19.3. Error Analysis of Householder Computations

It is well known that computations with Householder matrices are very stable. Wilkinson showed that the computation of a Householder vector, and the application of a Householder matrix to a given matrix, are both normwise stable, in the sense that the computed Householder vector is very close to the exact one and the computed update is the exact update of a tiny normwise perturbation of the original matrix [1233, 1965, pp. 153–162, 236], [1234, 1965]. Wilkinson also showed that the Householder QR factorization algorithm is normwise backward stable [1233, p. 236]. In this section we give a columnwise error analysis of Householder matrix computations. The columnwise bounds provide extra information over the normwise ones that is essential in certain applications (for example, the analysis of iterative refinement).

In the following analysis it is not worthwhile to evaluate the integer constants in the bounds explicitly, so we make frequent use of the notation

$$\tilde{\gamma}_k = \frac{cku}{1 - cku}$$

introduced in (3.8), where c denotes a small integer constant.

Lemma 19.1. *Let $x \in \mathbb{R}^n$. Consider the following two constructions of $\beta \in \mathbb{R}$ and $v \in \mathbb{R}^n$ such that $Px = \sigma e_1$, where $P = I - \beta vv^T$ is a Householder matrix with $\beta = 2/(v^T v)$:*

% "Usual" choice of sign, (19.1):	% Alternative choice of sign, (19.2):
% $\operatorname{sign}(\sigma) = -\operatorname{sign}(x_1)$.	% $\operatorname{sign}(\sigma) = \operatorname{sign}(x_1)$.
$v = x$	$v = x$
$s = \operatorname{sign}(x_1)\|x\|_2$ % $\sigma = -s$	$s = \operatorname{sign}(x_1)\|x\|_2$ % $\sigma = s$
$v_1 = v_1 + s$	Compute v_1 from (19.2)
$\beta = 1/(sv_1)$	$\beta = 1/(sv_1)$

In floating point arithmetic the computed $\widehat{\beta}$ and $\widehat{v}$ from both constructions satisfy $\widehat{v}(2\colon n) = v(2\colon n)$ and

$$\widehat{\beta} = \beta(1 + \tilde{\theta}_n), \qquad \widehat{v}_1 = v_1(1 + \tilde{\theta}_n),$$

where $|\tilde{\theta}_n| \le \tilde{\gamma}_n$.

Proof. We sketch the proof for the first construction. Each occurrence of δ denotes a different number bounded by $|\delta| \le u$. We compute $fl(x^Tx) = (1+\theta_n)x^Tx$, and then $fl(\|x\|_2) = (1+\delta)(1+\theta_n)^{1/2}(x^Tx)^{1/2} = (1+\theta_{n+1})\|x\|_2$ (the latter term $1+\theta_{n+1}$ is suboptimal, but our main aim is to keep the analysis simple). Hence $\widehat{s} = (1+\theta_{n+1})s$.

For notational convenience, define $w = v_1 + s$. We have $\widehat{w} = (v_1 + \widehat{s})(1+\delta) = w(1+\theta_{n+2})$ (essentially because there is no cancellation in the sum). Hence

$$\widehat{\beta} = fl(1/(\widehat{s}\widehat{w})) = \frac{(1+\delta)^2}{(1+\theta_{n+1})s(1+\theta_{n+2})w}$$
$$= \frac{(1+\delta)^2}{(1+\theta_{2n+3})sw} = (1+\theta_{4n+8})\beta.$$

The proof for the second construction is similar. □

For convenience we will henceforth write Householder matrices in the form $I - vv^T$, which requires $\|v\|_2 = \sqrt{2}$ and amounts to redefining $v := \sqrt{\beta}v$ and $\beta := 1$ in the representation of Lemma 19.1. We can then write, using Lemma 19.1,

$$\widehat{v} = v + \Delta v, \qquad |\Delta v| \le \tilde{\gamma}_m|v| \quad (v \in \mathbb{R}^m, \ \ \|v\|_2 = \sqrt{2}), \tag{19.5}$$

where, as required for the next two results, the dimension is now m.

The next result describes the application of a Householder matrix to a vector, and is the basis of all the subsequent analysis. In the applications of interest P is defined as in Lemma 19.1, but we will allow P to be an arbitrary Householder matrix. Thus v is an arbitrary, normalized vector, and the only assumption we make is that the computed $\widehat{v}$ satisfies (19.5).

