The University of Manchester

Probabilistic Rounding Error Analysis of Householder QR Factorization

Connolly, Michael P. and Higham, Nicholas J.

2022

MIMS EPrint: **2022.5**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

PROBABILISTIC ROUNDING ERROR ANALYSIS OF HOUSEHOLDER QR FACTORIZATION*

MICHAEL P. CONNOLLY† AND NICHOLAS J. HIGHAM†

Abstract. The standard worst-case normwise backward error bound for Householder QR factorization of an $m \times n$ matrix is proportional to mnu, where u is the unit roundoff. We prove that the bound can be replaced by one proportional to $\sqrt{mnu}$ that holds with high probability if the rounding errors are mean independent and of mean zero and if the normwise backward errors in applying a sequence of $m \times m$ Householder matrices to a vector satisfy bounds proportional to $\sqrt{mu}$ with probability 1. The proof makes use of a matrix concentration inequality. The same square rooting of the error constant applies to two-sided transformations by Householder matrices and hence to standard QR-type algorithms for computing eigenvalues and singular values. It also applies to Givens QR factorization. These results complement recent probabilistic rounding error analysis results for inner-product based algorithms and show that the square rooting effect is widespread in numerical linear algebra. Our numerical experiments, which make use of a new backward error formula for QR factorization, show that the probabilistic bounds give a much better indicator of the actual backward errors and their rate of growth than the worst-case bounds.

Key words. floating-point arithmetic, backward error analysis, backward error, probabilistic rounding error analysis, Givens QR factorization, Householder QR factorization, matrix concentration inequality

AMS subject classifications. 65G50, 65F05

1. Introduction.

It is well known that backward error bounds from worst-case rounding error analyses can greatly overestimate the backward error. Recently, it was proved that for inner product-based algorithms with backward error bounds of the form $f(n)u$, where n is the problem dimension and u is the unit roundoff, a bound proportional to $\sqrt{f(n)}u$ holds with high probability under suitable assumptions on the rounding errors [15]. These results apply to matrix multiplication, Cholesky factorization, LU factorization, and the solution of triangular systems, but not to algorithms based on orthogonal transformations.

A QR factorization of $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is a factorization $A = QR$ where $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{m\times n}$ is upper trapezoidal. A Householder matrix is a matrix of the form

$$P = I - \frac{2}{v^T v} vv^T, \quad 0 \neq v \in \mathbb{R}^m.$$

The vector v can be chosen so that in the product $y = Px$ all elements of y except the first are zero. By applying a sequence of such Householder matrices to A we can reduce it to upper trapezoidal form: $P_n \dots P_2 P_1 A = R$. We then have $Q = (P_n \dots P_2 P_1)^T$.

The standard rounding error analysis for Householder QR factorization is summarized in the following result [12, Thm. 19.4]. We define

$$\gamma_n = \frac{nu}{1-nu} \tag{1.1}$$

*Version of April 13, 2023.

Funding: This work was supported by the Royal Society, Engineering and Physical Sciences Research Council grant EP/P020720/1, and the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (michael.connolly-3@manchester.ac.uk, nick.higham@manchester.ac.uk).

and use the 2-norm $\|x\|_2 = (x^T x)^{1/2}$, the corresponding subordinate matrix norm, and the Frobenius norm $\|A\|_F = \mathrm{trace}(A^T A)^{1/2}$. We denote the jth column of A by a_j. Throughout this work we express our bounds in terms of c_0, c_1, c_2, ..., which denote integer constants of modest size.

THEOREM 1.1. *Let $\widehat{R} \in \mathbb{R}^{m\times n}$ be the computed upper trapezoidal QR factor of $A \in \mathbb{R}^{m\times n}$ $(m \ge n)$ obtained via the Householder QR algorithm, and assume that a condition of the form $c_0\gamma_{mn} < 1$ holds. There exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

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

where

$$\|\Delta a_j\|_2 \le c_1\gamma_{mn}\|a_j\|_2, \quad j = 1\!:\!n. \tag{1.2}$$

The columnwise bounds (1.2) yield the normwise backward error bound

$$\|\Delta A\|_F \le c_1\gamma_{mn}\|A\|_F. \tag{1.3}$$

The purpose of this work is to show that the worst-case backward error bounds (1.2) and (1.3) for Householder QR factorization can be replaced by probabilistic bounds in which the dimension-dependent constants are proportional to the square roots of the original constants under suitable assumptions on the rounding errors; essentially, we can replace $c_1\gamma_{mn}$ by $c_1'\gamma_{\sqrt{mn}} + O(u^2)$. This is not an immediate consequence of the analysis for inner product-based algorithms because the vector addition and scaling involved in applying Householder matrices need a different treatment.

We use the standard model of floating-point arithmetic, which assumes that the elementary operations satisfy

$$\mathrm{fl}(x \mathrm{\ op\ } y) = (x \mathrm{\ op\ } y)(1+\delta), \quad |\delta| \le u. \tag{1.4}$$

In a rounding error analysis products of $1+\delta$ terms arise, and their distance from 1 can be bounded using the following lemma [12, Lem 3.1].

LEMMA 1.2. *If $|\delta_i| \le u$ and $\rho_i = \pm 1$ for $i = 1\!:\!n$, and $nu < 1$, then*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \theta_n, \quad |\theta_n| \le \gamma_n. \tag{1.5}$$

An analogous probabilistic result [7, Thm. 4.6] is given in terms of the constant

$$\widetilde{\gamma}_n(\lambda) = \exp\left(\frac{\lambda\sqrt{n}u + nu^2}{1-u}\right) - 1 = \lambda\sqrt{n}u + O(u^2). \tag{1.6}$$

with $\lambda > 0$. Before stating the result, we need a definition.

DEFINITION 1.3. *The random variables $\delta_1, \delta_2, \ldots, \delta_n$ are* mean independent *if $\mathbb{E}(\delta_k \mid \delta_{k-1}, \ldots, \delta_1) = \mathbb{E}(\delta_k)$ for $k = 2\!:\!n$, where $\mathbb{E}$ denotes expectation.*

LEMMA 1.4. *Let $\delta_1, \delta_2, \ldots, \delta_n$ be mean independent random variables of mean zero such that $|\delta_i| \le u$ for all i, and let $\rho_i = \pm 1$, $i = 1\!:\!n$. Then for any constant $\lambda > 0$,*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda), \tag{1.7}$$

holds with probability at least $p(\lambda) = 1 - 2\exp(-\lambda^2/2)$.

Throughout this work we will use the following model of rounding errors, which was also used in [7] and [16].

MODEL 1.5 (probabilistic model of rounding errors). *Let the computation of interest generate rounding errors* $\delta_1, \delta_2, \ldots$ *(in that order) satisfying* (1.4). *The* δ_k *are mean independent random variables of mean zero.*

For the round to nearest rounding mode (the default in IEEE arithmetic) the error bounds obtained under Model 1.5 often reflect the behavior of the actual errors obtained, but in some circumstances the model is not valid because the rounding errors are highly dependent or have nonzero mean, and the bound can then be violated [7, sec. 7], [8, sec. 8], [15, sec. 4.2]. For stochastic rounding [7], [8], the rounding errors *always* satisfy the model. For further discussion of the applicability of the model see [7], [15].

In our probabilistic rounding error analysis it is more difficult than in worst-case rounding error analysis to bound terms of all orders in u so we will state the bounds to first order in u. We will, however, show that the higher order terms do not change the conclusions of the analysis. We also give explicit bounds for the probabilities with which the error bounds hold, and these show that the analysis does not introduce exponential worsening of probabilities with the dimensions. All our probabilistic results can informally be stated as saying that the given backward error bound holds with λ a modest constant with high probability.

