Homework #5
CS 575: Numerical Linear Algebra
Spring 2023

John Tran

March 20, 2023

Important Notes

- `Python` 3.11 was used to run the notebook (i.e., to use the `match` statement)
- the partially completed notebook was used to complete the coding assignment, so many of the things asked (e.g., verification and testing) was done for us and they were modified for the mat-mat portion
- the `README` was done in `Markdown` but the raw text can still be viewed

Problem 1

We are given A as a 2×2 lower triangular matrix in the system $Ax = b$. We are asked to show that the computed solution $\hat{x}$ by forward substitution satisfies $(A + F)\hat{x} = b$, where $f_{ij} \leq 2\epsilon|a_{ij}|$ (thus, forward substitution is stable).

First, we can write A, x, b as

$$A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

We covered the 1×1 case in class and the general approach will be outlined below, adapted for the first step of the forward substitution (i.e., with a_{11}:

$$\begin{aligned} a_{11}x_1 &= b_1 \\ \Rightarrow x_1 &= \frac{b_1}{a_{11}} \\ \mathrm{fl}(x_1) &= \mathrm{fl}\left(\frac{b_1}{a_{11}}\right) \\ \hat{x} &= \frac{b_1}{a_{11}}(1 + \delta) \\ \Rightarrow b_1 &= a_{11}\hat{x}\left(\frac{1}{1 + \delta}\right) \end{aligned}$$

where any floating point operation $x \odot y$ results in

$$x \odot y = (x \cdot y)(1+\delta)$$

where $(\odot)$ was taken to be multiplication $(\cdot)$ above, as an example

In the above case, $(\odot)$ was used in place for division, and as a recap,

$$\begin{aligned} |\delta| &\leq \epsilon \\ \frac{1}{1+\delta} &\approx 1+\delta' \quad \text{where } |\delta'| \leq \epsilon + O(\epsilon^2) \end{aligned}$$

where we can use a first order Taylor approximation for $(1+\delta)^{-1}$ and ignore the $O(\epsilon^2)$ term

Continuing where we left off,

$$\begin{aligned} b_1 &\approx a_{11}\hat{x}(1+\delta') \\ &= (a_{11} + a_{11}\delta')\hat{x} \\ &= (a_{11} + f_{11})\hat{x} \end{aligned}$$

So for the first step of the forward substitution, we have $f_{11} = \delta'|a_{11}| \leq \epsilon|a_{11}|$

Then for the second step,

$$\begin{aligned} a_{21}x_1 + a_{22}x_2 &= b_2 \\ \Rightarrow a_{22}x_2 &= b_2 - a_{21}x_1 \\ x_2 &= \frac{b_2 - a_{21}x_1}{a_{22}} \\ \text{fl}(x_2) &= \text{fl}\left(\frac{b_2 - a_{21}x_1}{a_{22}}\right) \\ \hat{x_2} &= \frac{(b_2 - (a_{21}\hat{x_1})(1+\delta))(1+\delta)}{a_{22}}(1+\delta) \\ &= \frac{(b_2 - (a_{21}\hat{x_1})(1+\delta))}{a_{22}}(1+\delta)^2 \\ \Rightarrow (a_{22}\hat{x_2})\frac{1}{(1+\delta)^2} &= b_2 - (a_{21}\hat{x_1})(1+\delta) \end{aligned}$$

where we apply a factor of $(1+\delta)$ for each floating point operation $(\odot)$ – one for the multiplication $a_{21}x_1$, one for the subtraction $b_2 - a_{21}x_1$, and one for the division with a_{22}

Again, we can use a first order Taylor approximation for $(1+\delta)^{-2}$ and ignore the $O(\epsilon^2)$ term

$$\begin{aligned} |\delta| &\leq \epsilon \\ \frac{1}{(1+\delta)^2} &\approx 1 + 2\delta' \quad \text{where } |\delta'| \leq \epsilon + O(\epsilon^2) \end{aligned}$$

Continuing where we left off,

$$\begin{aligned} (a_{22}\hat{x_2})(1+2\delta') &\approx b_2 - (a_{21}\hat{x_1})(1+\delta) \\ \Rightarrow b_2 &= (a_{21}\hat{x_1})(1+\delta) + (a_{22}\hat{x_2})(1+2\delta') \\ &= (a_{21} + a_{21}\delta)\hat{x_1} + (a_{22} + 2a_{22}\delta')\hat{x_2} \\ &= (a_{21} + f_{21})\hat{x_1} + (a_{22} + f_{22})\hat{x_2} \end{aligned}$$

From here, we can see that $f_{21} = \delta|a_{21}| \leq \epsilon|a_{21}|$ and $f_{22} = 2\delta|a_{22}| \leq 2\epsilon|a_{22}|$. If we look at the largest magnitude of f_{ij}, we see that f_{22} has the highest bound of $2\epsilon|a_{22}|$.

So, we can generalize our results based on the factor of 2 of the corresponding $a_{ij}\epsilon$ bound: $f_{ij} \leq 2\epsilon|a_{ij}|$.

