AMATH 584

HW 5

Nate Whybra

12.1) We have

$$101 = \|A\|_F = \sqrt{\sigma_1^2 + \sum_{i=2}^{202} \sigma_i^2}$$

$$= \sqrt{\|A\|_2^2 + \sum_{i=2}^{202} \sigma_i^2}$$

$$= \sqrt{100^2 + \sum_{i=2}^{202} \sigma_i^2} \quad (1)$$

$$\leftrightarrow \sum_{i=2}^{202} \sigma_i^2 = 101^2 - 100^2 = 201$$

suppose $\sigma_i = \sigma_j \ \forall \ i,j \geq 2$, then

$$201\, \sigma_i^2 = 201$$

$$\leftrightarrow \sigma_i = 1 \quad \text{(since singular values are positive)}$$

If all σ_i with $i \geq 2$ were > 1 the formula (1) would not hold. So it must be that $\exists i$ with $\sigma_i \leq 1$, so that the RHS of (1) ≤ 101.

However, equality happens by setting $\sigma_i = 1$ as demonstrated above, for all $i \geq 2$. So

$$K(A) = \frac{\sigma_1}{\sigma_{202}} \leq \frac{\sigma_1}{1} = \|A\|_2 = 100$$

and this is the tightest possible bound, because we have shown the smallest singular value of A is at least equal to 1.

21.1)

a) We have

$$\underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix}}_{U}$$

So $\det(A) = \det(L) \cdot \det(U)$, but since both matrices are triangular, their determinants are the products of the diagonals, hence,

$$\det(A) = (1)^4 \cdot (2)^3 \cdot 1 = 8$$

b) we have,

$$\underbrace{\begin{bmatrix} & & 1 & \\ & & & 1 \\ & 1 & & \\ 1 & & & \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & & & \\ 3/4 & 1 & & \\ 1/2 & -\frac{2}{7} & 1 & \\ 1/4 & -\frac{3}{7} & 1/3 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & 1/4 \\ & & -6/7 & -2/7 \\ & & & 2/3 \end{bmatrix}}_{U}$$

we have $\det(P) \cdot \det(A) = \det(L) \cdot \det(U)$. $\det(P) = -1$ since it is an odd number of row swaps from I. As L and U are triangular, the determinants are the products of the diagonals. So

$$\det(A) = -(1)^4 \cdot 8 \cdot \frac{\cancel{7}}{2\cancel{4}} \cdot -\frac{\cancel{6}^{2}}{\cancel{7}} \cdot \frac{\cancel{2}}{\cancel{3}} = -8 \cdot (-1) = +8$$

c) To compute $\det(A)$ from $PA = LU$ we can compute the product of the products of the diagonals of L and U, and multiply it by $(-1)^{\#\text{ of rowswaps for } P}$ however, L has 1's ~~on~~ the diagonals which simplifies things further, in general we can write

$$\det(A) = (-1)^{m - \sum_{i=1}^{m} P_{ii}} \cdot \prod_{i=1}^{m} \operatorname{diag}(U)_i$$

where $m - \sum_{i=1}^{m} P_{ii}$ counts how many columns of P are "not in the right place" considering if there were no permutations, P would be the identity I.

A1) If we make $\beta = 7$ with finite precision t, $\frac{8}{7}$ can be represented as 1.1, which can be verified by noticing that $1 \cdot 7^0 + 1 \cdot 7^{-1} = 1 + \frac{1}{7} = 8/7$.
There is no base β in which we can represent π with finite precision t exactly since it can be shown that irrational numbers have infinitely long non-repeating decimal representations in any integer base β, I proved this in combinatorics class once.

A2)

a) $$K = \frac{\|f'(1.2)\|}{\|f(1.2)\| / \|1.2\|} = (1.2)\cdot\frac{|\cos(1.2)|}{|\sin(1.2)|}$$

$$= (1.2)\cdot|\cot(1.2)| \approx 0.467 < 10^2$$

Since the relative condition number of this problem is small, it is well conditioned.

b) The plot does not behave as I expect. The error is large when h is large (as expected), but also large when h is small (not expected), and there seems to be a minimal error on the order of 10^{-10} when $10^{-10} < h < 10^{-5}$.

c) (See next page)

c) Assuming these rounding errors we have that our derivative approximation is

$$\approx \underbrace{\frac{f(x_0+h)-f(x_0)}{h}}_{\substack{(1)\\ \text{derivative approximation error}}} + \underbrace{O\left(\frac{\varepsilon_{machine}}{h}\right)}_{\substack{(2)\\ \text{rounding error}}}$$

when h is large our error from term (1) is large but the error from term (2) is small, and when h is small the error in term (1) is small but the error from term (2) is large. This implies there should be some value of h where the error is minimized (a value of h where the contribution of errors from both terms is jointly minimized), which explains our observations from (b).

d) Using this alternate formula, the results are more like what I would expect. The error is small when h is small and large when h is large, it is a linear function on this loglog scale.

