17.4 QR factorization.

When solving a square linear system $Ax = y$, the basic idea is to bring the system into a simpler equivalent form $Ax = LUx = y$ or $Ux = L^{-1}y$ where the matrix U is upper triangular. In this section we shall develop a similar technique for overdetermined or rectangular linear systems.

When trying to solve an overdetermined $m \times n$ system $Ax = y$ then at best, when the rank of the matrix A enlarged with the column y is $n + 1$ (we have already assumed that A has maximal column rank n), one can compute x so as to minimize the residual:

$$\min_{x \in \mathbb{R}^n} ||Ax - y||_2 = \min_{x \in \mathbb{R}^n} \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j - y_i \right)^2} \tag{17.12}$$

We know from *Lemma* 17.1 that the solution of ('`eq:lstsqrq`') is unique. To factorize the rectangular matrix A in a useful way, we need the notion of orthogonal matrix. A square matrix Q is called orthogonal if

$$Q^T Q = QQ^T = I$$

where the superscript ^{T} denotes the transpose and I is the identity matrix. In other words, the inverse of a square orthogonal matrix is given by its transpose. It is apparent that if Q is orthogonal, so is Q^T and that the product of orthogonal matrices is orthogonal. The column vectors of an $m \times m$ orthogonal matrix form an orthogonal basis for $\mathbb{R}^m$ because their dotproduct evaluates to the Kronecker delta. The dotproduct of the i^{th} and j^{th} column of Q is given by

$$\sum_{k=1}^{m} Q_{ki} Q_{kj} = (Q^T Q)_{ij} = \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Orthogonal matrices are special in the sense that they preserve norms:

$$||Qr||_2^2 = \sum_{i=1}^{m} (Qr)_i^2 = (Qr)^T (Qr) = r^T Q^T Qr = r^T r = ||r||_2^2$$

and this is useful in the context of (17.12) because for an orthogonal matrix Q^T,

$$||Ax - y||_2^2 = ||Q^T Ax - Q^T y||_2^2$$

Suppose we can determine an orthogonal $m \times m$ matrix Q^T such that $Q^T A$ has a rectangular upper triangular structure, meaning that the upper $n \times n$ submatrix

is upper triangular and the lower $m-n$ rows are filled with zeroes:

$$Q^TA = \begin{pmatrix} * & \dots & & * \\ 0 & * & \dots & * \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & * \\ 0 & \dots & & 0 \\ \vdots & & & \vdots \\ 0 & & \dots & 0 \end{pmatrix}$$

Then the unique solution x of

$$\begin{aligned} \min_{x\in\mathbb{R}^n} ||Ax-y||_2 &= \min_{x\in\mathbb{R}^n} ||Q^TAx - Q^Ty||_2 \\ &= \sqrt{\sum_{i=n+1}^{m} (Q^Ty)_i^2} \end{aligned}$$

because x can be determined so as to satisfy the first n equations and the lower $m-n$ rows of Q^TA do not contain nonzero elements. Such an orthogonal matrix Q^T for which $Q^TA = R$ with R rectangular upper triangular, apparently allows to factorize A as $A = Q(Q^TA) = QR$, from which the terminology QR factorization. A simple example of a 2×2 orthogonal matrix is

$$Q^T = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

and a general $m\times m$ example is given by

$$G(i,k,\theta_{kj}) = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & \cos\theta_{kj} & & & & \sin\theta_{kj} & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & -\sin\theta_{kj} & & & & \cos\theta_{kj} & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \tag{17.13}$$

where the entries $\cos\theta_{kj}$ appear in the diagonal positions (i,i) and (k,k) and the entries $\sin\theta_{kj}$ and $-\sin\theta_{kj}$ respectively in (i,k) and (k,i). Matrices like Q^T and $G(i,k,\theta_{kj})$ above are called rotation matrices. A similar matrix was involved in the rotation of a robot over an angle $-\phi$ in *Section* 14.1. The value of θ_{kj} in

the orthogonal matrix $G(i,k,\theta_{kj})$ will be determined such that in the product $G(i,k,\theta_{kj})A$ a zero is created in position (k,j).