We begin in section 2 with a probabilistic analysis of the rounding errors in constructing a Householder vector. In section 3 we state a matrix concentration inequality that we use in section 4 in our probabilistic rounding error analysis of the application of products of Householder matrices to a matrix and of Householder QR factorization. In section 5 we discuss the implications of the analysis for two-sided transformations, Givens QR factorization, and other approaches. In section 6 we give a new backward error formula for QR factorization that we use in section 7, where we carry out numerical experiments to compare the results of the analysis with the practical behavior. Conclusions are given in section 8.

2. Construction of a Householder vector. We first give a probabilistic error result for the construction of the Householder vectors needed in QR factorization. We begin by considering the evaluation of the vector 2-norm. We denote by $\mathrm{fl}(\mathit{expr})$ the computed result of evaluating the expression *expr*. Note that Model 1.5 assumes that (1.4) holds, so we are at liberty to use worst-case bounds within the probabilistic rounding error analysis, and we will do so for certain scalar operations.

LEMMA 2.1. *Under Model 1.5, the computed 2-norm of* $x \in \mathbb{R}^n$ *satisfies*

$$\mathrm{fl}(\|x\|_2) = \|x\|_2(1+\eta), \quad |\eta| \le c_1\widetilde{\gamma}_n(\lambda), \tag{2.1}$$

with probability at least $p_1(\lambda, n) = 1 - 2n\exp(-\lambda^2/2)$.

Proof. By standard inner-product analysis [7, Thm. 4.8], [15, Thm. 3.1],

$$\mathrm{fl}(x^Tx) = x^Tx(1+\theta_n), \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda), \tag{2.2}$$

holds with probability at least $p_1(\lambda, n)$. Then $\mathrm{fl}(\|x\|_2) = \|x\|_2\sqrt{1+\theta_n}(1+\delta)$, with $|\delta| \le u$. For $\sqrt{1+\theta_n} = 1+\alpha$, we certainly have $|\alpha| \le |\theta_n|$. Then $(1+\alpha)(1+\delta) = 1+\beta$, with $|\beta| \le |\theta_n| + u + |\theta_n|u$. Hence provided $|\theta_n| \le \widetilde{\gamma}_n(\lambda)$, we have $|\beta| \le \widetilde{\gamma}_n(\lambda) + u + \widetilde{\gamma}_n(\lambda)u \le c_1\widetilde{\gamma}_n(\lambda)$. The bound holds with probability at least the probability of (2.2) holding. ∎

We now derive a probabilistic version of the worst-case error bound for construction of a Householder vector [12, Lem. 19.1]. Here, $\mathrm{sign}(t) = -1$ if $t < 0$ or 1 if $t \geq 0$.

LEMMA 2.2. *Let $x \in \mathbb{R}^n$. Consider the construction of $\beta \in \mathbb{R}$ and $v \in \mathbb{R}^n$ such that $Px = -\mathrm{sign}(x_1)\|x\|_2 e_1$, where $P = I - \beta vv^T$ is a Householder matrix and $\beta = 2/(v^T v)$, by*

$$\begin{aligned} v(2\!:\!n) &= x(2\!:\!n), \\ s &= \mathrm{sign}(x_1)\|x\|_2, \\ v_1 &= x_1 + s, \\ \beta &= 1/(sv_1). \end{aligned}$$

Under Model 1.5 the computed $\widehat{\beta}$ and $\widehat{v}$ satisfy $\widehat{v}(2:n) = v(2:n)$ and

$$\widehat{v}_1 = v_1(1+\eta_1), \qquad \widehat{\beta} = \beta(1+\eta_2), \tag{2.3}$$

where $|\eta_i| \leq c_3\widetilde{\gamma}_n(\lambda) + O(u^2)$ holds for $i = 1\!:\!2$ with probability at least $p_1(\lambda, n)$.

Proof. From Lemma 2.1 we have that $\widehat{s} = s(1+\Delta s)$, where $|\Delta s| \leq c_1\widetilde{\gamma}_n(\lambda)$ holds with probability at least $p_1(\lambda, n)$. The rest of the computations, which determine v_1 and β, are scalar operations and so we use worst-case bounds for these. We have

$$\begin{aligned} \widehat{v}_1 &= (x_1 + \widehat{s})(1+\delta), \quad |\delta| \leq u \\ &= x_1 + s + \delta(x_1 + s) + s\Delta s(1+\delta) =: v_1 + \Delta v_1. \end{aligned}$$

Because x_1 and s have the same sign, $|s| \leq |x_1 + s|$, and so

$$\begin{aligned} |\Delta v_1| &= \left| (x_1 + s)\Big(\delta + \frac{s}{x_1 + s}\Delta s(1+\delta)\Big) \right| \\ &\leq |v_1|\big(u + c_1\widetilde{\gamma}_n(\lambda) + c_1\widetilde{\gamma}_n(\lambda)u\big) \leq (c_2\widetilde{\gamma}_n(\lambda) + O(u^2))|v_1|. \end{aligned}$$

So $\widehat{v}_1 = v_1(1+\eta_1)$, where $|\eta_1| \leq c_2\widetilde{\gamma}_n(\lambda) + O(u^2)$ holds with probability at least $p_1(\lambda, n)$. If this inequality holds we then have, using (1.4) and (1.5),

$$\widehat{\beta} = \mathrm{fl}(1/(\widehat{s}\,\widehat{v}_1)) = \frac{1+\theta_2}{s(1+\Delta s)v_1(1+\eta_1)} = \frac{1}{sv_1}(1+\eta_2),$$

where $|\eta_2| \leq c_3\widetilde{\gamma}_n(\lambda) + O(u^2)$, with $c_3 \geq c_2$. ∎

If we redefine the Householder matrix as $P = I - vv^T$ with $\|v\|_2 = \sqrt{2}$ then Lemma 2.2 amounts to the result

$$\widehat{v} = v + \Delta v, \quad |\Delta v| \leq \big(c_4\widetilde{\gamma}_n(\lambda) + O(u^2)\big)|v|, \tag{2.4}$$

where the bound holds with probability at least $p_1(\lambda, n)$.

3. Matrix concentration inequalities. The proof of Lemma 1.4 in [7] makes use of the scalar Azuma–Hoeffding inequality [17, Thm. 13.4]. Here, we will use a matrix version of that result due to Tropp [18, Thm. 7.1]. Our notation is as follows: we write (in this section only), $X \leq Y$, where X and Y are symmetric matrices, to mean that $Y - X$ is positive semidefinite; the expectation of a matrix is defined componentwise; and we denote by $\lambda_{\max}$ the largest eigenvalue of a symmetric matrix. Note that "with probability 1" is often expressed as "almost surely".

THEOREM 3.1. *Let $X_1, \dots, X_n$ be a sequence of random symmetric $d \times d$ matrices. If $\mathbb{E}(X_k \mid X_{k-1}, \dots, X_1) = 0$ and $X_k^2 \le A_k^2$ with probability 1 for all k, where $A_1, \dots, A_n$ is a fixed sequence of symmetric $d \times d$ matrices, then for all $t \ge 0$,*

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum_{k=1}^n X_k \right) \ge t \right\} \le d \exp\bigl(-t^2/(8\sigma^2)\bigr), \tag{3.1}$$

where

$$\sigma^2 = \left\| \sum_{k=1}^n A_k^2 \right\|_2. \tag{3.2}$$

We need a version of Theorem 3.1 for nonsymmetric matrices. Define

$$\phi(B) = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \tag{3.3}$$

to be the symmetric dilation of the rectangular matrix B. It is well known that

$$\lambda_{\max}(\phi(B)) = \|B\|_2. \tag{3.4}$$

The required result can be obtained by applying the theorem to the symmetric dilation of the sequence $X_1, \dots, X_n$ and using (3.4) [18, Rem. 7.3]. We also rescale by defining $\lambda = t/(2\sqrt{2}\sigma)$.

