Mixed-precision Block QR Decomposition on GPU

Jaidon Lybbert

February 2, 2023

Contents

1 QR Decomposition

The QR decomposition of an m-by-n matrix A with $m > n$, is the matrix product $A = QR$, where Q is an m-by-n unitary matrix, and R is upper triangular.

1.1 Matrix Q

The matrix Q is a transformation which preserves inner products of column vectors of R. If the inner product space is real, the matrix Q is equivalently orthogonal. One possibility of such a transformation is a rotation.

Another possibility of such an orthogonal transformation is a reflection. The matrix Q in general is a combination of rotations and reflections.

1.2 Matrix R

The matrix R is upper triangular, a form which has the following useful properties: (I) the determinant is equal to the product of the diagonal elements, (II) the eigenvalues are equal to the diagonal elements, (III) given the linear system $Rx = b$ it is easy to solve for x by back substitution.

2 Computation

In order to compute the decomposition of A, the matrix is iteratively transformed by unitary matrices $\{U_i : 0 < i < k\}$ until the product is upper triangular. This upper triangular matrix is the matrix R in $A = QR$

$$R = U_k U_{k-i} \dots U_1 A. \tag{1}$$

It follows, that the matrix Q is composed of the set of inverse transformations

$$Q = U_1^T U_2^T \dots U_k^T. \tag{2}$$

The key to solving for R is to choose transformations U_i which produce zeros below the diagonal of the matrix product

$$P = U_i \dots U_1 A, \tag{3}$$

and can iteratively be applied to converge to R as quickly as possible. Two choices for U_i are Householder reflections, and Givens rotations.

2.1 Householder Reflections

The Householder reflection is a unitary transformation represented by a matrix $H \in \mathbb{R}^{N \times N}$ which reflects a vector $x \in \mathbb{R}^N$ across a hyperplane defined by its unit normal vector $\{w \in \mathbb{R}^N : \|w\| = 1\}$. The transformation matrix is given by

$$H = I - 2\mathbf{w}\mathbf{w}^T \tag{4}$$

where $I \in \mathbb{R}^{N \times N}$ is the identity matrix. [1]

In order to get the upper triangular matrix $R \in \mathbb{R}^{N \times N}$ given a matrix $A \in \mathbb{R}^{M \times N}$ using householder reflections, each column is treated in sequence left to right.

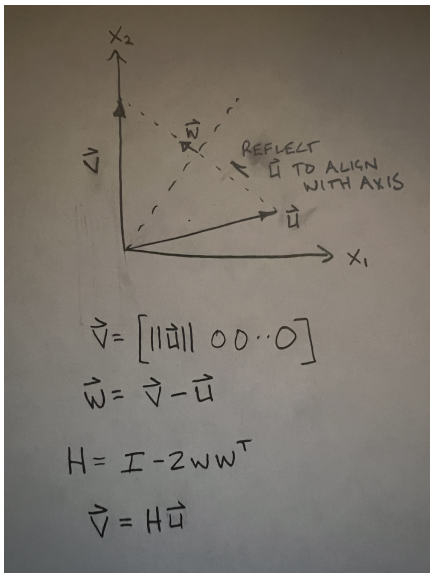

Figure 1: Geometric illustration of the reflection of a vector to an axis. The result of this transformation is that the vector now only has one non-zero component.

2.2 Givens Rotations

A Givens rotation is a unitary transformation which rotates a vector x counterclockwise in a chosen plane. For example, possible Givens rotation matrices in $\mathbb{R}^4$ include

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c & -s & 0 \\ 0 & s & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, or \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c & -s \\ 0 & 0 & s & c \end{bmatrix}, \tag{5}$$

where $c = \cos\theta$ and $s = \sin\theta$. Each of these examples have the effect of rotating the vector in different planes.

The Givens rotation is parallelizable because it effects only two dimensions of the input vector. For example, the second and last transformations above could simultaneously be computed on the vector x, then the results combined by selecting the first 2 dimensions of the first result, and the second two dimensions of the second result.

A Givens rotation can easily be computed to introduce zeros in the matrix P. The scalars c and s can be computed directly from elements in P in order to zero out targeted elements. For example, say we want to zero out element

a_{21} in the matrix

$$P = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. \tag{6}$$

We target the second dimension of the column vector, so we rotate on the plane spanned by the first two dimensions. We don't choose the plane spanned by the second and third dimensions, because we would end up losing the zero in the third row in the process. The Givens rotation to rotate on this plane is of the form

$$G = \begin{bmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

which will leave the third row of P unmodified. We are aligning the column vector with the axis of the first dimension, making the component of the vector along the second dimension zero. Below is a geometric illustration of the rotation.

