Numerical Analysis Assignment 2

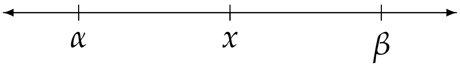
Name: Nate Stemen (20906566)
Email: nate.stemen@uwaterloo.ca

Due: Fri, Sep 25, 2020 5:00 PM
Course: AMATH 740

Problem 1

Floating point numbers.

Solution. It's helpful if we start with a picture. Let α, β be two adjacent numbers in our number system and x to be equidistant between them.



Note if we round x as $\text{fl}(x)$, then x will be rounded to either α or β depending on even-ness. By the definition of floating point number systems, the distance between α and β is β^{-t+1} ($\alpha - \beta = \beta^{-t+1}$), and because the way we positioned x we have $x = \alpha + \frac{1}{2}\beta^{-t+1} = \alpha + \mu$. If we take $\alpha = 1$ and $\beta = 1 + \texttt{eps}$, then we can see the distance between them is 2μ and hence $\texttt{eps} = 2\mu$.

Problem 2

Consider the expression

$$z_A = \frac{1}{\sqrt{1+x^2} - \sqrt{1-x^2}}.$$

(a) Explain why the formula for z_A is susceptible to catastrophic cancellation errors for x close to 0.

(b) Use reformulation to find an alternative expression z_B for expression z_A in a way that avoids catastrophic cancellation for x close to 0. (Hint: $p^2 - q^2 = (p-q)(p+q)$.)

Solution. **??** To see why z_A is susceptible to catastrophic cancellation we can use the Taylor expansion around $x = 0$ of $\sqrt{1+x} \approx 1 + \frac{x}{2}$.

$$\begin{aligned} z_A &= \frac{1}{\sqrt{1+x^2} - \sqrt{1-x^2}} \\ &\approx \frac{1}{1 + \frac{1}{2}x^2 - (1 - \frac{1}{2}x^2)} \end{aligned}$$

Just as we saw in lecture, when subtracting two numbers that are very close to each other (in this case $1+\varepsilon$ and $1-\varepsilon$), then we are susceptible to catastrophic cancellation.

??

In order to avoid catastrophic cancellation we can further the last line of z_A to express it as $\frac{1}{x^2}$. Written this was there is no catastrophic cancellation, just a blowup at $x = 0$.

Problem 3

Consider the matrix

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 3}$$

(a) Determine (by hand) the reduced QR factorization of A using the Gram-Schmidt algorithm. I.e., orthonormalise the column vectors of A, and write the result as

$$A = \widehat{Q}\widehat{R},$$

with $\widehat{Q} \in \mathbb{R}^{4\times 3}$ and $\widehat{R} \in \mathbb{R}^{3\times 3}$.

(b) Extend this to the full QR factorization of A:

$$A = QR$$

with $Q \in \mathbb{R}^{4\times 4}$ and $R \in \mathbb{R}^{4\times 3}$.

Solution. **??** This is mostly a computational question, so I'll just show that.

$$\mathbf{u}_1 = \mathbf{a}_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{u}_2 = \mathbf{a}_2 - \frac{\langle \mathbf{u}_1, \mathbf{a}_2 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{e}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{u}_3 = \mathbf{a}_3 - \frac{\langle \mathbf{u}_1, \mathbf{a}_3 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 - \frac{\langle \mathbf{u}_2, \mathbf{a}_3 \rangle}{\langle \mathbf{u}_2, \mathbf{u}_2 \rangle}\mathbf{u}_2 = \frac{1}{2}\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{e}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

We can then construct the reduced factorization as follows.

$$\widehat{Q} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 \end{bmatrix}$$

$$\widehat{R} = \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{a}_1 \rangle & \langle \mathbf{e}_1, \mathbf{a}_2 \rangle & \langle \mathbf{e}_1, \mathbf{a}_3 \rangle \\ 0 & \langle \mathbf{e}_2, \mathbf{a}_2 \rangle & \langle \mathbf{e}_2, \mathbf{a}_3 \rangle \\ 0 & 0 & \langle \mathbf{e}_3, \mathbf{a}_3 \rangle \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{2}{\sqrt{2}} & \frac{3}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

?? To extend this to the full QR factorization we can add another orthonormal vector onto the end of $\widehat{Q}$, and add a row of zeros to the bottom of $\widehat{R}$.

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad R = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{2}{\sqrt{2}} & \frac{3}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 \end{bmatrix}$$

Problem 4

Computational cost of Gram-Schmidt algorithm.

