Justin Hexem
AMATH 584

Homework #5

1.) Proof: Let $A \in \mathbb{C}^{m \times n}$, $m \geq n$, be a full rank matrix. Let $\hat{x}$ be the minimizer of $\|Ax-b\|_2$ over all choices of $x \in \mathbb{C}^n$. Consider the $(m+n)\times(m+n)$ system

$$\begin{bmatrix} I_{mm} & A \\ A^* & O_{nn} \end{bmatrix} y = \begin{bmatrix} b \\ O_n \end{bmatrix} \iff \begin{bmatrix} I_{mm} & A \\ A^* & O_{nn} \end{bmatrix} \begin{bmatrix} y_m \\ y_n \end{bmatrix} = \begin{bmatrix} b \\ O_n \end{bmatrix}$$

where O_n is a vector of n zeros, O_{nn} is an $n \times n$ matrix of all zeros, and I_{mm} is the $m \times m$ identity matrix. Show $\hat{x}$ will be a submatrix of y. We see that the above system is equivalent to

$$\begin{bmatrix} I_{mm} & A \\ A^* & O_{nn} \end{bmatrix} \begin{bmatrix} y_m \\ y_n \end{bmatrix} = \begin{bmatrix} b \\ O_n \end{bmatrix} \iff \begin{bmatrix} y_m + Ay_n \\ A^* y_m \end{bmatrix} = \begin{bmatrix} b \\ O_n \end{bmatrix} \iff \begin{aligned} y_m + Ay_n &= b \quad (1) \\ A^* y_m &= 0. \quad (2) \end{aligned}$$

Let us multiply both sides of $y_m + Ay_n = b$ by A^* on the left.

$$y_m + Ay_n = b \iff A^*(y_m + Ay_n) = A^* b \iff A^* y_m + A^* A y_n = A^* b$$

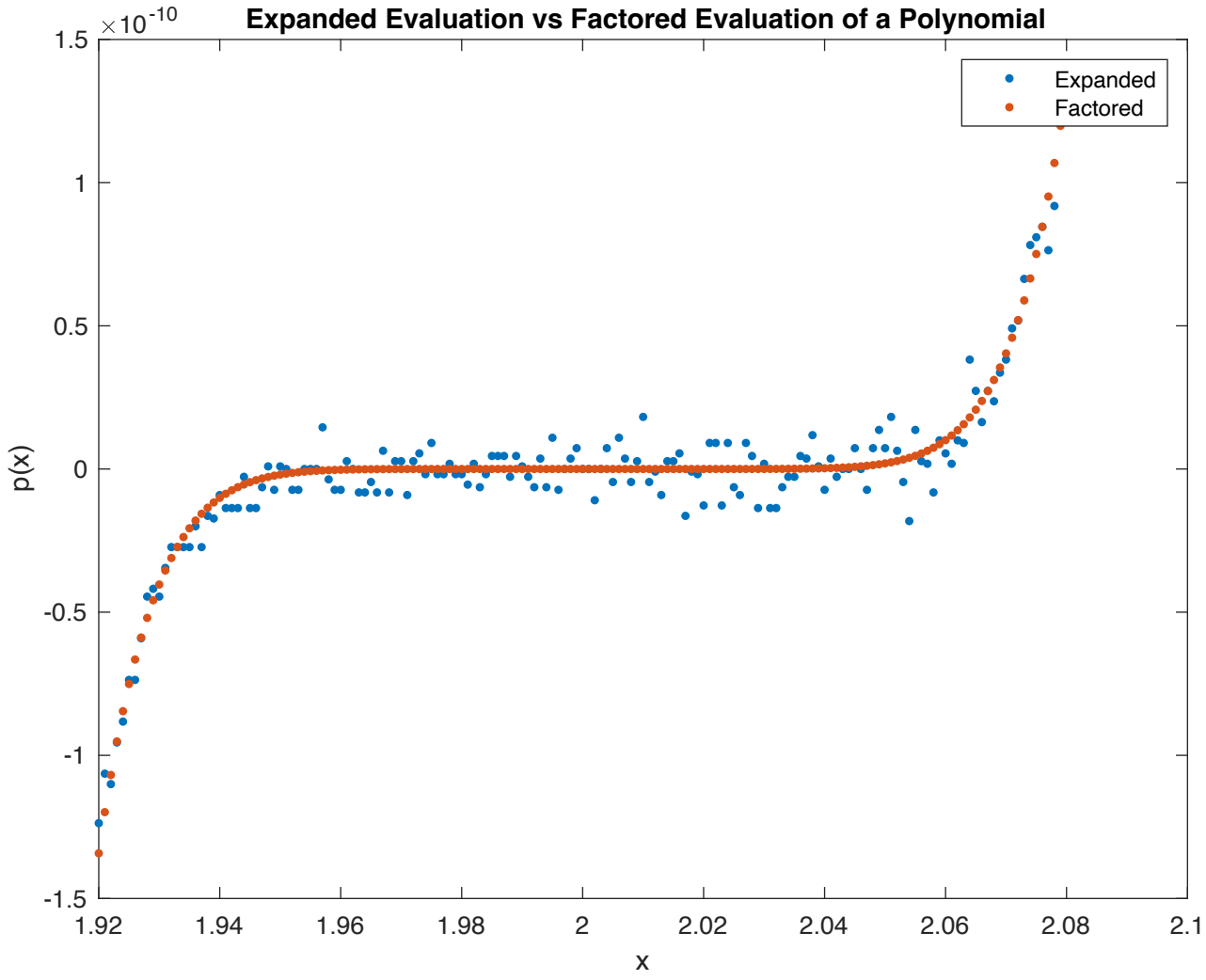
Since $A^* y = 0$ (equation (2)) we see (1) becomes $A^* A y_n = A^* b$. Since $\hat{x}$ is the minimizer of $\|Ax-b\|_2$ for all $x \in \mathbb{C}^n$, theorem 11.1 says $\hat{x}$ must satisfy $A^* A x = A^* b$. If $y_n = \hat{x}$, then (1) is clearly satisfied, so $y = [y_m \ \hat{x}]^T$ satisfies the system above. Thus $\hat{x}$ is a submatrix of y in the system above. ☒