THEOREM 3.2. *Let $X_1, \dots, X_n$ be a sequence of random $d_1 \times d_2$ matrices. If $\mathbb{E}(X_k \mid X_{k-1}, \dots, X_1) = 0$ for all k, and $X_k X_k^T \le A_k A_k^T$ and $X_k^T X_k \le A_k^T A_k$ with probability 1 for all k, where $A_1, \dots, A_n$ is a fixed sequence of $d_1 \times d_2$ matrices, then for all $\lambda \ge 0$,*

$$\mathbb{P}\left\{ \left\| \sum_{k=1}^n X_k \right\|_2 \ge 2\sqrt{2}\sigma\lambda \right\} \le (d_1 + d_2) \exp(-\lambda^2), \tag{3.5}$$

where

$$\sigma^2 = \max\left(\left\| \sum_{k=1}^n A_k A_k^T \right\|_2, \left\| \sum_{k=1}^n A_k^T A_k \right\|_2 \right) \le \sum_{k=1}^n \|A_k\|_2^2. \tag{3.6}$$

This is the key result that we need in the next section.

4. Application of a sequence of Householder matrices. Our probabilistic rounding error analysis for Householder QR factorization follows a similar strategy as for the worst-case analysis in [12, sec. 19.3]: we analyze the application of a sequence of general Householder matrices to a vector, then the application of the same sequence to a matrix, and finally specialize to the Householder matrices used in QR factorization.

For a single application of an $m \times m$ Householder matrix our backward error analysis gives an error constant proportional to $m^{1/2}$. A straightforward inductive argument for the application of n Householder matrices would lead to a backward error bound with a constant proportional to $nm^{1/2}$, which is unsatisfactory as it is linear in n, as for the worst-case analysis. Our key observation is that by applying the matrix concentration inequality in Theorem 3.2 we can obtain a bound with a constant of order $n^{1/2}m^{1/2}$.

In the first two lemmas we consider the application of the Householder matrices to a vector b. In Lemma 4.3 and Theorem 4.4, b is taken to be each of the columns of A in turn. In the first three lemmas we take the Householder vectors to be exact.

LEMMA 4.1. *Consider the (given) Householder matrices*

$$P_i = I - v_i v_i^T \in \mathbb{R}^{m\times m}, \quad v_i^T v_i = 2, \quad i = 1\!:\!r,$$

and the product

$$b_{r+1} = P_r \dots P_2 P_1 b, \quad b = b_1 \in \mathbb{R}^m,$$

computed as $b_{j+1} = P_j b_j = b_j - (v_j^T b_j) v_j$, $j = 1\!:\!r$. *Under Model 1.5, the computed* $\widehat{b}_{r+1}$ *satisfies*

$$\widehat{b}_{r+1} = (P_r + \Delta P_r)\dots(P_1 + \Delta P_1) b, \tag{4.1}$$

where

$$\|\Delta P_j\|_2 \le c_5 \widetilde{\gamma}_m(\lambda) \tag{4.2}$$

holds for all j *with probability at least* $p_2(\lambda, m, r) = 1 - 2rm\exp(-\lambda^2/2)$.

Proof. Define the scalars $s_j = v_j^T b_j$. From standard rounding error analysis [12, sec. 3.1], [15, Proof of Thm. 3.1], the computed s_j satisfy $\widehat{s}_j = v_j^T D_j \widehat{b}_j$, where

$$D_j = \operatorname{diag}\big(d_k^{(j)}\big), \quad d_k^{(j)} = \prod_{i=1}^{m}\big(1 + \delta_{k,i}^{(j)}\big), \quad \big|\delta_{k,i}^{(j)}\big| \le u, \quad i,k = 1\!:\!m.$$

Then we form

$$\widehat{b}_{j+1} = \mathrm{fl}(\widehat{b}_j - \widehat{s}_j v_j) = \big(I + \Lambda_1^{(j)}\big)\big(\widehat{b}_j - \big(I + \Lambda_2^{(j)}\big)\widehat{s}_j v_j\big),$$

where

$$\Lambda_k^{(j)} = \operatorname{diag}\big(\epsilon_{k,i}^{(j)}\big), \quad \big|\epsilon_{k,i}^{(j)}\big| \le u, \quad k = 1, 2.$$

Note that $\Lambda_1^{(j)}$ depends on $\Lambda_2^{(j)}$, and that since we are using Model 1.5 the $\delta_{k,i}^{(j)}$ and $\epsilon_{k,i}^{(j)}$ are mean independent random variables of mean zero. Hence

$$\begin{aligned}
\widehat{b}_{j+1} &= \big(I + \Lambda_1^{(j)}\big)\big(\widehat{b}_j - \big(I + \Lambda_2^{(j)}\big) v_j^T D_j \widehat{b}_j \cdot v_j\big) \\
&= \big(I + \Lambda_1^{(j)}\big)\big(I - \big(I + \Lambda_2^{(j)}\big) v_j v_j^T D_j\big)\widehat{b}_j \\
&= \Big[I - \big(I + \Lambda_2^{(j)}\big)\big(v_j v_j^T + v_j v_j^T (D_j - I)\big) + \Lambda_1^{(j)}\big(I - \big(I + \Lambda_2^{(j)}\big) v_j v_j^T D_j\big)\Big]\widehat{b}_j \\
&= (P_j + \Delta P_j)\widehat{b}_j, \qquad (4.3)
\end{aligned}$$

where

$$\Delta P_j = -\Lambda_2^{(j)} v_j v_j^T - \underbrace{\big(I + \Lambda_2^{(j)}\big) v_j v_j^T (D_j - I)}_{(*)} + \Lambda_1^{(j)}\big(I - \big(I + \Lambda_2^{(j)}\big) v_j v_j^T D_j\big), \tag{4.4}$$

and so

$$\begin{aligned}
\|\Delta P_j\|_2 &\le 2u + 2(1+u)\|D_j - I\|_2 + u\big(1 + 2(1+u)\|D_j\|_2\big) \\
&= 3u + 2(1+u)\|D_j - I\|_2 + 2u(1+u)\|D_j\|_2.
\end{aligned}$$

From Lemma 1.4 we have that for any constant $\lambda > 0$ and any particular j,

$$d_k^{(j)} = 1 + \psi_j^{(k)}, \quad |\psi_j^{(k)}| \le \widetilde{\gamma}_m(\lambda), \tag{4.5}$$

holds for any particular k and j with probability at least $p(\lambda) = 1 - 2\exp(-\lambda^2/2)$, or equivalently it fails to hold with probability at most $1 - p(\lambda)$. By the inclusion–exclusion principle [21, p. 39] (4.5) fails to hold for at least one of the pairs (j,k) with probability at most $rm(1-p(\lambda))$, which means that it holds for all j and k with probability at least $p_2(\lambda, m, r) = 1 - rm(1-p(\lambda)) = 1 - 2rm\exp(-\lambda^2/2)$.

Hence $\|I - D_j\|_2 \le \widetilde{\gamma}_m(\lambda)$ holds for all j with probability at least $p_2(\lambda, m, r)$ and hence

$$\|\Delta P_j\|_2 \le 3u + 2(1+u)\widetilde{\gamma}_m(\lambda) + 2u(1+u)(1+\widetilde{\gamma}_m(\lambda)) \le c_5\widetilde{\gamma}_m(\lambda) \tag{4.6}$$

holds for all j with probability at least $p_2(\lambda, m, r)$. ∎

The remaining results in this section assume that the bound (4.2) holds with probability 1. We discuss the reason for this assumption after Theorem 4.4.