From the structure of the rotation matrix $G(i,k,\theta_{kj})$ it is easy to see that computing the product $G(i,k,\theta_{kj})A$ only affects the i^{th} and k^{th} rows of A. They are given by

$$\begin{pmatrix} \cos\theta_{kj} & \sin\theta_{kj} \\ -\sin\theta_{kj} & \cos\theta_{kj} \end{pmatrix} \begin{pmatrix} a_{i1} & \dots & a_{in} \\ a_{k1} & \dots & a_{kn} \end{pmatrix}$$

Choosing θ_{kj} such that an entry $\left(G(i,k,\theta_{kj})A\right)_{kj} = 0$ with $k > j$, which is beneath the diagonal, can be done by choosing

$$\cos\theta_{kj} = a_{ij}/\sqrt{a_{ij}^2 + a_{kj}^2} \qquad \sin\theta_{kj} = a_{kj}/\sqrt{a_{ij}^2 + a_{kj}^2}$$

which implies that

$$-a_{ij}\sin\theta_{kj} + a_{kj}\cos\theta_{kj} = 0$$

In order to avoid the danger of overflow or underflow while computing $a_{ij}^2 + a_{kj}^2$, the value θ_{kj} is better obtained as follows. If $|a_{kj}| \leq |a_{ij}|$ then set $\tan\theta_{kj} = a_{kj}/a_{ij}$ which is in magnitude at most 1. The values $\cos\theta_{kj}$ and $\sin\theta_{kj}$ are then given by

$$\begin{aligned} \cos\theta_{kj} &= 1/\sqrt{1+\tan^2\theta_{kj}} \\ \sin\theta_{kj} &= \tan\theta_{kj}\ \cos\theta_{kj} \end{aligned} \tag{17.14}$$

If $|a_{ij}| < |a_{kj}|$ then $\cot\theta_{kj} = a_{ij}/a_{kj}$ is computed with (17.14) still valid. When creating one zero at position (k,j) involves the rows i and k when premultiplying with $G(i,k,\theta_{kj})$, then the order in which the zeroes are created is important. One should not fill in previously zeroed positions.

With $i = k-1$, the columns j and rows k should be traversed from left to right and bottom up. In other words, in column number j with $j = 1,\dots,n$ one deals with the the rows $k = m,\dots,j+1$ in that order. So the first rotation matrix to be used is $G(m-1,m,\theta_{m1})$ to create a zero entry on position $(m,1)$, followed by premultiplication with rotation matrices $G(k-1,k,\theta_{k1})$ where $k = m-1,\dots,2$ creating zeroes in the first column beneath the diagonal. Then zeroes beneath the diagonal in the second column are created, to top it off in the end with $G(n,n+1,\theta_{n+1,n})$ which creates a zero entry in position $(n+1,n)$ just beneath the diagonal in the last column. The orthogonal matrix Q^T we are looking for is given by the product of all rotation matrices involved:

$$Q^T = \prod_{k=n+1}^{m} G(k-1,k,\theta_{kn}) \times \dots \prod_{k=2}^{m} G(k-1,k,\theta_{k1})$$

Finally

$$||Ax - y||_2 = ||Q^T(Ax-y)||_2 = ||Rx - Q^T y||_2$$

where the rectangular matrix R consists of an upper triangular square matrix U and $m-n$ bottom zero rows. Let us denote this by

$$R = \begin{pmatrix} U \\ 0 \end{pmatrix}$$

and let us introduce the notations z_1 for the vector consisting of the first n entries of $Q^T y$ and z_2 for the vector consisting of the bottom $m-n$ entries of $Q^T y$, which can be explicited as

$$Q^T y = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

If $\tilde{x}$ is the computed solution of $Ux = z_1$, then

$$||A\tilde{x} - y||_2^2 = \left|\left|\begin{pmatrix} U \\ 0 \end{pmatrix}\tilde{x} - \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right|\right|_2^2 = ||U\tilde{x} - z_1||_2^2 + ||z_2||_2^2$$

This implies that

$$||A\tilde{x} - y||_2 \geq ||z_2||_2$$

while at the same time

$$||A\tilde{x} - y||_2 \leq ||z_2||_2 + ||U\tilde{x} - z_1||_2$$

17.5 Conditioning and stability.

For square matrices condition numbers were defined in (14.3). A condition number is large if a matrix is close to singular, not the other way around. The mere fact of a large condition number does not imply that a matrix is nearly singular. So for rectangular matrices a definition of condition number should again be inversely proportional to the distance of the matrix to its closest rank-deficient matrix. Remember that the rank of an $m \times n$ matrix is at most $\min(m, n)$ and hence for $m \gg n$, the rank of A is at most n. A rank-deficient $m \times n$ matrix therefore has rank less than n. Consequently a definition for the condition number of a rectangular matrix A is

$$\kappa_2(A) = \frac{1}{\min_{rank(A+D)<n} ||D||_2/||A||_2}$$

An estimate for $\kappa_2(A)$ can be obtained from an equivalent definition for $\kappa_2(A)$, based on the notion of singular values.