The scalars c and s of matrix G are computed directly from the values in matrix P by the equations

$$c = \frac{a_{11}}{r}, \tag{8}$$

$$s = -\frac{a_{21}}{r}, \tag{9}$$

where

$$r = \sqrt{a_{11}^2 + a_{21}^2} \tag{10}$$

The transformation to introduce the zero is then

$$P = GP_{prior} = \begin{bmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{11}$$

$$P = GP_{prior} = \begin{bmatrix} a_{11}/r & a_{21}/r & 0 \\ -a_{21}/r & a_{11}/r & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{12}$$

where . Multiplying through

$$P = GP_{prior} = \begin{bmatrix} a_{11}/r & a_{21}/r & 0 \\ -a_{21}/r & a_{11}/r & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{13}$$

$$P = \begin{bmatrix} \frac{a_{11}a_{11}+a_{21}a_{21}}{r} & \frac{a_{11}a_{12}+a_{21}a_{22}}{r} & \frac{a_{11}a_{13}+a_{21}a_{23}}{r} \\ \frac{-a_{21}a_{11}+a_{11}a_{21}}{r} & \frac{-a_{21}a_{12}+a_{11}a_{22}}{r} & \frac{-a_{21}a_{13}+a_{11}a23}{r} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{14}$$

the zero is introduced in the desired location.

References

[1] Bhaskar Dasgupta. *Applied Mathematical Methods*. Pearson, 1986.

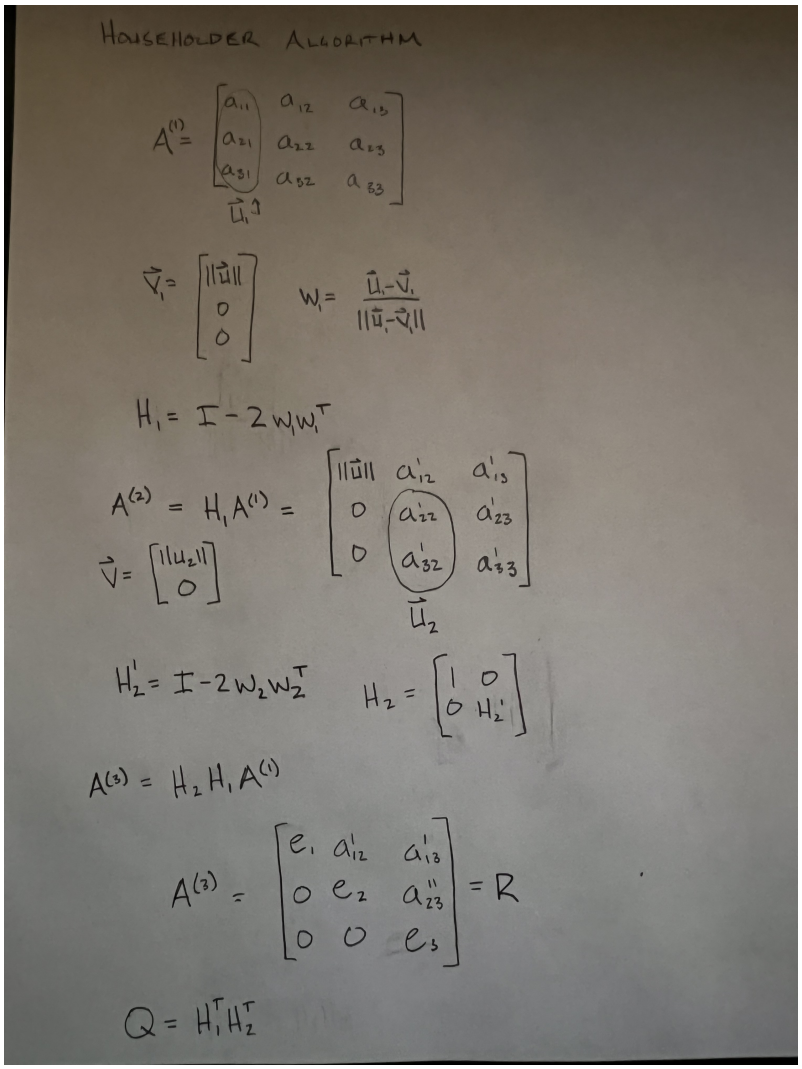

Figure 2: QR factorization algorithm with Householder reflections

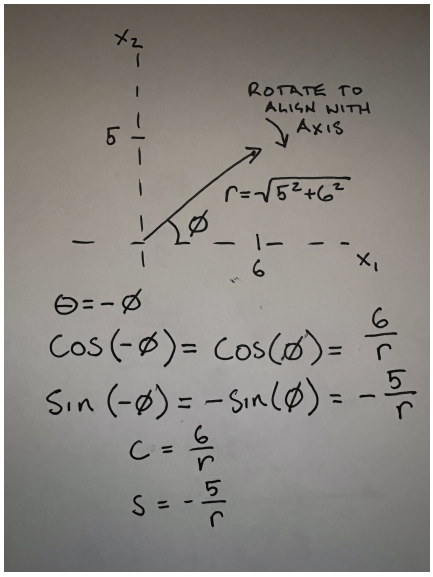

Figure 3: Geometric illustration of the rotation of a vector in $\mathbb{R}^3$ about the axis of basis vector $x3$ to align with the basis vector $x1$. The result of this transformation is that the component of the transformed vector in the direction of the basis vector $x2$ is zero, corresponding to a zero introduced in the transformed matrix.