LEMMA 4.2. *With the same notation as in Lemma 4.1, assume that the bound (4.2) holds with probability 1 for all j. Then the computed $\widehat{b}_{r+1}$ satisfies*

$$\widehat{b}_{r+1} = Q^T(b + \Delta b), \quad \|\Delta b\|_2 \le c_6\lambda\sqrt{r}\widetilde{\gamma}_m(\lambda)\|b\|_2 + O(u^2) \tag{4.7}$$

with probability at least $p_3(\lambda, m) = 1 - 2m\exp(-\lambda^2)$, where $Q = (P_r \dots P_2 P_1)^T$.

Proof. We can rewrite (4.1) as

$$\begin{aligned}\widehat{b}_{r+1} &= b_{r+1} + \left(\sum_{j=1}^{r} P_r \dots P_{j+1}\Delta P_j P_{j-1}\dots P_1 + O(u^2)\right) b \\ &= b_{r+1} + Q^T\left(\sum_{j=1}^{r} F_j + O(u^2)\right) b, \end{aligned} \tag{4.8}$$

where

$$F_j = P_1 \dots P_j \Delta P_j P_{j-1} \dots P_1. \tag{4.9}$$

Our aim is to show that $\mathbb{E}(F_j \mid F_{j-1}, \dots, F_1) = 0$, so that we can apply Theorem 3.2 with $X_k = F_k$. When we substitute the expressions for F_j and then ΔP_j into this expression and use the linearity of the expectation we obtain a sum of expectations, each of which we need to show is zero. We will just consider two of the terms, which come from $(*)$ in (4.4), as the other terms are treated similarly.

The matrix D_j contains the rounding errors from the computation of $v_j^T b_j$, whereas all the other terms in F_j contain rounding errors from later computations.

For the first part of the term $(*)$ in (4.4), namely $v_j v_j^T(D_j - I)$, we have

$$\begin{aligned}&\mathbb{E}(P_1 \dots P_j v_j v_j^T (I - D_j) P_{j-1} \dots P_1 \mid F_{j-1}, \dots, F_1) \\ &\qquad = P_1 \dots P_j v_j v_j^T \mathbb{E}(I - D_j \mid F_{j-1}, \dots, F_1) P_{j-1} \dots P_1,\end{aligned} \tag{4.10}$$

since the v_j and P_j are constant. We need the general form of the law of total expectation (LTE), $\mathbb{E}(\mathcal{X} \mid \mathcal{Y}) = \mathbb{E}(\mathbb{E}(\mathcal{X} \mid \mathcal{Z}) \mid \mathcal{Y})$, for any pair of measurable sets $\mathcal{Y}$ and $\mathcal{Z}$ such that $\mathcal{Y} \subseteq \mathcal{Z}$ [2, Thm. 34.4]. We define the set

$$\mathcal{Z}_s = \left\{ \left\{ \delta_{k,i}^{(\ell)} : i,k = 1:m,\ \ell = 1:s \right\}, F_{j-1}, \dots, F_1 \right\}. \tag{4.11}$$

By the LTE

$$\mathbb{E}(D_j \mid F_{j-1}, \dots, F_1) = \mathbb{E}\big(\mathbb{E}(D_j \mid \mathcal{Z}_{j-1}) \mid F_{j-1}, \dots, F_1\big) = I,$$

by [7, Lem. 6.4], in view of Model 1.5. Hence the expectation (4.10) is zero.

For the second part of the term $(*)$ in (4.4), namely $\Lambda_2^{(j)} v_j v_j^T (D_j - I)$, we have

$$\begin{aligned}\mathbb{E}(P_1 \dots P_j \Lambda_2^{(k)} v_j v_j^T (I - D_j) P_{j-1} \dots P_1 \mid F_{j-1}, \dots, F_1) \\ = P_1 \dots P_j \mathbb{E}\Big(\Lambda_2^{(j)} v_j v_j^T (I - D_j) \mid F_{j-1}, \dots, F_1\Big) P_{j-1} \dots P_1.\end{aligned}$$

By the LTE,

$$\begin{aligned}\mathbb{E}\Big(\Lambda_2^{(j)} v_j v_j^T (I - D_j) \mid F_{j-1}, \dots, F_1\Big) &= \mathbb{E}\Big(\mathbb{E}\Big(\Lambda_2^{(j)} v_j v_j^T (I - D_j) \mid \mathcal{Z}_j\Big) \mid F_{j-1}, \dots, F_1\Big) \\ &= \mathbb{E}\Big(\mathbb{E}\Big(\Lambda_2^{(j)} \mid \mathcal{Z}_j\Big) v_j v_j^T (I - D_j) \mid F_{j-1}, \dots, F_1\Big) \\ &= 0,\end{aligned}$$

since $\mathbb{E}\big(\Lambda_2^{(j)} \mid \mathcal{Z}_j\big) = 0$ by Model 1.5.

We now bound $\|F_j\|_2$, so that we can apply Theorem 3.2. By (4.2), (4.9), and the assumption of the lemma,

$$\|F_j\|_2^2 = \|\Delta P_j\|_2^2 \le c_5^2 \widetilde{\gamma}_m(\lambda)^2, \quad j = 1 : r,$$

with probability 1. Hence we can take $X_j = F_j$, $A_j = c_5 \widetilde{\gamma}_m(\lambda) I_{d_1,d_2}$, and $\sigma^2 = r c_5^2 \widetilde{\gamma}_m(\lambda)^2$ in Theorem 3.2 and set $E = \sum_{j=1}^r F_j$, to obtain

$$\mathbb{P}\big(\|E\|_2 \ge 2\sqrt{2} c_5 \lambda \sqrt{r}\, \widetilde{\gamma}_m(\lambda)\big) \le 2m \exp(-\lambda^2), \tag{4.12}$$

or equivalently

$$\|E\|_2 \le c_6 \lambda \sqrt{r}\, \widetilde{\gamma}_m(\lambda), \tag{4.13}$$

holding with probability at least $1 - 2m\exp(-\lambda^2)$.

From (4.8) we have

$$\widehat{b}_{r+1} = Q^T b + Q^T E b + O(u^2) = Q^T (b + \Delta b)$$

with $\Delta b = Eb + O(u^2)$, and so

$$\|\Delta b\|_2 \le \|Eb\|_2 + O(u^2) \le c_6 \lambda \sqrt{r}\, \widetilde{\gamma}_m(\lambda) \|b\|_2 + O(u^2). \tag{4.14}$$

∎

Now we consider the backward error in applying a sequence of Householder matrices to a matrix.

LEMMA 4.3. *Consider the sequence of transformations*

$$A_{k+1} = P_k A_k, \quad k = 1 : r.$$

where $A_1 = A \in \mathbb{R}^{m \times n}$, each $P_k \in \mathbb{R}^{m \times m}$ is a Householder matrix. Under Model 1.5 and the assumption of Lemma 4.2, the computed matrix $\widehat{A}_{r+1}$ satisfies

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

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

$$\|\Delta a_j\|_2 \le c_6 \lambda \sqrt{r}\, \widetilde{\gamma}_m(\lambda) \|a_j\|_2 + O(u^2), \quad j = 1 : n, \tag{4.15}$$

holds with probability at least $p_4(\lambda, m, n) = 1 - 2mn\exp(-\lambda^2)$.

Proof. Lemma 4.2 shows that for each j the bound (4.7) holds with $b = a_j$ with probability at least $p_3(\lambda, m)$. The probability that it holds for all columns is given by 1 minus the probability that it fails for at least one, which is at least $1 - n(1 - p_3(\lambda, m)) = 1 - 2mn\exp(-\lambda^2)$. ∎

The probabilistic result for Householder QR factorization, analogous to the worst-case rounding error result Theorem 1.1, now follows.