Solution. In order to get the dominant term here we only need to look at the inner most loop of the algorithm. So lets first just compute the computational work required for the `i=1:j-1` loop which we will denote by W_{inner}.

$$\begin{aligned} W_{\text{inner}} &= \sum_{i=1}^{j-1} m\mathrm{M} + (m-1)\mathrm{A} + m\mathrm{M} + m\mathrm{A} \\ &= \sum_{i=1}^{j-1} 4m - 1 \\ &= (j-1)(4m-1) \approx 4m(j-1) \end{aligned}$$

Now because this loop lives inside the outer loop we have to repeat this a few times.

$$\begin{aligned} W_{\text{dominant}} = \sum_{j=1}^{n} W_{\text{inner}} &= \sum_{j=1}^{n} 4m(j-1) \\ &= 4m \sum_{j=2}^{n} (j-1) \\ &= 4m \sum_{j=1}^{n-1} j \\ &= 4m\frac{1}{2}n(n-1) \\ &= 2mn^2 - 2mn \approx 2mn^2 \end{aligned}$$

Just want to say thanks for asking for the dominant term, it's so much easier, and seems like that's mostly what matters in practice.

Problem 5

Let

$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 3 & 10 \end{bmatrix}.$$

(a) Construct the first Householder reflection matrix, Q_1, which reflects the first column of A, $\mathbf{a}_1$ to a vector

$$Q_1\mathbf{a}_1 = \begin{bmatrix} \pm\|\mathbf{a}_1\| \\ 0 \end{bmatrix},$$

i.e., choose the sign according to the rule used to ensure numerical stability, determine vector $\mathbf{v}_1$ and its normalised version $\mathbf{u}_1$, then the matrix Q_1.

(b) Verify that Q_1 is an orthogonal matrix.

(c) Make a sketch in the $\mathbb{R}^2$ plane indicating the vectors $\mathbf{x} = \begin{pmatrix} x & y \end{pmatrix}^\intercal \in \mathbb{R}^2$ that arise in **????**: the original vector $\mathbf{a}_1$, and it's image $Q_1\mathbf{a}_1$ under the reflection. Also indicate the line about which the vectors are reflected.

(d) Compute Q_1A and write down the QR decomposition of A.

Solution. **??** First we have to compute the vector about which we will rotate $\mathbf{a}_1$.

$$\mathbf{v}_1 = \mathbf{a}_1 - \|\mathbf{a}_1\|\mathbf{e}_1 = \begin{bmatrix} 4 \\ 3 \end{bmatrix} - \begin{bmatrix} 5 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$$

And now to normalize it.

$$\mathbf{u}_1 = \frac{1}{\sqrt{10}} \begin{bmatrix} -1 \\ 3 \end{bmatrix}$$

With this we can now construct our Householder transformation as follows.

$$Q_1 = \mathbb{1} - 2\mathbf{u}_1\mathbf{u}_1^\intercal = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{5}\begin{bmatrix} 1 & -3 \\ -3 & 9 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix}$$

We can now apply this to $\mathbf{a}_1$ to verify we have the desired result.

$$Q_1\mathbf{a}_1 = \frac{1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix}\begin{bmatrix} 4 \\ 4 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 25 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

Now we see it's worked out, but we should change the sign of Q_1 so that the sign of $Q_1\mathbf{a}_1$ is opposite that of the first non-zero component of $\mathbf{a}_1$. So we take

$$Q_1 = \frac{-1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix}.$$

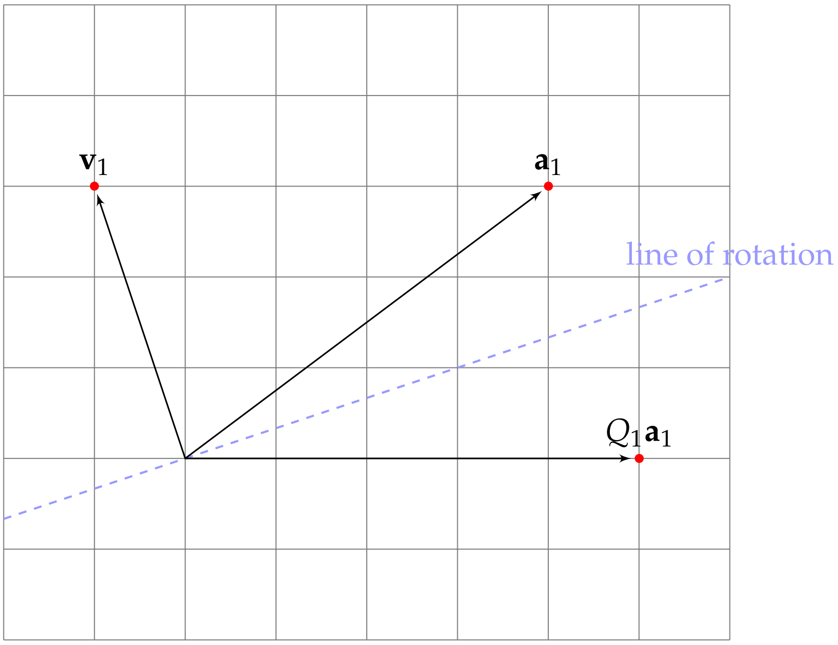
?? To check orthogonality all we need to do is verify $Q_1Q_1^\intercal = \mathbb{1}$[1].

$$\frac{-1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix} \cdot \frac{-1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix} = \frac{1}{25}\begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix} = \mathbb{1}$$

[1] As opposed to $QQ^\intercal = \mathbb{1} = Q^\intercal Q$ because Q_1 is symmetric so $Q_1Q_1^\intercal = Q_1^2 = Q_1^\intercal Q_1$.