2.) Let $A \in \mathbb{C}^{2024 \times 202}$ with $\|A\|_2 = 100$ and $\|A\|_F = 101$. Then we know from theorem 5.3 that $\|A\|_2 = \sigma_1 = 100$ and $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_{202}^2} = 101$. We also know that $\sigma_2 \geq \sigma_{202}$, and if $\sigma_2 = \sigma_{202}$ then $\sigma_2 = \sigma_3 = \dots = \sigma_{201} = \sigma_{202}$. Let us assume this is the case this is the largest value possible for σ_{202}. Then we see that

$$\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_{202}^2} = 101 \implies \sqrt{100^2 + \sigma_{202}^2 + \sigma_{202}^2 + \dots + \sigma_{202}^2} = 101 \implies 100^2 + 201\sigma_{202}^2 = 101^2$$

$$\implies 201\sigma_{202}^2 = 101^2 - 100^2 \implies \sigma_{202} = \sqrt{\frac{101^2 - 100^2}{201}} = \sqrt{\frac{201}{201}} = 1.$$

By (12.16) in the textbook $\kappa(A)$ in this case is $\kappa(A) = \frac{\sigma_1}{\sigma_{202}} = \frac{100}{1} = 100$. Since this is the largest σ_{202} can be, this means 100 must be the smallest $\kappa(A)$ can be. Thus $\kappa(A) \geq 100$. ☒


Expanded Evaluation vs Factored Evaluation of a Polynomial
×10^-10
1.5
1
0.5
0
-0.5
-1
-1.5
p(x)
1.92
1.94
1.96
1.98
2
2.02
2.04
2.06
2.08
2.1
x
Expanded
Factored

4.) The digits of x from parts d, e, and f seem to be similar but all digits of x from part a seem to be incorrect except for the first two entries. This indicates that solving the least squares problem via the normal equations might exhibit some instability. From my code $k_2(R) = 1.1718 \cdot 10^8$ and $k_2(A^*A) = 1.3535 \cdot 10^{16}$, which indicates that the normal equations are much more unstable when compared to the condition number for R. Both condition numbers are greater than 10^6, so both of these seem to be ill-conditioned if we define large to mean greater than 10^6. Theorem 12.2 states that computing $x = A^{-1}b$ has condition number $k(A) = \|A\| \|A^{-1}\|$, so since our computed $k(R)$ is on the order of 10^8, we will lose $\log_{10} k(R) = 8$ digits of accuracy for R and $\log_{10} k(A^*A) = 16$ digits of accuracy for the normal equations. From this analysis, we can conclude that the QR method gives 8 decimal points of accuracy and the normal equations will give 0 decimal points of accuracy if double precision numbers have 16 decimal points of accuracy.

```
    "x from part a is: "

1.0000000307358379
-0.0000090460514431
-7.9996596611461346
-0.0050308928693223
10.7053389299906918
-0.1752295133667835
-5.1909997733973592
-0.9080564462088736
2.6749369516676724
-0.7146396516660373
-0.0743126430932392
0.0340180777420736
```

```
ans =

    "x from part d is: "

1.0000000009966072
-0.0000004227430514
-7.9999812356841176
-0.0003187632591137
10.6694307961079513
-0.0138202889027496
-5.6470756248757352
-0.0753160286666863
1.6936069684431345
0.0060321052024426
-0.3742417019919809
0.0880405758115372
```

```
    "x from part e is: "

1.0000000009966059
-0.0000004227434936
-7.9999812356706981
-0.0003187634226586
10.6694307971883813
-0.0138202932422330
-5.6470756136825262
-0.0753160476154302
1.6936069893872410
0.0060320906481438
-0.3742416962208316
0.0880405748157388
```

```
    "x from part f is: "

1.0000000009966050
-0.0000004227430403
-7.9999812356842277
-0.0003187632578376
10.6694307960983181
-0.0138202888600174
-5.6470756249921275
-0.0753160284652890
1.6936069682204147
0.0060321053554234
-0.3742417020516586
0.0880405758216802
```

```
k2R =

   1.1718e+08


k2AA =

   1.3535e+16
```

```
m = 50;
n = 12;
t = linspace(0, 1, m);

A = fliplr(vander(t));
A = A(:, 1:n);
b = cos(4 * t');

xa = (A' * A) \ (A' * b);
"x from part a is: "
fprintf('%.16f\n', xa)

[Q, R] = qr(A, "econ");
xd = R \ (Q' * b);
"x from part d is: "
fprintf('%.16f\n', xd)

xe = A \ b;
"x from part e is: "
fprintf('%.16f\n', xe)

[U, S, V] = svd(A, "econ");
w = S \ (U' * b);
xf = V * w;
"x from part f is: "
fprintf('%.16f\n', xf)

xall = [xa, xd, xe, xf];

SR = svd(R);
k2R = SR(1, 1) / SR(end, end)

SAA = svd(A' * A);
k2AA = SAA(1, 1) / SAA(end, end)
```