THEOREM 4.4. *Let $\widehat{R} \in \mathbb{R}^{m\times n}$ be the computed upper trapezoidal QR factor of $A \in \mathbb{R}^{m\times n}$ obtained via the Householder QR algorithm. Under Model 1.5 and the assumption of Lemma 4.2, there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

$$A + \Delta A = Q\widehat{R}, \tag{4.16}$$

where

$$\|\Delta a_j\|_2 \le c_6\lambda\sqrt{n}\widetilde{\gamma}_m(\lambda)\|a_j\|_2 + O(u^2), \quad j = 1:n, \tag{4.17}$$

holds with probability at least $p_4(\lambda, m, n) = 1 - 2mn\exp(-\lambda^2)$.

Proof. The theorem follows from Lemma 4.3 as long as we note two subtleties. First, the Householder matrices in the lemma are completely general, yet for QR factorization they are chosen to introduce zeros into vectors and we explicitly set those elements to zero rather than compute them. This essentially forces rows of ΔP_i in (4.4) to be zero, which does not increase $\|\Delta P_i\|_2$, so the bounds still hold.

The second subtlety is that for Householder QR factorization the Householder vector v_j in Lemma 4.1 depends on previous computed quantities and is computed itself, so is subject to rounding error. The fact that v_j is no longer a constant as regards the conditional probabilities can be dealt with by adding "$v_j, v_{j-1}, \dots, v_1$" to $\mathcal{Z}_s$ in (4.11). The key point is that D_j, $\Lambda_1^{(j)}$, and $\Lambda_2^{(j)}$ depend on rounding errors that occur later than those on which v_j depends. The result (2.4) for the construction of a Householder vector, together with (4.4), shows that the error in v_j adds a perturbation of order $uc_4\widetilde{\gamma}_m(\lambda) = O(u^2)$ to ΔP_j, so it adds an $O(u^2)$ term to the bound (4.6) and hence does not change (4.17). ∎

Theorem 4.4 bounds the backward error to first order, while the analogous worst-case result, Theorem 1.1, bounds all orders. We argue that the exclusion of higher order terms is inconsequential. Set $r = n$ in Lemma 4.2. The $O(u^2)$ term that we dropped from the analysis after the expansion of (4.1) at the beginning of the proof of Lemma 4.2 comprises $\binom{n}{2}$ terms of order u^2 and in view of (4.2) is bounded by (omitting constants c_i and λ) $n^2\widetilde{\gamma}_m(\lambda)^2\|b\|_2 \approx n^2mu^2\|b\|_2$. The first-order term has $\sqrt{mn}u$ dependence, while the second-order term has mn^2u^2 dependence. Thus for problem sizes with $m^{1/2}n^{3/2}u > 1$, second-order terms could dominate. This, in effect, imposes the requirement

$$m^{1/2}n^{3/2}u < 1 \tag{4.18}$$

for our analysis to imply that the error grows like $\sqrt{mn}u$. The higher order terms in the expansion (4.1) do not affect this argument. Indeed the ratio of the bound for the $(k+1)$st-order terms divided by the bound for the kth-order terms is

$$\frac{\binom{n}{k+1}(m^{1/2}u)^{k+1}}{\binom{n}{k}(m^{1/2}u)^k} = \frac{\frac{n!}{(n-k-1)!(k+1)!}m^{(k+1)/2}u^{k+1}}{\frac{n!}{(n-k)!k!}m^{k/2}u^k} = \frac{n-k}{k+1}m^{1/2}u,$$

and this ratio is less than 1 for $k \geq 2$ given (4.18). However, (4.18) is no more restrictive than existing assumptions about the worst-case analysis, as Theorem 1.1 has the assumption

$$mnu < 1. \tag{4.19}$$

Since we have $m \geq n$, and in many applications $m \gg n$, (4.18) can be significantly less restrictive than (4.19). While we ideally would have probabilistic bounds for the higher order terms, our analysis has still shown that we can significantly tighten the error bounds for the same problem sizes for which the worst-case analysis holds.

Ideally, we would remove the assumption in Lemmas 4.2 and 4.3 and Theorem 4.4 that the bound (4.2) holds with probability 1 for all j. We have effectively replaced a probability of $1 - 2rm\exp(-\lambda^2/2)$, which is extremely close to 1 for modest λ as explained in Section 5.1, with a probability of 1, and this suggests that our analysis should still provide bounds that are indicative of the practical behavior. To remove this assumption we need a version of Theorem 3.2 that allows $X_k X_k^T \leq A_k A_k^T$, and $X_k^T X_k \leq A_k^T A_k$ to hold with a certain probability less than 1 rather than with probability 1. Such a result is not, to our knowledge, available in the literature on concentration inequalities, and deriving one requires further research that is beyond the scope of this paper.

The columnwise bound (4.17) implies a normwise one:

$$\|\Delta A\|_F \leq c_6 \lambda \sqrt{n}\, \widetilde{\gamma}_m(\lambda) \|A\|_F + O(u^2). \tag{4.20}$$

The equivalent bound for the worst-case analysis is (1.3).

The conclusion from our analysis is that the constant mnu in the worst-case backward error bound for Householder QR factorization reduces to $\sqrt{mnu}$ in the probabilistic bound.

Finally, we derive a probabilistic error bound on the loss of orthogonality in the computed factor $\widehat{Q}$ formed as the product of the Householder matrices, computed in the more efficient right to left order.

THEOREM 4.5. *Let $\widehat{Q} \in \mathbb{R}^{m\times m}$ be the computed orthogonal QR factor of $A \in \mathbb{R}^{m\times n}$ obtained by forming $Q = P_1 P_2 \dots P_n$ in the right to left order. Under Model 1.5 and the assumption of Lemma 4.2, $\widehat{Q} = Q + \Delta Q$, where the jth column of ΔQ is bounded by*

$$\|\Delta q_j\|_2 \leq c_6 \lambda \sqrt{n}\, \widetilde{\gamma}_m(\lambda) + O(u^2), \quad j = 1:m, \tag{4.21}$$

with probability at least $p_4(\lambda, m, n) = 1 - 2mn\exp(-\lambda^2)$. Moreover,

$$\|\widehat{Q}^T\widehat{Q} - I\|_F \leq c_7 \lambda \sqrt{mn}\, \widetilde{\gamma}_m(\lambda) + O(u^2) \tag{4.22}$$

holds with the same probability.

Proof. Applying Lemma 4.2 with the P_i taken in the reverse order, with $r = n$ and b the jth column of the $m \times m$ identity matrix, gives (4.21), where the probability follows by the same argument as in the proof of Lemma 4.3.

By [13, Thm. 8.17], with U denoting the orthogonal polar factor of $\widehat{Q}$, which is the nearest orthogonal matrix to $\widehat{Q}$ in the Frobenius norm [13, Thm. 8.4], we have

$$\begin{aligned}
\|\widehat{Q}^T\widehat{Q} - I\|_F &\leq (1 + \|\widehat{Q}\|_2)\|\widehat{Q} - U\|_F \\
&\leq (1 + \|\widehat{Q}\|_2)\|\widehat{Q} - Q\|_F \\
&\leq c_7 \lambda \sqrt{mn}\, \widetilde{\gamma}_m(\lambda) + O(u^2).
\end{aligned}$$

∎

TABLE 5.1

The value of $1-p_5(\lambda,m,m)$ for various choices of λ and m. The underlined entries correspond to negative probabilities.

	m			
λ	10^2	10^4	10^6	10^8
6.0	3.0460e−04	3.0460e+00	3.0460e+04	3.0460e+08
7.0	4.5795e−07	4.5795e−03	4.5795e+01	4.5795e+05
8.0	2.5328e−10	2.5328e−06	2.5328e−02	2.5328e+02
9.0	5.1535e−14	5.1535e−10	5.1535e−06	5.1535e−02
10.0	3.8575e−18	3.8575e−14	3.8575e−10	3.8575e−06
11.0	1.0622e−22	1.0622e−18	1.0622e−14	1.0622e−10
12.0	1.0760e−27	1.0760e−23	1.0760e−19	1.0760e−15

The worst-case bound corresponding to (4.21) has constant mnu [12, p. 360], so again the dimensions are square-rooted.