Lemma 19.2. *Let $b \in \mathbb{R}^m$ and consider the computation of $y = \widehat{P}b = (I - \widehat{v}\widehat{v}^T)b = b - \widehat{v}(\widehat{v}^Tb)$, where $\widehat{v} \in \mathbb{R}^m$ satisfies* (19.5). *The computed $\widehat{y}$ satisfies*

$$\widehat{y} = (P + \Delta P)b, \qquad \|\Delta P\|_F \le \tilde{\gamma}_m,$$

where $P = I - vv^T$.

Proof. (Cf. the proof of Lemma 3.9.) We have

$$\widehat{w} := fl(\widehat{v}(\widehat{v}^Tb)) = (\widehat{v} + \Delta\widehat{v})(\widehat{v}^T(b + \Delta b)),$$

where $|\Delta\widehat{v}| \le u|\widehat{v}|$ and $|\Delta b| \le \gamma_m|b|$. Hence

$$\widehat{w} = (v + \Delta v + \Delta\widehat{v})(v + \Delta v)^T(b + \Delta b) =: v(v^Tb) + \Delta w,$$

where $|\Delta w| \le \tilde{\gamma}_m|v||v^T||b|$. Then

$$\widehat{y} = fl(b - \widehat{w}) = b - v(v^Tb) - \Delta w + \Delta y_1, \qquad |\Delta y_1| \le u|b - \widehat{w}|.$$

We have

$$|-\Delta w + \Delta y_1| \le u|b| + \tilde{\gamma}_m|v||v^T||b|.$$

Hence $\widehat{y} = Pb + \Delta y$, where $\|\Delta y\|_2 \le \tilde{\gamma}_m\|b\|_2$. But then $\widehat{y} = (P + \Delta P)b$, where $\Delta P = \Delta y b^T/b^Tb$ satisfies $\|\Delta P\|_F = \|\Delta y\|_2/\|b\|_2 \le \tilde{\gamma}_m$. □

Next, we consider a sequence of Householder transformations applied to a matrix. Again, each Householder matrix is arbitrary and need have no connection to the matrix to which it is being applied. In the cases of interest, the Householder matrices P_k have the form (19.4), and so are of ever-decreasing effective dimension, but to exploit this property would not lead to any significant improvement in the bounds. Since the P_j are applied to the columns of A, columnwise error bounds are to be expected, and these are provided by the next lemma.

We will assume that

$$r\tilde{\gamma}_m < \frac{1}{2}, \tag{19.6}$$

where r is the number of Householder transformations. We will write the jth columns of A and ΔA as a_j and Δa_j, respectively.

Lemma 19.3. *Consider the sequence of transformations*

$$A_{k+1} = P_k A_k, \qquad k = 1\colon r,$$

where $A_1 = A \in \mathbb{R}^{m\times n}$ and $P_k = I - v_k v_k^T \in \mathbb{R}^{m\times m}$ is a Householder matrix. Assume that the transformations are performed using computed Householder vectors $\widehat{v}_k \approx v_k$ that satisfy (19.5). *The computed matrix $\widehat{A}_{r+1}$ satisfies*

$$\widehat{A}_{r+1} = Q^T(A + \Delta A), \tag{19.7}$$

where $Q^T = P_r P_{r-1} \dots P_1$ and

$$\|\Delta a_j\|_2 \le r\tilde{\gamma}_m \|a_j\|_2, \qquad j = 1\colon n. \tag{19.8}$$

In the special case $n = 1$, so that $A \equiv a$, we have $\widehat{a}^{(r+1)} = (Q + \Delta Q)^T a$ with $\|\Delta Q\|_F \le r\tilde{\gamma}_m$.