A3)

a) The condition number is $44.8023 < 10^2$, so yeah, the matrix is well-conditioned.

b) we have $\| x - x_{ge} \|_2 \approx 8.05$, this is saying the GE solution is noticeably different from our exact solution, so I don't trust this solution.

c) we have $\| x - x_{qr} \|_2 \approx 2.7e-14$ which is very very small, so I trust this solution.

d) we have $\| x - x_{geop} \|_2 \approx 3.4e-15$ which is very very small, so I trust this solution.

Editor - C:\Users\Nwhybra\Desktop\UW AMATH Masters\AMATH 584\HW\HW5\problem_A2.m

problem_A2.m | problem_A3.m | gecp.m | +

```
x0 = 1.2;
f_prime_exact = cos(x0);
h = 10.^(0:-1:-16);
f_prime_approx = (sin(x0 + h) - sin(x0)) ./ h;
better_approx = (2 * cos(x0 + h/2) .* sin(h/2)) ./ h;
error_1 = abs(f_prime_exact - f_prime_approx);
error_2 = abs(f_prime_exact - better_approx);

figure;
loglog(h, error_1);
xlabel("h")
ylabel("Error (b)")
title("Error (b) vs. h")

figure;
loglog(h, error_2);
xlabel("h")
ylabel('Error (d)')
title('Error (d) vs. h')
```

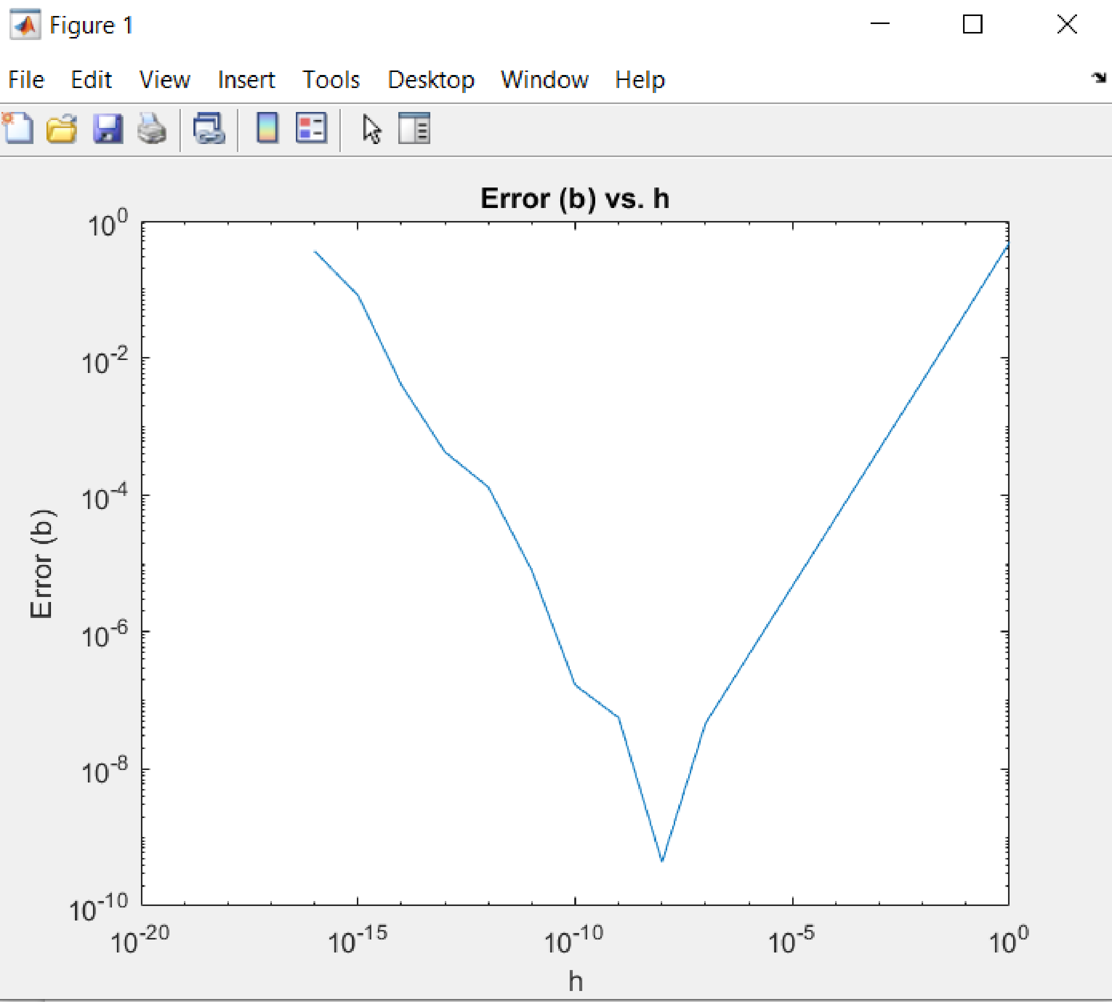
Figure 1
File Edit View Insert Tools Desktop Window Help
Error (b) vs. h
Error (b)
h
10^0
10^-2
10^-4
10^-6
10^-8
10^-10
10^-20
10^-15
10^-10
10^-5
10^0