5. Discussion. We now discuss several aspects and implications of the error analysis.

5.1. Choice of λ. We begin with a brief discussion of the probabilities associated with the bounds in Lemma 4.1 and Theorem 4.4. As noted in previous work [7], [15], [16], these probabilities are typically pessimistic and setting $\lambda = 1$ almost always provides bounds that hold in practice. Define $1 - p_5(\lambda, m, n) = 2mn\big(\exp(-\lambda^2) + \exp(-\lambda^2/2)\big)$, which is an upper bound on the probabilities in Lemma 4.1 and Theorem 4.4 that the respective bounds do not hold. Table 5.1 shows values of $1 - p_5(\lambda, m, m)$. For certain values the upper bounds are negative. However, for $\lambda = 10$ and all the problem sizes shown in Table 5.1, p_5 is within 4×10^{-6} of it.

5.2. Aggregated Householder transformations. In practice, Householder transformations are usually aggregated in order to express the computation primarily in terms of matrix multiplication. A common form of aggregation is the WY representation [3], [12, sec. 19.5]. We represent the product $Q_r = P_r P_{r-1} \dots P_1$ of b Householder matrices $P_i = I - v_i v_i^T$ as

$$Q_r = I + W_b Y_b^T, \quad W_b, Y_b \in \mathbb{R}^{m \times b}.$$

This is done through the recurrence

$$W_1 = -v_1, \quad Y_1 = v_1, \qquad W_i = \begin{bmatrix} W_{i-1} & -v_i \end{bmatrix}, \quad Y_i = \begin{bmatrix} Y_{i-1} & Q_{i-1}^T v_i \end{bmatrix}.$$

We partition A as

$$A = \begin{bmatrix} A_1 & B \end{bmatrix}, \quad A_1 \in \mathbb{R}^{m \times r}$$

and transform A_1 to upper trapezoidal form by forming $P_r \dots P_1 A_1$, accumulating the product $P_r \dots P_1 = I + W_r Y_r^T$. The matrix B is then updated via $B \leftarrow B + W_r(Y_r^T B)$, and we repeat this process on the remaining rows of B.

We do not perform a full analysis of the aggregated algorithm. We simply note that the two core operations of the aggregated Householder QR factorization are the application of a sequence of Householder matrices, in forming $P_r \dots P_1 A_1$, and matrix multiplication in forming $B \leftarrow B + W_r(Y_r^T B)$. Both of these operations have been shown to have probabilistic bounds whose constants are the square roots of those in the worst-case bounds, so we can expect the same to be true for the aggregated algorithm.

5.3. Mixed-precision QR factorization. Yang, Fox, and Sanders [22, Thm. 4.1] consider a mixed precision Householder QR factorization algorithm in which the working precision is u_{low} and the inner products are computed at precision u_{high}. They obtain a normwise backward error bound of order $n(u_{\mathrm{low}}+mu_{\mathrm{high}})$. Our analysis is readily adapted for this algorithm by replacing $\widetilde{\gamma}_m = \widetilde{\gamma}_m(u)$ in (4.2) by $\widetilde{\gamma}_m(u_{\mathrm{high}})+O(u_{\mathrm{low}})$, and the resulting probabilistic error bound is of order $n^{1/2}(u_{\mathrm{low}} + m^{1/2}u_{\mathrm{high}})$.

5.4. Two-sided transformations. Householder matrices and Householder QR factorization are tools in many algorithms, including the reduction of matrices to tridiagonal or Hessenberg form for the QR algorithm for eigenvalues and to bidiagonal form for the QR algorithm for singular values [11]. Can we use our results to obtain probabilistic backward error bounds with reduced constants for these reductions? For the reduction of $A \in \mathbb{R}^{n\times n}$ to Hessenberg form H we have $H = P_{n-2}\dots P_1 A P_1^T \dots P_{n-2}^T$. The reduction is carried out in the order

$$H = P_{n-2}(\dots(P_1 A P_1^T)\dots)P_{n-2}^T, \tag{5.1}$$

because P_{j+1} depends on $P_j \dots P_1 A P_1^T \dots P_j^T$. Our analysis does not directly apply to this two-sided case. However, consider applying all the transformations on the left before applying those on the right:

$$H = (P_{n-2}\dots P_1 A)P_1^T \dots P_{n-2}^T \equiv (PA)P^T. \tag{5.2}$$

Two applications of Lemma 4.3 give

$$\begin{aligned}\widehat{H} &= \bigl(P(A+\Delta A_1)+\Delta A_2\bigr)P^T \\ &= P(A+\Delta A)P^T, \qquad \Delta A = \Delta A_1 + P^T\Delta A_2.\end{aligned}$$

The probabilistic bounds on $\|\Delta A_1\|_F$ and $\|\Delta A_2\|_F$ from the lemma are proportional to nu and hence so is the bound for $\|\Delta A\|_F$. The usual worst-case bound obtained by Wilkinson is proportional to n^2u [20, pp. 160–161] (we have accounted for the fact that Wilkinson assumes that inner products are accumulated at twice the working precision).

There are two reasons why the probabilistic backward error bound for (5.2) should be indicative of the backward error for (5.1). First, (5.2) does many more operations, since it incurs substantial fill-in; in particular $P_{n-2}\dots P_1 A$ is full, apart from the first column. We have carried out extensive numerical experiments, in which the backward error for (5.2) was always of the same, or smaller, order of magnitude as that for (5.1). The second reason is that in practice the Householder matrices are aggregated, so that the actual computation lies somewhere between (5.1) and (5.2).

For the multishift Hessenberg QR iteration the bulge chasing is implemented using Householder matrices [5], [6], so Lemma 4.3 can be applied again for sufficiently large bulges, and for small bulges one can apply the worst-case bounds.

5.5. Other forms of QR factorization. The modified Gram–Schmidt algorithm applied to $A \in \mathbb{R}^{m\times n}$ with $m \ge n$ is known to be equivalent both mathematically and numerically to Householder QR factorization applied to the augmented matrix $\left[\begin{smallmatrix}0\\A\end{smallmatrix}\right]$ [4], [12, sec. 19.8]. Hence we can apply Theorem 4.4, and since the augmented matrix is $(m+n)\times n$ the constant obtained will be proportional to $\sqrt{(m+n)nu} \approx \sqrt{mnu}$ since $m \ge n$. (As explained in the above references, some additional work is needed to obtain a backward error result for modified Gram–Schmidt and the constant will be increased.)

In [1], a randomized process for computing the QR factorization of $A \in \mathbb{R}^{m\times n}$ is presented. It uses a randomized Gram-Schmidt process, with the approach based on the dimension reduction technique of random sketching. The authors give a rounding error analysis under the assumption that the rounding errors are mean independent random variables of zero mean. They exploit mixed-precision arithmetic with two precisions $u_{\mathrm{fine}} < u_{\mathrm{crs}}$, where u_{crs} is used for computing the projections and u_{fine} for everything else. The analogous result to Theorem 4.4 is [1, Thm. 3.2] $\|A - \widehat{Q}\widehat{R}\|_F \le cu_{\mathrm{crs}}n^{3/2}\|A\|_F$, where $\widehat{Q} \in \mathbb{R}^{m\times n}$ is the computed orthonormal QR factor.