Proof. The jth column of A undergoes the transformations $a_j^{(r+1)} = P_r \dots P_1 a_j$. By Lemma 19.2 we have

$$\widehat{a}_j^{(r+1)} = (P_r + \Delta P_r)\dots(P_1 + \Delta P_1)a_j, \tag{19.9}$$

where each ΔP_k depends on j and satisfies $\|\Delta P_k\|_F \le \tilde{\gamma}_m$. Using Lemma 3.7 we obtain

$$\begin{aligned} &\widehat{a}_j^{(r+1)} = Q^T(a_j + \Delta a_j), \\ &\|\Delta a_j\|_2 \le \left((1+\tilde{\gamma}_m)^r - 1\right)\|a_j\|_2 \le \frac{r\tilde{\gamma}_m}{1 - r\tilde{\gamma}_m}\|a_j\|_2 = r\tilde{\gamma}'_m\|a_j\|_2, \end{aligned} \tag{19.10}$$

using Lemma 3.1 and assumption (19.6). Finally, if $n = 1$, so that A is a column vector, then (as in the proof of Lemma 19.2) we can rewrite (19.7) as $\widehat{a}^{(r+1)} = (Q + \Delta Q)^T a$, where $\Delta Q^T = (Q^T \Delta a)a^T/a^T a$ and $\|\Delta Q\|_F = \|\Delta a\|_2/\|a\|_2 \le r\tilde{\gamma}_m$. □

Recall that columnwise error bounds are easily converted into normwise ones (Lemma 6.6). For example, (19.8) implies $\|\Delta A\|_F \le r\tilde{\gamma}_m \|A\|_F$.

Lemma 19.3 yields the standard backward error result for Householder QR factorization.

Theorem 19.4. *Let $\widehat{R} \in \mathbb{R}^{m\times n}$ be the computed upper trapezoidal QR factor of $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ obtained via the Householder QR algorithm (with either choice of sign, (19.1) or (19.2)). Then there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

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

where

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2, \qquad j = 1\colon n. \tag{19.11}$$

The matrix Q is given explicitly as $Q = (P_nP_{n-1}\dots P_1)^T$, where P_k is the Householder matrix that corresponds to the exact application of the kth step of the algorithm to $\widehat{A}_k$.

Proof. This is virtually a direct application of Lemma 19.3, with P_k defined as the Householder matrix that produces zeros below the diagonal in the kth column of the computed matrix $\widehat{A}_k$. One subtlety is that we do not explicitly compute the lower triangular elements of $\widehat{R}$, but rather set them to zero explicitly. However, it is easy to see that the conclusions of Lemmas 19.2 and 19.3 are still valid in these circumstances; the essential reason is that the elements of ΔPb in Lemma 19.2 that correspond to elements that are zeroed by the Householder matrix P are forced to be zero, and hence we can set the corresponding rows of ΔP to zero too, without compromising the bound on $\|\Delta P\|_F$. □

We note that for Householder QR factorization $\Delta P_k = 0$ for $k > j$ in (19.9), and consequently the factor $\tilde{\gamma}_{mn}$ in (19.11) can be reduced to $\tilde{\gamma}_{mj}$.

Theorem 19.4 is often stated in the weaker form $\|\Delta A\|_F \leq \tilde{\gamma}_{mn}\|A\|_F$ that is implied by (19.11) (see, e.g., [509, 1996, §5.2.1]). For a matrix whose columns vary widely in norm this normwise bound on ΔA is much weaker than (19.11). For an alternative way to express this backward error result define B by $A = BD_C$, where $D_C = \operatorname{diag}(\|A(:,j)\|_2)$; then the result states that there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that

$$(B + \Delta B)D_C = Q\widehat{R}, \qquad \|\Delta B(:,j)\|_2 \leq \tilde{\gamma}_{mn}, \tag{19.12}$$

so that $\|\Delta B\|_2/\|B\|_2 = O(u)$.

Note that the matrix Q in Theorem 19.4 is not computed by the QR factorization algorithm and is of purely theoretical interest. It is the fact that Q is exactly orthogonal that makes the result so useful. When Q is explicitly formed, two questions arise:

1. How close is the computed $\widehat{Q}$ to being orthonormal?
2. How large is $A - \widehat{Q}\widehat{R}$?

Both questions are easily answered using the analysis above.

