Michael Goforth
CAAM 550
HW 4
9/22/2021

Problem 1
part a

$$\boldsymbol{G^T G} = \begin{bmatrix} g_1^T \\ g_2^T \\ \dots \\ g_m^T \end{bmatrix} \begin{bmatrix} g_1 & g_2 & \dots & g_m \end{bmatrix}$$

where $g_1, g_2, \dots, g_m$ are the columns of $\boldsymbol{G}$. Then

$$\boldsymbol{G^T G} = \begin{bmatrix} g_1^T g_1 & g_1^T g_2 & \dots & g_1^T g_m \\ g_2^T g_1 & g_2^T g_2 & \dots & g_2^T g_m \\ \dots & \dots & \dots & \dots \\ g_m^T g_1 & g_m^T g_2 & \dots & g_m^T g_m \end{bmatrix}$$

and for a given element $q_a, b \in \boldsymbol{G^T G}$, for any $a, b \neq j, k$,

$$q_{a,b} = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

Also for any element $a \neq j, k$, $q_{a,j} = 0$, $q_{a,k} = 0$ and any element $b \neq j, k$, $q_{j,b} = 0$, $q_{k,b} = 0$. Finally,

$$q_{j,j} = q_{k,k} = cos^2(\theta) + sin^2(\theta) = 1$$

$$q_{j,k} = q_{k,j} = cos(\theta)sin(\theta) - cos(\theta)sin(\theta) = 0$$

Combining all of this together gives

$$q_{a,b} = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} = \boldsymbol{I}$$

$\boldsymbol{G^T G} = \boldsymbol{I}$ so a Givens rotation is orthogonal.

part b
Consider a matrix $A \in \mathbb{R}^{2x2}$ where

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and Givens rotation matrix

$$\boldsymbol{G} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

If we want to use a Givens rotation to make A upper triangular, then we want

$$(GA)_{1,2} = -a\ \sin(\theta) + c\ \cos(\theta) = 0$$

$$a\ \sin(\theta) = c\ \cos(\theta)$$
$$\theta = \arctan(c/a)$$

In a similar way, we can use Givens rotations to zero out all the lower elements of a generic matrix $\boldsymbol{A} \in \mathbb{R}^{nxn}$. For every element in a_i below the diagonal of $\boldsymbol{A}$, a matrix $\boldsymbol{G_j}$ can be formed such that $\boldsymbol{A_j} = \boldsymbol{G_j A}$ and $(A_j)_i = 0$. Repeat this for every subdiagonal element and then combine the G matrices such that

$$G = \prod_{i=1}^{n} \boldsymbol{G_i}$$

where each $\boldsymbol{G_i}$ is a Givens rotation that zeros out one of the n subdiagonal elements of $\boldsymbol{A}$. Then

$$\boldsymbol{GA} = (\boldsymbol{G_n} \dots \boldsymbol{G_2 G_1})\boldsymbol{A} = \boldsymbol{R}$$

which is upper triangular. Also, as shown in part a, Givens rotations are orthogonal so

$$\boldsymbol{G^T GA} = \boldsymbol{G^T R}$$

and finally

$$\boldsymbol{A} = \boldsymbol{G^T R} = \boldsymbol{QR}$$

One Givens rotation will be needed for each subdiagonal element. When $n \leq m$, this will be the $(n-1)$th triangular number, which can be found as $\frac{1}{2}(n^2 - n)$. When $m > n$, we will have require the same number as in the previous case plus an additional $(m-n)$ rows of length n. Combining these yields the number of Givens rotations, i, as

$$i = \begin{cases} \frac{n}{2}(n-1), & n \leq m \\ \frac{n}{2}(n-1) + (m-n)n, & m > n \end{cases}$$

As shown above,

$$\boldsymbol{A} = \boldsymbol{G^T R} = \boldsymbol{QR}$$

so

$$\boldsymbol{Q} = \boldsymbol{G^T}$$

part c
See Jupyter notebook for implementation and output.

