MA 423 Theory Assignment 3

Group 4

4 November 2020

1. An $n \times n$ matrix $A = [a_{ij}]$ is said to be upper Hessenberg if $a_{ij} = 0$ whenever $i > j + 1$. Prove the following for such a matrix.

   (a) Both factors of a QR decomposition of A can be computed in $O(n^2)$ flops.

   We need to find the flop count for computing both factors of QR decomposition for a hessenberg matrix. We show that we can find them in $O(n^2)$ flops using householder reflectors.

   Let $x \in \mathbb{R}^n \setminus 0$.We know there exits a householder matrix $Q = I_n - \gamma uu^T \in \mathbb{R}^{n\times n}$ s.t $Qx = y = [-\tau\ 0 \ldots\ 0]^T$ where $\tau = \pm||x||_2$,. Also $u = \frac{x-y}{x_1+\tau} \in \mathbb{R}^n$ and $\gamma = \frac{x_1+\tau}{\tau}$. u,γ and τ can be computed in $O(n)$ flops.

   The structure of our reflectors Q_i for an $n \times m$ matrix is such that they make all elements below the diagonal elements in i-th column zero. Then

$$Q_i = \left[\begin{array}{c|c} I_{i-1} & 0 \\ \hline 0 & I_{n-i+1} - \gamma uu^T \end{array}\right] = \left[\begin{array}{c|c} I_{i-1} & 0 \\ \hline 0 & \tilde{Q_i} \end{array}\right] \ \forall i = 1, \ldots, m$$

   Let A $= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ 0 & a_{23} & \cdots & a_{3n} \\ 0 & 0 & & a_{4n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots 0\ a_{n,n-1} & a_{nn}. \end{bmatrix}$. A is a hessenberg matrix as for $1 \le i, j \le n$ if $i > j + 1$, $a_{ij} = 0$.

   By multiplying orthogonal reflectors, we made all elements below diagonal elements in each column 0, to get an upper triangular matrix R. In case of hessenberg matrix, for i-th column, we only need to make (i,i+1) entry zero as all other entries are already zero due to the structure of hessenberg matrix.Therefore, $\tilde{Q}$ is a $(n - i + 1) \times (n - i + 1)$ matrix which only affects first 2 rows and leaves the rest unchanged. Therefore,

$$\tilde{Q_i} = \left[\begin{array}{c|c} Q_i' & 0 \\ \hline 0 & I_{n-i+1-2} \end{array}\right]$$

   where, $Q_i' = I_2 - \gamma uu^T$ where Q_1 is a 2×2 matrix and $u \in \mathbb{R}^2 \implies Q_i'$ can be computed in $O(1)$ flops which means that $\tilde{Q_i}$ and by extension Q_i is computed in $O(1)$ flops.

   $x = [a_{ii}\ a_{i,i+1}]^T \in \mathbb{R}^2$

   $Q_i'x = Q_i' \begin{bmatrix} a_{ii} \\ a_{i,i+1} \end{bmatrix} = \begin{bmatrix} \pm||A(i : i + 1, : i)||_2 \\ 0 \end{bmatrix} \ \forall i = 1, \ldots, n$

   and, $A = Q_1 \ldots Q_m R$, where R is an upper triangular matrix and Q_i is defined with $\tilde{Q_i}$, Q_i' as shown above.

   Computing Upper Triangular Matrix R
   As discussed in lectures, approximately $4nm$ flops are required to multiply a $n \times n$ reflector matrix with a $n \times m$ matrix, when done efficiently. For finding the upper triangular matrix, in the i-th column we made all the entries below the diagonal element zero by multiplying the reflector matrix, $\forall i = 1, \ldots, m$. Therefore flop count is

$$\sum_{i=1}^{m} 4(n - i + 1)(m - i)$$

   . In case of a hessenberg matrix, we have 2×2 square matrix as active reflector (Q_i') and a $2 \times (m - i)$ matrix (rest of the rows of matrix which needs to transformed as a result of application of reflector Q_i). Accounting for these conditions in the flop count expression above, we get