DEFINITION 17.1

If $A \in \mathbb{R}^{m\times n}$ then orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ exist such that A can be factored as $A = U\Sigma V^T$ where Σ is a diagonal matrix. The nonzero entries $\sigma_i, i = 1, \ldots, \min(n, m)$ on the diagonal of Σ are called the singular values of A.

These singular values σ_i are all positive. For a square matrix A, it can be shown that $||A||_2 = \max_{1\le i\le n} \sigma_i$. For a rectangular $m \times n$ matrix A one can prove that

$$\kappa_2(A) = \frac{\max_{i=1,\ldots,\min(n,m)} \sigma_i}{\min_{i=1,\ldots,\min(n,m)} \sigma_i}$$

The QR factorization obeys the same nice property as the LU decomposition. A backward error analysis indicates that for the computed upper triangular factor $\tilde{R}$ there exists an orthogonal $m \times m$ matrrix Q' such that $\tilde{R}$ can be regarded as the exact factor of a slightly perturbed rectangular matrix A:

$$(A + E) = Q'\tilde{R} \qquad ||E||_2 \le cn^{3/2}m \,\texttt{ULP}\, ||A||_2$$

Note that the above statement is made about the difference of the factorizations, $E = Q'\tilde{R} - QR$ and not about the difference of the factors $\tilde{R}$ and R! Hence the QR factorization is a stable numerical procedure.

After computing the factorization, remains to perform the backsubstitution part of the Gaussian elimination procedure when computing

$$\tilde{R}\tilde{x} = Q'^T y$$

The effect of the condition number $\kappa_2(A)$ of the rectangular matrix A on the quality of the solution as outlined in *Theorem* 17.1, is also similar to the effect detailed in *Theorem* 14.1.

THEOREM 17.1

If $x^\star$ denotes the exact solution of (17.8) and if its computed solution $\tilde{x}$ satisfies

$$||(A + E)\tilde{x} - y||_2 \text{ minimal} \qquad \begin{array}{l} ||E||_2 \le c(n, m)\texttt{ULP}\, ||A||_2 \\ c(n, m)\texttt{ULP} \; < 1/\kappa_2(A) \end{array}$$

then, if $A + E$ has maximal column rank n,

$$||x^\star - \tilde{x}||_2 \le c(n, m)\texttt{ULP}\, \kappa_2(A)||x^\star||_2 \left(2 + \frac{||Ax^\star - y||_2(\kappa_2(A) + 1)}{||A||_2||x^\star||_2}\right)$$

When the given data y can be approximated by an appropriate model, making the columns of A almost linearly dependent, or $r = Ax^\star - y \approx 0$, then *Theorem* 17.1 simplifies to

$$\frac{||x^\star - \tilde{x}||_2}{||x^\star||_2} = O\left(c(n,m)\texttt{ULP}\ \kappa_2(A)\right)$$

Solving a square linear system of equations does not introduce a truncation error, because Gaussian elimination never introduces an approximation to the given matrix or right hand side. It deals with the machine representation of the linear system throughout the entire algorithm. When solving an overdetermined linear system of equations by means of QR factorization, the situation is very similar. The problem to be solved is a minimization problem and QR factorization is able to construct its solution without introducing approximations to the original data or problem. The vector z_2 consisting of the bottom $m - n$ components of the transformed right hand side $Q^T y$, is not a truncation error. In exact arithmetic

$$\min_{x \in \mathbb{R}^n} ||Ax - y||_2 = ||z_2||_2$$

But of course $||z_2||_2$ is influenced by the choice of the basis functions $f_j(x)$ in (17.3) which can be more or less appropriate for the data y_i as we can learn from the next example.