Note: for larger sizes of A, we have to be careful about the order in which we do the subtraction and the stacking of floating point operations – better to rely on a general approach than looking at specific components

Problem 2

(a) If we write down the relationship in terms of $W_N, W_O, andW_{a,b}$ (corresponding to N, O, N_aO_b, respectively), we have the simple expression

$$aW_N + bW_O = W_{a,b}$$

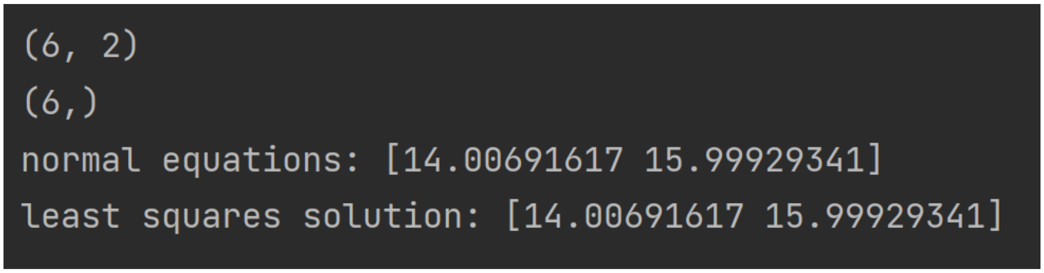
(b) Using the expression we have above, we can simply plug in the values for $a, b, W_{a,b}$ provided to us where a, b are derived from the chemical formulas N_aO_b and $W_{a,b}$ from the molecular weight for each data row

$$\begin{aligned} W_N + W_O &= 30.006 \\ 2W_N + W_O &= 44.013 \\ W_N + 2W_O &= 46.006 \\ 2W_N + 3W_O &= 76.012 \\ 2W_N + 4W_O &= 92.011 \\ 2W_N + 5W_O &= 108.010 \end{aligned}$$

where we have the normal equations $A^T Ax = A^T b$ for

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \\ 2 & 3 \\ 2 & 4 \\ 2 & 5 \end{bmatrix}, \qquad x = \begin{bmatrix} W_N \\ W_O \end{bmatrix}, \qquad b = \begin{bmatrix} 30.006 \\ 44.013 \\ 46.006 \\ 76.012 \\ 92.011 \\ 108.010 \end{bmatrix}$$

(c) Below are the molecular weights for x defined above, trying both the Normal Equations using the built-in solver and built-in least squares fit in `Python`



```
(6, 2)
(6,)
normal equations: [14.00691617 15.99929341]
least squares solution: [14.00691617 15.99929341]
```

Figure 1: The least squares fit comparing the built-in `numpy.linalg.lstsq` (direct least squares fit) and `numpy.linalg.solve` functions (using the Normal Equations) – see code for more details

> For this simple system, we found that both methods seemed to produce the same results.

Problem 3

We are asked to show that if $Ax = b$ has a solution, then this set of solutions must be equal to the set of solutions for $A^T Ax = A^T b$.

If we use the definition for the Normal Equations, then there exists some vector $y \in \mathbb{R}^n$ (not to be confused with the vector x in the solution to $Ax = b$) that minimizes $\|b - Ay\|_2$ where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

The definition states that $\|b - Ay\|_2$ is minimized ***if and only if*** $A^T Ax = A^T b$. Since we are given x as a solution to $Ax = b$, x produces the norm $\|b - Ay\|_2$ to be 0. This minimizes the norm since norms can only be non-negative values, so x must also be a solution to $A^T Ax = A^T b$.

Problem 4

We are given $A = QR$ where Q is an $m \times m$ orthogonal matrix and R is an $m \times m$ upper triangular matrix.

(a) We are asked to show that $|\det Q| = 1$ (adapted from textbook)

$$\begin{aligned}
1 &= \det I && \text{property of I} \\
&= \det Q^{-1}Q \\
&= \det Q^{T}Q && \text{property of } Q \\
&= \det Q^{T} \det Q && \text{separability of determinant} \\
&= (\det Q)^{2} && \text{det same for transpose} \\
\Rightarrow \det Q &= \pm 1 \\
\Leftrightarrow |\det Q| &= 1
\end{aligned}$$

(b) We are asked to show that $\|a_j\|_2 = \|r_j\|_2$ for any jth columns of A and R, respectively.

$$\begin{aligned}
A &= QR \\
&= [Qr_1 Qr_2 \dots Qr_m] && \text{composition of } Q \text{ times the columns of } R \\
&\Leftrightarrow \\
a_j &= Qr_j \quad \forall j \in \{1, 2, \dots, m\} \\
&\Leftrightarrow \\
\|a_j\|_2 &= \|Qr_j\|_2 \\
\|a_j\|_2 &= \|r_j\|_2 && \text{property of } Q\text{, preserve lengths (2-norm)}
\end{aligned}$$

