On the Choice of Sign Defining Householder Transformations

Michael L. Overton[*]   Pinze Yu[†]

October 7, 2023

Abstract

It is well known that, when defining Householder transformations, the correct choice of sign in the standard formula is important to avoid cancellation and hence numerical instability. In this note we point out that when the "wrong" choice of sign is used, the extent of the resulting instability depends in a somewhat subtle way on the data leading to cancellation.

AMS Subject Classification: 65F05

1   Introduction

The QR factorization is a standard tool in numerical linear algebra, and Householder transformations provide the best general method to compute it. Following [Hig02, Sec. 19.1], a Householder transformation (or Householder reflector) has the form

$$P = I - \frac{2}{v^T v} v v^T, \tag{1}$$

where I is the identity matrix and v is a nonzero vector. It is easily verified that P is an orthogonal matrix, i.e., $P^T P = I$. The first step in the Householder reduction of an $m \times n$ matrix A, with $m \geq n$, to triangular form is to define a Householder transformation P_1 that maps x, the first column of A, to a multiple of the first coordinate vector $e_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^m$. Since $P_1 x$ must have the same Euclidean length as x, we require $P_1 x = \sigma \|x\| e_1$, where $\sigma = \pm 1$ and $\| \cdot \|$ denotes the 2-norm. Thus we need

$$P_1 x = x - \frac{2 v^T x}{v^T v} v = \sigma \|x\| e_1$$

which implies that v is a scalar multiple of $x - \sigma \|x\| e_1$, and since P_1 is independent of $\|v\|$, without loss of generality we can choose

$$v = x - \sigma \|x\| e_1. \tag{2}$$

---

[*]Courant Institute of Mathematical Sciences, New York University, `mo1@nyu.edu`
[†]Courant Institute of Mathematical Sciences, New York University, `py2050@nyu.edu`

To avoid numerical cancellation in (2), it is generally recommended to use

$$\sigma = -\text{sgn}(x_1) \tag{3}$$

where x_1 is the first component of the vector x and sgn is the standard sign function, which for convenience we define to be $+1$ if its argument is zero. The transformation P_1 is then applied to the remaining columns of A as well, exploiting the formula (1) for efficiency, yielding the matrix $P_1 A$ whose first column has all zeros except in the first position. The factorization is completed by repeating the process for every column of A, working with only with the data in rows k through m and columns k through n at the kth step, yielding a total of n Householder transformations $P_1, P_2 \dots, P_n$, along with the upper triangular final matrix R. Then in exact arithmetic, $A = QR$, with $Q = P_1 P_2 \dots P_n$.

In this note we examine exactly what occurs if the "wrong" sign[1]

$$\sigma = \text{sgn}(x_1) \tag{4}$$

is used to compute v in (2).

2 Observation

We consider the following experiment. We would like to choose A so that using the wrong sign (4) results in as much cancellation as possible; an easy way to do this is to choose the first column to have much smaller entries, in magnitude, than the (1,1) entry, so that $\text{sgn}(x_1)\|x\|$ approximately cancels with x_1 in (2). Here, we report the results of an experiment computing Q and R using both choices of sign for a 3×2 matrix A with $a_{11} = 1$, $a_{21} = \delta$, $a_{32} = 0$ and the second column chosen randomly, for δ taking the successive values 10^{-p}, $p = 1, 2, \dots, 16$. The experiment was conducted using MATLAB on a MacBook Pro, for which the machine epsilon ϵ_{mch} (the gap between 1 and the next larger floating point number) is approximately 10^{-16} (as MATLAB uses IEEE double precision by default).

Figure 1 shows the computed 2-norm $\|A - QR\|$ for each choice of δ and for three algorithms: using the correct sign (blue circles), the wrong sign (red asterisks), and using MATLAB's built-in qr (cyan crosses); note the log-log scaling. Unsurprisingly, the results using the correct choice of sign or the built-in qr are, for all δ, approximately ϵ_{mch}. Surprisingly, however, the results using the wrong sign appear in an inverted-V pattern with respect to δ. This is somewhat reminiscent of the well-known V pattern that is often used, for example in [Ove01, Chap. 11], to show how the truncation error and rounding error respectively dominate the error in the approximation of a derivative of a function f at a point x by a finite difference quotient $\frac{f(x+h)-f(x)}{h}$, the former dominant for large h and the latter dominant for small h. The comparison even extends