??

Vectors of Interest



I couldn't draw it on the picture, but it's worth noting that $\mathbf{v}_1$ is perpendicular to the line of rotation as expected.

?? We can now compute Q_1A.

$$Q_1A = \frac{-1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix}\begin{bmatrix} 4 & 5 \\ 3 & 10 \end{bmatrix} = \frac{-1}{5}\begin{bmatrix} 25 & 50 \\ 0 & -25 \end{bmatrix} = \begin{bmatrix} -5 & -10 \\ 0 & 5 \end{bmatrix}$$

Thus we can write the complete QR factorization as

$$A = QR = \frac{-1}{5}\begin{bmatrix} 4 & 3 \\ 3 & -4 \end{bmatrix}\begin{bmatrix} -5 & -10 \\ 0 & 5 \end{bmatrix}$$

Problem 6

Determinant inequality.

Solution. Let's start with a helpful corollary.

Corollary. If $x \in \mathbb{R}^n$ then $\|x\|_\infty \leq \|x\|_2$.

Proof.

$$\|x\|_\infty = \max_i |x_i| = \left(\max_i x_i^2\right)^{1/2} \leq \left(\sum_i x_i^2\right)^{1/2} = \|x\|_2$$

We'll use that in a bit, in the meantime let's return to the problem.

Starting with the fact that A is invertible, we know we can factor A as $A = QR$ where Q is an orthogonal matrix, and R is upper triangular. Using the the fact that the determinant of product of matrices is the product of the determinants we can write $\det A = \det Q \cdot \det R$. The orthogonality of Q means $\det Q = \pm 1$, and hence $|\det A| = |\det R|$. Since R is upper triangular we know the determinant will be the product of the diagonal elements: $\det R = \prod_i r_{ii}$. Using the Gram-Schmidt version of QR decomposition we know the diagonal terms are given by $\langle \mathbf{e}_i, \mathbf{a}_i \rangle$[2] and hence $\det R = \prod_i \langle \mathbf{e}_i, \mathbf{a}_i \rangle = \prod_i (\mathbf{a}_i)_i$ where $(\mathbf{a}_i)_i$ is the ith component of vector $\mathbf{a}_i$.

So far we have shown $|\det A| = \prod_i (\mathbf{a}_i)_i$ so now we need to show $\prod_i (\mathbf{a}_i)_i \leq \prod_i \|\mathbf{a}_i\|_2$. If we are able to show $(\mathbf{a}_i)_i \leq \|\mathbf{a}_i\|_2$ is true for each i, then surely the product statement holds as well so we'll do that.

$$(\mathbf{a}_i)_i \leq \max_j (\mathbf{a}_i)_j = \|\mathbf{a}_i\|_\infty \leq \|\mathbf{a}_i\|_2$$

With that we've shown that the ith component of a vector is smaller than the 2 norm of it (which sounds obvious put that way), and hence $\prod_i (\mathbf{a}_i)_i \leq \prod_i \|\mathbf{a}_i\|_2$.

As a recap we used the QR factorization of A to show $|\det A| = |\det R|$ and R's upper triangular form to show $|\det A| = |\prod_i (\mathbf{a}_i)_i|$. We then used the corollary from above to prove

$$|\det A| \leq \prod_i \|\mathbf{a}_i\|_2.$$

[2] Here we are using the convention that $\mathbf{e}_i$ is the vector with 1 in the ith spot and 0 elsewhere.

Problem 7

Norm inequality.

Solution. The first thing we'll do is rewrite the inequality.

$$\begin{aligned} \frac{1}{\kappa(A)} &\leq \frac{\|A-B\|}{\|A\|} \\ \frac{1}{\|A^{-1}\|} &\leq \|A-B\| \\ 1 &\leq \|A^{-1}\| \, \|A-B\| \end{aligned}$$

So our goal will be to show this expression on the right is greater than one. We can use the submultiplicative property of matrix norms[3] to write $\left\|A^{-1}(A-B)\right\| \leq \left\|A^{-1}\right\|\|A-B\|$. Now let's take an $x \in \ker(B)$ with $\|x\| = 1$[4] and send it through $A^{-1}(A-B)$.

$$A^{-1}(A-B)x = A^{-1}(Ax - Bx) = A^{-1}Ax = x$$

Taking the (vector) norm of both sides yields $\left\|A^{-1}(A-B)x\right\| = 1$. With the way the matrix norm is defined as the maximum over a set of vectors that x is in, we can conclude $\left\|A^{-1}(A-B)\right\| \geq 1$.

Because $\left\|A^{-1}(A-B)\right\| \leq \left\|A^{-1}\right\|\|A-B\|$ we can thus conclude $1 \leq \|A^{-1}\| \, \|A-B\|$ by combining these two inequalities.

[3] $\|AB\| \leq \|A\|\|B\|$.

[4] This is always possible because the kernel of an operator is a subspace and hence must be amenable to scaling.