Flop count $= \sum_{i=1}^{n} 4(2)(n-i)$

$= 8(\sum_{i=1}^{n} n - \sum_{i=1}^{n} i)$

$= 8(n^2 - \frac{n(n+1)}{2}) = 4n^2 - 4n \implies O(n^2)\ flops$

Computing Orthogonal Matrix Q
We know that $A = Q_1 Q_2 \dots Q_m R = \hat{Q}R$. The isometry $Q = [\hat{Q}e_1 \dots \hat{Q}e_m]$. From the structure of Q_i given above, we can easily observe that each the first $i-1$ rows and $i-1$ columns of each Q_i is same as an $n \times n$ matrix. Using this fact we can see that $\hat{Q}e_1 = Q_1 Q_2 \dots Q_m e_1 = Q_1 e_1$ as $Q_i e_1 = e_1$ for $i > 1$ and similarly this holds $\forall\ i = 1, \dots m$. Therefore we have,

$$\hat{Q}e_k = Q_1 Q_2 \dots Q_k e_k$$

Recall the fact that $4nm$ flops are required to multiply a $n \times n$ reflector matrix with a $n \times m$ matrix, when done efficiently. When we multiply $Q_k e_k$, we are effectively multiplying e_k with the active reflector $\tilde{Q_k}$ (which is of size $(n-k+1) \times (n-k+1)$) which means that we essentially need, $4(n-k+1)(1)$ flops to do this. Then similarly we multiply reflector Q_{k-1} then Q_{k-2} and so on. Therefore we need $\sum_{j=1}^{k} 4(n-j+1)$ flops to compute $\hat{Q}_k e_k$, which is k-th column of isometry Q. Therefore, to compute all columns of Q, we need $\sum_{k=1}^{m} \sum_{j=1}^{k} 4(n-j+1)$ flops.
Now, for hessenberg matrix, we only needed to make only one element below the diagonal zero as opposed to $n-i$ in the general case. Therefore, our active reflector $Q_i^{'}$ (which is a component of the $\tilde{Q}_i$ matrix for which the summation was derived for) is always of the size 2×2. Plugging this in the summation above, we get

$$\sum_{k=1}^{n} \sum_{j=1}^{k} 4(2) = \sum_{k=1}^{n} 8k = 4n(n+1) \implies O(n^2)\ flops$$

Therefore we can compute both Q and R in $O(n^2)$ flops.

(b) If A is additionally tridiagonal, that is, it has nonzero entries only on the main diagonal and on the first super-diagonal, and the first sub-diagonal, then the cost of computing Q and R is $O(n)$ flops.

Consider the structure of a tridiagonal matrix,

$$\begin{bmatrix} a_{11} & a_{12} & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & & \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & a_{n-1n} \\ 0 & & & a_{nn-1} & a_{nn} \end{bmatrix}$$

Since a tridiagonal matrix is also hessenberg, we can assume that we can similarly compute the reflector matrices Q_i in $O(1)$ flops. Now, while computing R, due to the structure of tridiagonal matrix we can further make another simplification. When we multiply our active reflector matrix Q_i^i with the $2 \times (n-1)$ matrix, we see the all the entries in this matrix beyond column 2 are zero. Therefore we call this matrix A_i and partition it as

$$A_i = [\ A_i^{'} \mid 0\ ]$$

where $A_i^{'}$ is 2×2 matrix.

Therefore $Q_i^{'} A_i = Q_i^{'}[A_i^{'} \mid 0] = Q_i^{'} A_i^{'}\ \forall i = 1, \dots, n$
Therefore, in the flop count expression for computing R in part (a) we can replace (n-i) with 2.
$\implies$ flop count $= \sum_{i=1}^{n} 4(2)(2) = 16n \implies O(n)\ flops$

2. Let A be any $n \times m$ real matrix. Suppose B is a real $m \times n$ matrix such that $ABA = A$, $BAB = B$, $(AB)^T = AB$ and $(BA)^T = BA$. Prove that B is the Moore Penrose pseudoinverse of A.

