmetic. Assume x is representable exactly in floating-point. Next, suppose the algorithm returns a computed sum, $\hat{s}(x)$.

Assume the standard model of floating-point arithmetic where ϵ is machine epsilon and the result of a floating-point addition, $\texttt{fl}(a+b)$, satisfies $\texttt{fl}(a+b) = (a+b)(1+\delta)$ for some $\delta = \delta(a,b)$ that depends on a and b and where $|\delta| \leq \epsilon$. Show that

$$|\hat{s}(x) - s(x)| \lesssim \big(n|x_1| + (n-1)|x_2| + (n-2)|x_3| + \cdots + |x_n|\big)\epsilon. \qquad (1)$$

(b) Assuming the above statement is true, propose a modification to the algorithm that can reduce the effect of rounding errors when summing a large number of floating-point numbers. Explain how your method works and be sure to compare its complexity to the original algorithm.

2 Problem 2

Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and $b \in \mathbb{R}^m$. Consider linear least squares problems of the form

$$\min_x \|Ax - b\|_2.$$

Answer the following questions.

(1) Derive the normal equations for the least squares problem. Explain why this approach can be numerically unstable for ill-conditioned matrices.

(2) Explain the difference between *thin* (economy) QR and *full* QR factorizations. In which contexts is the thin QR sufficient for least squares?

(3) Define the numerical rank of a matrix. Explain how it can be estimated using either QR with pivoting or the SVD.

(4) Show that among all least squares solutions, the SVD-based solution has minimum Euclidean norm. State the key property that guarantees this result.

(5) Suppose A is nearly rank-deficient. Describe a practical regularization strategy based on the SVD, and explain its effect on the solution.

3 Problem 3

Let $A \in \mathbb{R}^{n \times n}$ be nonsingular and $b \in \mathbb{R}^n$. Consider the Richardson iterative method for solving the linear system $Ax = b$:

$$x^{(k+1)} = x^{(k)} + \omega\big(b - Ax^{(k)}\big),$$

where $\omega > 0$.

1. Show that this iteration can be written in the fixed-point form

   $$x^{(k+1)} = Gx^{(k)} + c,$$

   and identify the iteration matrix G. Show that the method converges for any initial guess if and only if $\rho(G) < 1$. (You may cite the standard convergence result for stationary linear iterations.)

2. Assume that all eigenvalues of A have strictly positive real parts. Show that there exists a real $\omega > 0$ such that the Richardson iteration converges. Give a convergence condition on ω in terms of the spectrum $\sigma(A)$, and argue why such an ω exists under the given assumption.

3. Now assume that A is symmetric positive definite.

   (a) Determine the range of ω for which the method converges.

   (b) Find the value of ω that minimizes the asymptotic convergence factor $\rho(G)$.

   (c) Express the resulting convergence rate in terms of the condition number $\kappa_2(A)$.

4. Assume A is strictly diagonally dominant (so Jacobi iteration converges), and choose ω so that Richardson also converges. Explain **qualitatively** which of the two methods (Jacobi or Richardson) you would expect to converge faster or be more robust for a diagonally dominant system. Justify your answer using the facts that (i) diagonal dominance makes the off-diagonal-to-diagonal ratios small, and (ii) Richardson uses a single global scalar ω to scale all rows at once. You may reason using matrix splittings and norm-based arguments.

4 Problem 4

Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix and $b \in \mathbb{C}^n$ be a nonzero starting vector. We consider the Krylov subspace of dimension k:

$$\mathcal{K}_k(A, b) = \text{span}\{b, Ab, A^2b, \dots, A^{k-1}b\}$$

1. **Arnoldi vs. Lanczos Structure**
   The Arnoldi iteration generates an orthonormal basis $Q_k = [q_1, \dots, q_k]$ for $\mathcal{K}_k(A, b)$ satisfying the relation:

   $$AQ_k = Q_{k+1}\widetilde{H}_k$$

   where $\widetilde{H}_k$ is a $(k + 1) \times k$ upper Hessenberg matrix.

   (a) Prove that if A is Hermitian ($A = A^*$), the matrix $\widetilde{H}_k$ reduces to a real, symmetric tridiagonal matrix (ignoring the last row).

   (b) Based on this structural reduction, explain the key difference in **storage complexity** between the standard Arnoldi algorithm and the Lanczos algorithm.

2. **GMRES Convergence Bound**
   The Generalized Minimal Residual (GMRES) method finds an approximate solution $x_k \in \mathcal{K}_k(A, b)$ that minimizes the residual norm $\|r_k\|_2 = \|b - Ax_k\|_2$.

   Assume A is diagonalizable, $A = V\Lambda V^{-1}$. Derive the following upper bound for the relative residual norm:

   $$\frac{\|r_k\|_2}{\|r_0\|_2} \le \kappa_2(V) \inf_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{\lambda\in\sigma(A)} |p(\lambda)|$$

   where $\kappa_2(V)$ is the condition number of the eigenvector matrix, $\sigma(A)$ is the spectrum of A, and $\mathcal{P}_k$ is the set of polynomials of degree at most k.

3. **Finite Termination Property**
   Using the bound derived in Part 2, prove the following:

   If the matrix A has exactly m distinct eigenvalues (where $m \le n$), then the GMRES method will converge to the exact solution (in exact arithmetic) in at most m iterations.