Finally, we consider Givens QR factorization. Givens transformations effectively operate on vectors of length 2, so there is no benefit to a probabilistic approach in analyzing the application of a single rotation. The backward error analysis [12, sec. 19.6] shows that each individual rotation introduces a backward error bounded by a small constant and that the $m+n-2$ products of disjoint rotations needed for a QR factorization lead to a result of the same form as Theorem 1.1 but with

$$\|\Delta a_j\|_2 \le c_7\gamma_{m+n-2}\|a_j\|_2, \quad j = 1\colon n.$$

A probabilistic analysis analogous to that in the proof of Lemma 4.1, using the matrix concentration inequality, leads to a result of the same form as Theorem 4.4 but with probabilistic bound

$$\|\Delta a_j\|_2 \le c_8(m+n)^{1/2}\,u\,\|a_j\|_2 + O(u^2), \quad j = 1\colon n.$$

6. Backward error for QR factorization. In our numerical experiments in the next section we need to compare the probabilistic backward error bounds with the actual backward errors. How to compute the backward error matrix $\Delta A = A - Q\widehat{R}$ for a given $\widehat{R}$, though, is not clear, since the orthogonal matrix Q in Theorems 1.1 and 4.4 is unknown. We will focus on the backward error measure

$$\begin{aligned}(6.1)\quad \mu(\widehat{R}) &= \min\left\{ \left(\sum_{j=1}^{n} \|d_j\Delta a_j\|_2^2\right)^{1/2} : A + \Delta A = Q\widehat{R},\ Q \in \mathbb{R}^{m\times m},\ Q^TQ = I_m \right\} \\ &= \min\bigl\{ \|(A - Q\widehat{R})D\|_F : Q \in \mathbb{R}^{m\times m},\ Q^TQ = I_m,\ D = \mathrm{diag}(d_j) \bigr\}.\end{aligned}$$

For $D = \mathrm{diag}(\|a_j\|_2^{-1})$ we have a columnwise backward error and for $D = \|A\|_F^{-1}I$ the normwise relative backward error.

The next result shows how to compute $\mu(\widehat{R})$. Recall that the polar decomposition of $A \in \mathbb{R}^{n\times n}$ is a factorization $A = UH$, where U is orthogonal and H is symmetric positive semidefinite.

THEOREM 6.1. *Let $A, B \in \mathbb{R}^{m\times n}$ and let $D = \mathrm{diag}(d_i) \in \mathbb{R}^{n\times n}$ be nonsingular. Then $\min\{\, \|(A - QB)D\|_F : Q \in \mathbb{R}^{m\times m},\ Q^TQ = I_m \,\}$ is obtained for $Q = U^T$, where U is the orthogonal polar factor of the matrix BD^2A^T.*

Proof. For $F, G \in \mathbb{R}^{m\times n}$, the orthogonal Procrustes problem has the form $\min\bigl\{\, \|F - GW\|_F : W \in \mathbb{R}^{n\times n},\ W^TW = I_n \,\bigr\}$ and any orthogonal polar factor of G^TF is a solution [13, Thm. 8.6]. Writing $\|(A - QB)D\|_F = \|AD - QBD\|_F = \|DA^T - DB^TQ^T\|_F$ therefore gives the result on taking $F = DA^T$ and $G = DB^T$. ∎

By Theorem 6.1, $\mu(\widehat{R}) = \|(A - Q\widehat{R})D\|_F$, where Q^T is an orthogonal polar factor of $\widehat{R}D^2A^T$. If $\widehat{R}D^2A^T = U\Sigma V^T$ is a singular value decomposition then we can take $Q^T = UV^T$.

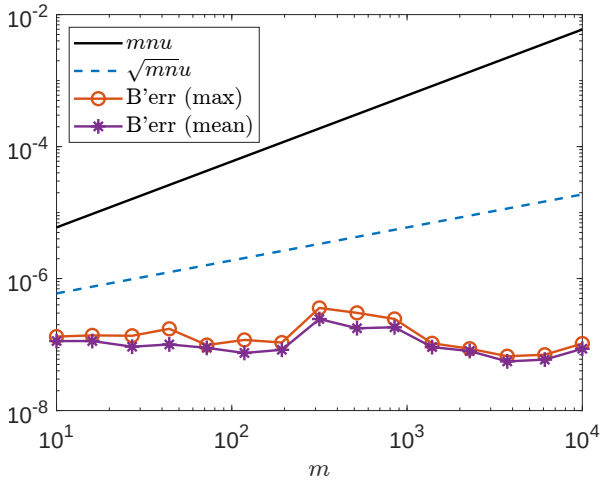

FIG. 7.1. *Normwise backward errors and bounds for Householder QR factorization for* $n = 10$ *and various* m*, for* $m \times n$ *matrices with elements sampled uniformly from* $[0, 1]$.

We note that one can express $\mu(\widehat{R})^2 = \|AD\|_F^2 + \|\widehat{R}D\|_F^2 - 2\sum_{i=1}^n \sigma_i(\widehat{R}D^2A^T)$, where σ_i denotes the ith largest singular value. However, this formula suffers from severe cancellation, and rounding errors can cause it to evaluate as negative, so it is better to use the expression $\|(A - Q\widehat{R})D\|_F$. In fact, even the latter expression does not necessarily give a result of the correct order of magnitude, so it is best to compute the backward error at twice the working precision. Hence in our experiments we take single precision as the working precision and compute the backward error in double precision.

7. Numerical experiments. For all our all numerical experiments we set the parameter $\lambda = 1$ and set all dimension-independent constants c_i in the error bounds (worst-case or probabilistic) to be 1. We use MATLAB R2021b and take IEEE single precision as the working precision. Normwise relative backward errors, as defined by (6.1) with $D = \|A\|_F^{-1} I$, are computed as described in section 6, in double precision.

We use round to nearest in all the experiments. We have tried stochastic rounding, which ensures that the assumptions of Model 1.5 are satisfied [7], and found that the numerical results reported are virtually identical to those for round to nearest in these experiments.

In sections 7.1, 7.3, and 7.4 we use random $m \times n$ matrices with entries drawn uniformly from the interval $[0, 1]$ and we sample 10 different matrices for each pair of dimensions. In section 7.2 we use real-life matrices from the SuiteSparse collection.

7.1. Householder QR factorization for random matrices. In this experiment we test the backward error bound in Theorem 4.4. In order to study how the error grows with n and m independently, we first fix a value of n and vary m and then fix m and vary n. For each pair of dimensions we plot the maximum and mean normwise backward errors for Householder QR factorization along with associated worst-case and probabilistic bounds obtained from (1.3) and (4.20). The results are given in Figures 7.1 and 7.2. We see that the probabilistic bound $\sqrt{mnu}$ proves a much better indicator than the worst-case bound mnu of the size of the error and its rate of growth.

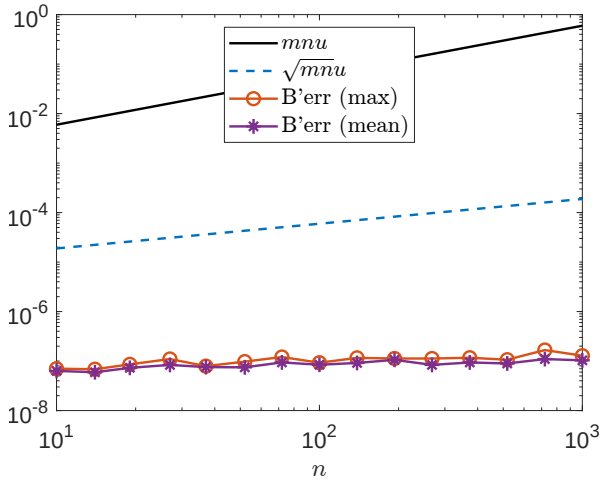

FIG. 7.2. *Normwise backward errors and bounds for Householder QR factorization for* $m = 10^4$ *and various* n*, for* $m \times n$ *matrices with elements sampled uniformly from* $[0, 1]$.