We suppose that $Q = P_1P_2\dots P_n$ is evaluated in the more efficient right to left order. Lemma 19.3 gives (with $A_1 = I_m$)

$$\widehat{Q} = Q(I_m + \Delta I), \qquad \|\Delta I(:,j)\|_2 \leq \tilde{\gamma}_{mn}, \quad j = 1\colon n.$$

Hence

$$\|\widehat{Q} - Q\|_F \leq \sqrt{n}\,\tilde{\gamma}_{mn}, \tag{19.13}$$

showing that $\widehat{Q}$ is very close to an orthonormal matrix. Moreover, using Theorem 19.4,

$$\begin{aligned}\|(A-\widehat{Q}\widehat{R})(:,j)\|_2 &= \|(A-Q\widehat{R})(:,j)+((Q-\widehat{Q})\widehat{R})(:,j)\|_2\\ &\le \tilde{\gamma}_{mn}\|a_j\|_2+\|Q-\widehat{Q}\|_F\|\widehat{R}(:,j)\|_2\\ &\le \sqrt{n}\,\tilde{\gamma}_{mn}\|a_j\|_2.\end{aligned}$$

Thus if Q is replaced by $\widehat{Q}$ in Theorem 19.4, so that $A+\Delta A=\widehat{Q}\widehat{R}$, then the backward error bound remains true with an appropriate increase in the constant.

Finally, we consider use of the QR factorization to solve a linear system. Given a QR factorization of a nonsingular matrix $A\in\mathbb{R}^{n\times n}$, a linear system $Ax=b$ can be solved by forming Q^Tb and then solving $Rx=Q^Tb$. From Theorem 19.4, the computed $\widehat{R}$ is guaranteed to be nonsingular if $\kappa_2(A)n^{1/2}\tilde{\gamma}_{mn}<1$.

Theorem 19.5. *Let $A\in\mathbb{R}^{n\times n}$ be nonsingular. Suppose we solve the system $Ax=b$ with the aid of a QR factorization computed by the Householder algorithm. The computed $\widehat{x}$ satisfies*

$$(A+\Delta A)\widehat{x}=b+\Delta b,$$

where

$$\|\Delta a_j\|_2\le\tilde{\gamma}_{n^2}\|a_j\|_2,\quad j=1\colon n,\qquad \|\Delta b\|_2\le\tilde{\gamma}_{n^2}\|b\|_2.$$

Proof. By Theorem 19.4, the computed upper triangular factor $\widehat{R}$ satisfies $A+\Delta A=Q\widehat{R}$ with $\|\Delta a_j\|_2\le\tilde{\gamma}_{n^2}\|a_j\|_2$. By Lemma 19.3, the computed transformed right-hand side satisfies $\widehat{c}=Q^T(b+\Delta b)$, with $\|\Delta b\|_2\le\tilde{\gamma}_{n^2}\|b\|_2$. Importantly, the same orthogonal matrix Q appears in the equations involving $\widehat{R}$ and $\widehat{c}$.

By Theorem 8.5, the computed solution $\widehat{x}$ to the triangular system $\widehat{R}x=\widehat{c}$ satisfies

$$(\widehat{R}+\Delta R)\widehat{x}=\widehat{c},\qquad |\Delta R|\le\gamma_n|\widehat{R}|.$$

Premultiplying by Q yields

$$(A+\Delta A+Q\Delta R)\widehat{x}=b+\Delta b,$$

that is, $(A+\overline{\Delta A})\widehat{x}=b+\Delta b$, where $\overline{\Delta A}=\Delta A+Q\Delta R$. Using $\widehat{R}=Q^T(A+\Delta A)$ we have

$$\begin{aligned}\|\overline{\Delta a_j}\|_2 &\le \|\Delta a_j\|_2+\gamma_n\|\widehat{r}_j\|_2\\ &= \|\Delta a_j\|_2+\gamma_n\|a_j+\Delta a_j\|_2\\ &\le \tilde{\gamma}_{n^2}\|a_j\|_2.\qquad\square\end{aligned}$$

The proof of Theorem 19.5 naturally leads to a result in which b is perturbed. However, we can easily modify the proof so that only A is perturbed: the trick is to use the last part of Lemma 19.3 to write $\widehat{c}=(Q+\Delta Q)^Tb$, where $\|\Delta Q\|_F\le\tilde{\gamma}_{n^2}$, and to premultiply by $(Q+\Delta Q)^{-T}$ instead of Q in the middle of the proof. This leads to the result

$$(A+\Delta A)\widehat{x}=b,\qquad \|\Delta a_j\|_2\le\tilde{\gamma}_{n^2}\|a_j\|_2,\quad j=1\colon n. \tag{19.14}$$

An interesting application of Theorem 19.5 is to iterative refinement, as explained in §19.7.