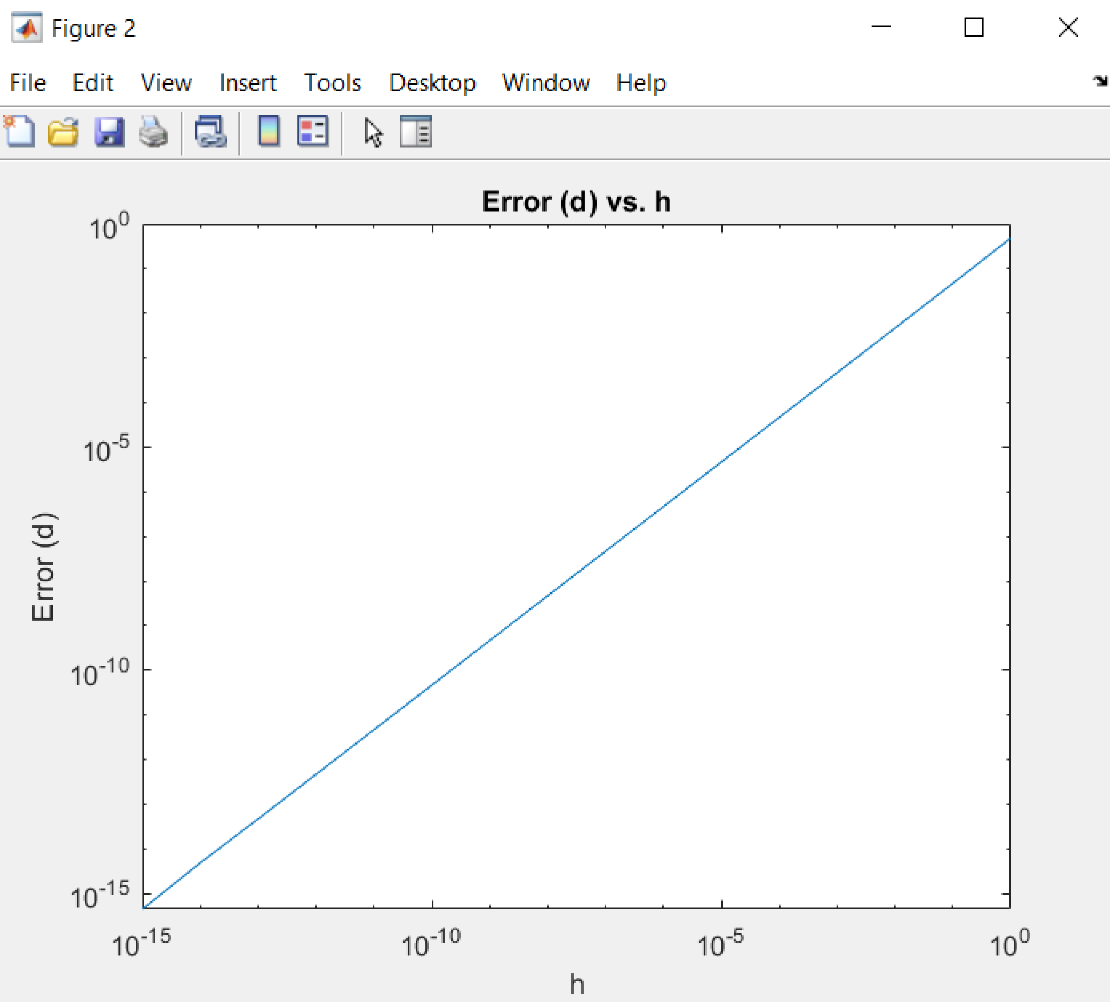
Figure 2
File Edit View Insert Tools Desktop Window Help
Error (d) vs. h
Error (d)
h
10^0
10^{-5}
10^{-10}
10^{-15}

Editor - C:\Users\Nwhybra\Desktop\UW AMATH Masters\AMATH 584\HW\HW5\problem_A3.m

problem_A2.m | problem_A3.m | gecp.m | +

```matlab
% Make the matrix A as defined in the problem.
n = 100;
A = -1 * tril(ones(n), -1) + eye(n);
A(:, end) = 1;

% Make the random vector.
x = randn(100, 1);

% Calculate b.
b = A * x;

% Part (a).
k = cond(A);

% Part (b).
x_ge = A \ b;
norm_error_1 = norm(x - x_ge, 2);

% Part (c).
[Q,R] = qr(A, 0);
x_qr = R \ (Q' * b);
norm_error_2 = norm(x - x_qr, 2);

% Part (d).
% P*A*Q = L*U...
[L, U, P, Q] = gecp(A);
x_gecp = Q * inv(U) * inv(L) * P * b;
norm_error_3 = norm(x - x_gecp, 2);
```