Problem 2
$x + (y+z) = (x+y) + z$ does not hold in floating point arithmetic due to the rounding that takes place at each step. For example, if

$\bar{x} = 5.112$

$\bar{y} = 5.112$

$\bar{z} = 5.113$

$\text{fl}(\text{fl}(\bar{x} + \bar{y}) + \bar{z}) = \text{fl}(\text{fl}(10.224) + 5.113) = \text{fl}(1.022 * 10^1 + 5.113) = \text{fl}(15.333) = 1.533 * 10^1$

$\text{fl}(\bar{x} + \text{fl}(\bar{y} + \bar{z})) = \text{fl}(5.112 + \text{fl}(10.225)) = \text{fl}(5.112 + 1.023 * 10^1) = \text{fl}(15.342) = 1.534 * 10^1$

Problem 3
part i

$$
\begin{aligned}
\sigma &= [\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2]^{1/2} \\
&= [\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\bar{x}^2)]^{1/2} \\
&= [\frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2\bar{x}x_i+\sum_{i=1}^{n}\bar{x}^2)]^{1/2} \\
&= [\frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-2\bar{x}\frac{n}{n}\sum_{i=1}^{n}x_i+n\bar{x}^2)]^{1/2} \\
&= [\frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-2\bar{x}n\bar{x}+n\bar{x}^2)]^{1/2} \\
&= [\frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)]^{1/2}
\end{aligned}
$$

part ii See Jupyter notebook for code and results.

Problem 4
part i

$$
\begin{aligned}
A1 &= \mathrm{fl}(4*\mathrm{fl}(\mathrm{fl}(\pi)*\mathrm{fl}(\mathrm{fl}(r)*\mathrm{fl}(r)))) \\
&= \mathrm{fl}(4*\mathrm{fl}(\mathrm{fl}(3.142)*\mathrm{fl}(\mathrm{fl}(6370)*\mathrm{fl}(6370)))) \\
&= \mathrm{fl}(4*\mathrm{fl}(3.142*\mathrm{fl}(6.370*10^3*6.370*10^3))) \\
&= \mathrm{fl}(4*\mathrm{fl}(3.142*\mathrm{fl}(40,576,900))) \\
&= \mathrm{fl}(4*\mathrm{fl}(3.142*4.058*10^7)) \\
&= \mathrm{fl}(4*\mathrm{fl}(127,502,360)) \\
&= \mathrm{fl}(4*1.275*10^8) \\
&= \mathrm{fl}(510,000,000) \\
&= 5.100*10^8
\end{aligned}
$$

part ii

$$\begin{aligned}
A2 &= \text{fl}(4 * \text{fl}(\text{fl}(\pi) * \text{fl}(\text{fl}(r) * \text{fl}(r)))) \\
&= \text{fl}(4 * \text{fl}(\text{fl}(3.142) * \text{fl}(\text{fl}(6371) * \text{fl}(6371)))) \\
&= \text{fl}(4 * \text{fl}(3.142 * \text{fl}(6.371 * 10^3 * 6.371 * 10^3))) \\
&= \text{fl}(4 * \text{fl}(3.142 * \text{fl}(40,589,641))) \\
&= \text{fl}(4 * \text{fl}(3.142 * 4.059 * 10^7)) \\
&= \text{fl}(4 * \text{fl}(127,533,780)) \\
&= \text{fl}(4 * 1.275 * 10^8) \\
&= \text{fl}(510,000,000) \\
&= 5.100 * 10^8
\end{aligned}$$

$$A2 - A1 = \text{fl}(5.100 * 10^8 - 5.100 * 10^8) = \text{fl}(0) = 0$$

part iii

$$\begin{aligned}
\frac{d}{dr}A &= \text{fl}(8\text{fl}(\pi r)) \\
&= \text{fl}(8\text{fl}(3.142 * 6.370 * 10^3)) \\
&= \text{fl}(8\text{fl}(20,014.54)) \\
&= \text{fl}(8 * 2.001 * 10^4)) \\
&= \text{fl}(160,080)) \\
&= 1.601 * 10^5
\end{aligned}$$

part iv
Part iii is more accurate.

part v
Part iii is more accurate because it is computing the difference directly. Part i and ii are the same in floating point arithmetic because the magnitude of the areas is much greater than the difference between them, so when the result of part i is subtracted from the result of part ii, catastrophic cancellation occurs. If floating point arithmetic with a longer mantissa is used, the difference between parts i and ii will be more accurate.