(c) We are asked to give an algebraic proof of Hadamard's inequality:

$$|\det A| \leq \prod_{j=1}^{m} \|a_j\|_2$$

If we look at the hints provided, the last two hints follow naturally if we just use $A = QR$. However, if we want to use the transpose of a given matrix and interweave the determinant of the identity, then a likely approach would be to look at $A^T A$:

$$\begin{aligned}
\det A^T A &= \det (QR)^T (QR) & \\
\det A^T \det A &= \det R^T Q^T Q R & \text{hint 3} \\
(\det A)^2 &= \det R^T Q^T Q R & \text{hint 2} \\
&= \det R^T Q^{-1} Q R & \text{property of } Q^T \\
&= \det R^T I R & \\
&= \det R^T \det I \det R & \text{hint 3} \\
&= \det R^T \det R & \text{hint 1} \\
&= (\det R)^2 & \text{hint 2} \\
\Rightarrow |\det A| &= |\det R| &
\end{aligned}$$

From the last step, we can now use the final hint given to use since R is an upper ***triangular matrix*** and expand the definition of the 2-norm to produce the inequality we need:

$$\begin{aligned}
|\det R| &= \left| \prod_{j=1}^{m} r_{jj} \right| & \text{hint 4} \\
&= \prod_{j=1}^{m} |r_{jj}| & \\
&\leq \prod_{j=1}^{m} \left(\sqrt{\sum_{i=1}^{m} |r_{ij}|^2} \right) & * \\
&= \prod_{j=1}^{m} \|r_j\|_2 & \text{def. of 2-norm} \\
&= \prod_{j=1}^{m} \|a_j\|_2 & \text{result from part (b)}
\end{aligned}$$

*for a given column r_j, the corresponding diagonal element is bounded above by the length of the column vector (i.e., equality when the only nonzero element is along the diagonal, else contributions from other elements would increase the length)

So, we have our final proof by combining the two parts above $|\det A| \leq \prod_{j=1}^{m} \|a_j\|_2$

Problem 5

We are asked to show that if $Q \in \mathbb{R}^{m \times m}$ is orthogonal, then for any $A \in \mathbb{R}^{m \times m}$

$$\|AQ\|_2 = \|QA\|_2 = \|A\|_2$$

We can use the property of orthogonal matrices Q that preserve lengths

$$\|Qx\|_2 = \|x\|_2 \quad \forall x \in \mathbb{R}^m$$

to prove the statement above along with the definition of an induced matrix norm $\|A\| = \max_{\|x\|=1} \|Ax\|$ (substitute in the 2-norm for our case)

For $\|AQ\|_2$,

$$\begin{aligned} \|AQ\|_2 &= \max_{\|x\|_2=1} \|AQx\|_2 \\ &= \max_{\|Qx\|_2=1} \|AQx\|_2 && Q \text{ preserve 2-norm in max condition} \\ &= \max_{\|y\|_2=1} \|Ay\|_2 && \text{substitute } y = Qx \\ &= \|A\|_2 && \text{apply def. of induced matrix norm} \end{aligned}$$

For $\|QA\|_2$, we cannot use the same trick we had above since Q does not directly interact with x anymore, being on the other side of A. However, if we look at the definition of the norm induced by $\|\cdot\|_2$,

$$\|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$$

Even though the eigenvalue might pose a problem, we will see that it does not matter when we expand $(QA)^T(QA)$

$$\begin{aligned}
\|QA\|_2 &= \sqrt{\lambda_{max}((QA)^T(QA))} \\
&= \sqrt{\lambda_{max}(A^TQ^TQA)} \\
&= \sqrt{\lambda_{max}(A^TQ^{-1}QA)} && \text{property of } Q, Q^T = Q^{-1} \\
&= \sqrt{\lambda_{max}(A^TA)} \\
&= \|A\|_2 && \text{apply def. of induced matrix } \|\cdot\|_2
\end{aligned}$$

Problem 6

We are asked to find a $u \in \mathbb{R}^2$ with $\|u\|_2 = 1$ such that,

$$Q_u \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

We can use a proposition discussed in class where we can let $x, y \in \mathbb{R}^m$, with $\|x\|_2 = \|y\|_2$ and $x \neq y$. Then there exists a vector $u \in \mathbb{R}^m$ with $\|u\|_2 = 1$ such that

$$Q_u x = y$$

and u is uniquely determined as

$$u = \pm \frac{x - y}{\|x - y\|_2}$$

(adapted from the textbook)

Before we can use this, we need to verify the two conditions for x, y are met. For our case, $m = 2$ and we can let

$$x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

First, $x \neq y$ is trivial by inspection. Then $\|x\|_2 = \|y\|_2$, we can see that the same elements are contained in x, y, but in different orders. Since the

2-norm relies only on the sum of the square of the elements, the norms for x, y are the same.

With both conditions satisfied, we can plug in x, y above to find u

$$\begin{aligned}
x - y &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
\|x - y\|_2 &= \sqrt{|1|^2 + |-1|^2} \\
&= \sqrt{1 + 1} \\
&= \sqrt{2} \\
u &= \pm \frac{x - y}{\|x - y\|_2} \\
&= \pm \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}
\end{aligned}$$