---

[1] It is pointed out in [Hig02, Sec. 19.1] that the sign (4) may be used if the formula for v is rearranged; see [Par71] for details. While this is useful if consistent signs are preferred in computing the QR factorization, it is not relevant to the subsequent discussion.
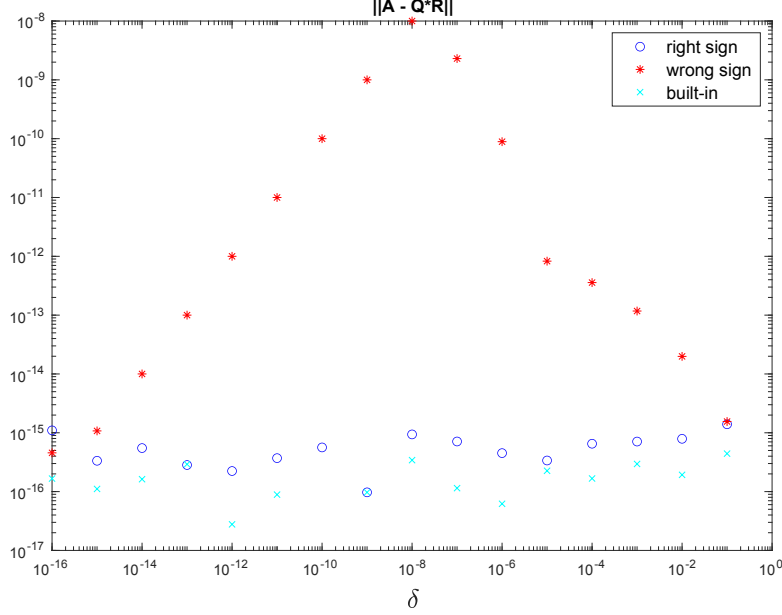
Figure 1: The 2-norm of $A - QR$, where Q and R are the computed Q and R factors of a 3×2 matrix A with first column $[1, \ \delta, \ 0]^T$, using Householder reduction with the correct choice of sign (3) (blue circles), the wrong choice of sign (4) (red asterisks) and MATLAB's built-in `qr` function (cyan crosses), all plotted as a function of δ.

to noting that the right side of the inverted V is ragged, indicating dominance by rounding error, while the left side is a straight line, indicating purely linear dependence; in the finite-difference example, the roles of left and right are reversed. Note that the choice of $\delta \approx \epsilon_{\mathrm{mch}}^{1/2}$, the square root of the machine precision, gives the most inaccurate result, while in the finite difference example, it is well known that $h \approx \epsilon_{\mathrm{mch}}^{1/2}$ is the best choice, assuming appropriately scaled data. The results shown in Figure 1 are essentially unchanged if much larger matrices are used.

3 Explanation

The right side of the inverted V, where the error increases as δ decreases, is what we expected as the cancellation error in (2) becomes more dominant. But what about the left side, where the error decreases as δ continues to decrease? In fact, this is easily explained. In the experiment, the first column of A is $[1, \ \delta, \ 0]^T$, whose 2-norm is $\sqrt{1 + \delta^2}$, so for δ somewhat less than $\epsilon_{\mathrm{mch}}^{1/2}$, the computed 2-

norm is precisely 1. This results in the first component of the vector v defining the first Householder transformation being zero. The second component of v is δ and the third is zero, so the normalized vector $v/\|v\|$ is the second unit vector. This means that the first Householder transformation is the identity except with -1 instead of $+1$ in the (2,2) position. Thus the first column of Q, the product of all (in this case two) Householder transformations, is the first unit vector. Since the computed matrix R is upper triangular, this means the first column of the computed product QR is $[1,\ 0,\ 0]^T$. Thus, the norm of the first column of $A - QR$ is exactly δ. There is no reason for $\|A - QR\|$ to be more than δ, so the result is that the error $\|A - QR\|$ decreases linearly as δ drops below $\epsilon_{\mathrm{mch}}^{1/2}$; although cancellation occurs, the result is to give an increasingly accurate answer as δ is reduced. An interesting consequence is that the cancellation apparently cannot result in arbitrarily poor results; the example illustrated here suggests that, for A with norm one, $\|A - QR\|$ will perhaps never be significantly greater than $\epsilon_{\mathrm{mch}}^{1/2}$ when the wrong sign is used, compared to ϵ_{mch} when the correct sign is used (a standard result in numerical linear algebra, e.g.[Hig02, Theorem 19.4], [TB97, Theorem 16.1]).

4   History

According to both Higham [Hig02] and Stewart [Ste98], the first known use of Householder transformations was by Turnbull and Aitken in 1932. Stewart writes "Householder, who discovered the transformations independently [in 1958], was the first to realize their computational significance." Stewart also writes "Householder seems to have missed the fact that there are two transformations that will reduce a vector to a multiple of [the first unit vector] and that the natural construction of one of them is unstable. This oversight was corrected by Wilkinson [in 1960]." In Householder's 1964 book [Hou75] he writes "a singularity would arise with one choice of sign" (when the two terms cancel exactly) and hence he recommends the other choice of sign, but, rather surprisingly, he does not mention possible cancellation. Virtually all later books on numerical linear algebra focus on the latter issue, motivating the choice (3), but we are not aware of any discussion of the "inverted V" phenomenon discussed here. Nor is there any hint that the error $\|A - QR\|$ may be bounded by approximately $\epsilon_{\mathrm{mch}}^{1/2}$ when A has norm one and the wrong sign is used. Of course, we are not arguing that using the wrong sign is acceptable. There is no reason to do so, and indeed, even if the worst case error is bounded by $\epsilon_{\mathrm{mch}}^{1/2}$, this is still unacceptable when using the correct sign results in a perfectly stable algorithm.

References

[Hig02] Nicholas J. Higham. *Accuracy and stability of numerical algorithms.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2002.

[Hou75] Alston S. Householder. *The theory of matrices in numerical analysis.* Dover Publications, Inc., New York, 1975. Reprint of 1964 edition.

[Ove01] Michael L. Overton. *Numerical computing with IEEE floating point arithmetic.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[Par71] B. N. Parlett. Analysis of algorithms for reflections in bisectors. *SIAM Rev.*, 13:197–208, 1971.

[Ste98] G. W. Stewart. *Matrix algorithms. Vol. I.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998. Basic decompositions.

[TB97] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.