Proof:
We first show that if $A^\dagger$ is the Moore-Penrose inverse of then it satisfies the following

(i) $A^\dagger = V_r \Sigma_r^{-1} U_r$
(ii) $(AA^\dagger)^* = AA^\dagger$
(iii) $(A^\dagger A)^* = A^\dagger A$
(iv) $A^\dagger AA^\dagger = AA^\dagger$
(v) $AA^\dagger A = A$

We know that $A^\dagger = V\Sigma^\dagger U^*$, $\Sigma^\dagger = diag(\sigma_1^{-1}, \cdots \sigma_r^{-1}, 0 \cdots 0) \in \mathbb{R}^{m\times n}$

Proof(i): Since $A^\dagger = V\Sigma^\dagger U^*$ is a SVD decomposition as V is a $\mathbb{F}^{m\times m}$ unitary matrix and U is a $\mathbb{F}^{n\times n}$ unitary matrix and $\Sigma^\dagger = diag(\sigma_1^{-1}, \cdots \sigma_r^{-1}, 0 \cdots 0) \in \mathbb{F}^{m\times n}$ is a singular matrix therefore we can write it as

$$A^\dagger = V_r \Sigma_r^{-1} U_r$$

Proof(ii): $(AA^\dagger)^* = (U_r\Sigma_r V_r^* V_r \Sigma_r^{-1} U_r^*)^*$ We have

$$V_r^* V_r = \begin{bmatrix} v_1^* \\ \vdots \\ v_r^* \end{bmatrix} \begin{bmatrix} v_1^* & \cdots & v_r^* \end{bmatrix} = I_{r\times r}$$

So $(AA^\dagger)^* = (U_r\Sigma_r I_{r\times r} \Sigma_r^{-1} U_r^*)^* = (U_r U_r^*)^* = U_r U_r^*$

$U_r U_r^* = U_r \Sigma_r V_r^* V_r \Sigma_r^{-1} U_r^* = AA^\dagger$
Proof(iii): Similiar to (ii)
$(A^\dagger A)^* = (V_r\Sigma_r^{-1}U_r^* U_r \Sigma_r V_r^*)^* = (V_r \Sigma_r^{-1} I_{r\times r} \Sigma_r V_r^*)^* = (V_r V_r^*)^* = V_r V_r^*$
$V_r V_r^* = (V_r \Sigma_r^{-1} U_r^* U_r \Sigma_r V_r^*) = A^\dagger A$

Proof(iv): $A^\dagger AA^\dagger = (V_r\Sigma_r^{-1}U_r^*)(U_r\Sigma_r V_r^*)(V_r\Sigma_r^{-1}U_r^*)$
$A^\dagger AA^\dagger = (V_r\Sigma_r^{-1}\Sigma_r\Sigma_r^{-1}U_r^*) = V_r\Sigma_r^{-1}U_r^* = A^\dagger$

Proof(v): Similiar to (iv)
$AA^\dagger A = (U_r\Sigma_r V_r^*)(V_r\Sigma_r^{-1}U_r^*)(U_r\Sigma_r V_r^*)$
$AA^\dagger A = (U_r\Sigma_r\Sigma_r^{-1}\Sigma_r V_r^*) = U_r\Sigma_r V_r^* = A$

Now that we have proven these 5 properties of Moore Penrose Pseudoinverse we'll use them in our main proof

$$\begin{aligned} AA^\dagger &= (ABA)A^\dagger = (AB)(AA^\dagger) \\ &= (AB)^T(AA^\dagger)^T = B^T(AA^\dagger A)^T = B^T A^T \\ &= (AB)^T = AB \end{aligned} \tag{1}$$

$$\begin{aligned} A^\dagger A &= A^\dagger(ABA) = (A^\dagger A)(BA) \\ &= (A^\dagger A)^T(BA)^T = (AA^\dagger A)^T B^T = A^T B^T \\ &= (BA)^T = BA \end{aligned} \tag{2}$$

$$\begin{aligned} A^\dagger &= A^\dagger AA^\dagger \\ &= A^\dagger AB \quad (\text{ using 1}) \\ &= BAB \quad (\text{ using 2}) \\ &= B \end{aligned} \tag{3}$$

Hence proved that B is the Moore Penrose Pseudoinverse of A