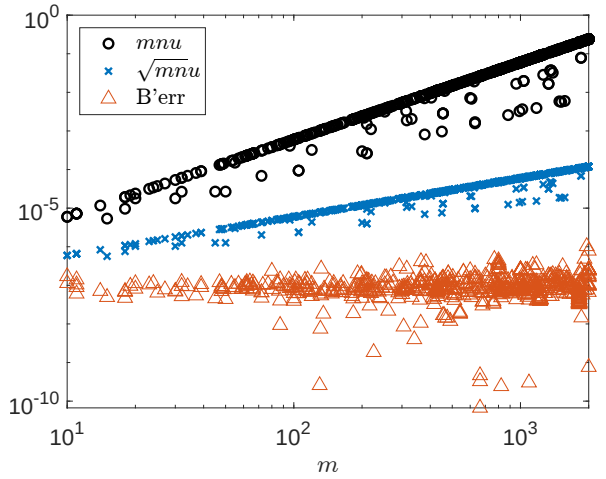

FIG. 7.3. *Normwise backward errors and bounds for* 774 $m \times n$ *matrices from the SuiteSparse collection.*

7.2. Householder QR factorization for SuiteSparse matrices. We also consider Householder QR factorization of some matrices from the SuiteSparse Matrix Collection [9], [10]. We select all matrices from the collection with $10 \leq m, n \leq 2 \times 10^3$ and $m \geq n$. This results in 842 matrices. We plot the same error quantities as for the random matrices. Some of the matrices in the collection have zero columns, so we filter these out. There are a few cases where the reported error exceeds even the deterministic bound, which we suspect is an underflow issue; a similar observation is made in [15, sec. 4.5]. We also filter out these cases from the reported results, which results in 774 matrices. We show the observed backward errors in Figure 7.3. Again, the probabilistic bound is satisfied and proves closer to the observed error than the worst-case bound by several orders of magnitude.

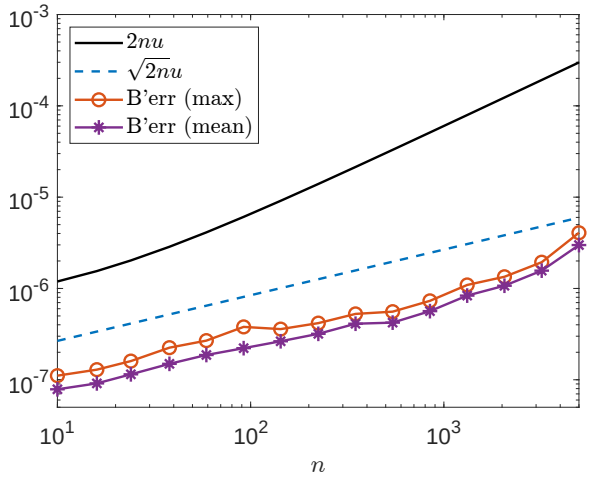

FIG. 7.4. *Normwise backward error and backward error bound for Givens QR factorization for $n \times n$ matrices with elements sampled uniformly from $[0, 1]$.*

7.3. Givens rotations. Givens QR factorization is typically used for structured matrices, such as tridiagonal or upper Hessenberg matrices, but here we wish to see how the backward error of the factorization behaves for dense matrices. For random $n \times n$ matrices we plot in Figure 7.4 the maximum and mean normwise backward errors, the worst-case error bound $2nu$, and the probabilistic error bound $\sqrt{2n}\,u$. The backward error grows at a rate very similar to that of the probabilistic bound.

7.4. Reduction to Hessenberg form. Finally, we consider Householder reduction matrix to Hessenberg form: $A = QHQ^T$. Figure 7.5 shows normwise backward errors, computed as $\|A - \widehat{Q}H\widehat{Q}^T\|_F / \|A\|_F$, where $\widehat{Q}$ is the computed product of Householder matrices (we do not have an explicit backward error formula such as that in Theorem 6.1 in this two-sided case), and the worst-case and probabilistic bounds, n^2u and nu respectively. We see that the backward error satisfies the probabilistic bound, and again the probabilistic bound is a better indicator than the worst-case bound of the rate of growth of the backward error with n.

8. Conclusions. In a classic 1961 paper, Wilkinson [19, p. 318] carries out rounding error analyses of LU factorization, Givens QR factorization, and Householder QR factorization. He notes that "The bounds we have obtained are in all cases strict upper bounds. In general, the statistical distribution of the rounding errors will reduce considerably the function of n occurring in the relative errors. We might expect in each case that this function should be replaced by something which is no bigger than its square root and is usually appreciably smaller." Recent probabilistic rounding error analysis has provided a rigorous foundation for Wilkinson's statement for LU factorization and other inner product-based computations. Our work does the same for Householder QR factorization-based methods, as well as for Givens QR factorization, under the technical assumption in Lemma 4.2.

A significant feature of the probabilistic backward error bounds is that they bound the likely rate of growth of the backward error as the problem dimensions increase. The rate of growth, along with blocking, exploiting architectural features of the hardware, and using other techniques to improve the accuracy of the computations, is what determines our ability to solve problems at extreme scale and possibly low precision

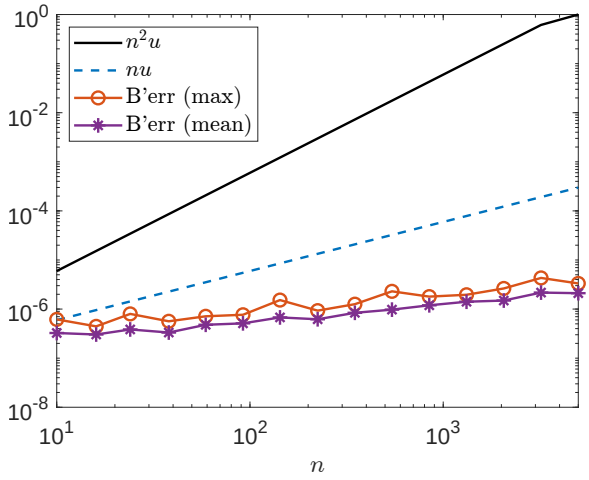

FIG. 7.5. *Normwise backward errors and bounds for reduction to Hessenberg form for $n \times n$ matrices with elements sampled uniformly from $[0, 1]$.*

in a numerically stable way [14].

Acknowledgments. We thank Joel Tropp, Neil Walton, and Marcus Webb for advice on concentration inequalities, and Theo Mary for comments on a draft manuscript. We also thank the referees and associate editor for their suggestions.

REFERENCES

[1] Oleg Balabanov and Laura Grigori. Randomized Gram–Schmidt process with application to GMRES. *SIAM J. Sci. Comput.*, 44(3):A1450–A1474, 2022.

[2] Patrick Billingsley. *Probability and Measure*. Third edition, Wiley, New York, USA, 1995. ISBN 0-471-00710-2.

[3] Christian H. Bischof and Charles F. Van Loan. The WY representation for products of Householder matrices. *SIAM J. Sci. Statist. Comput.*, 8(1):s2–s13, 1987.

[4] Åke Björck and C. C. Paige. Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, 13(1):176–190, 1992.

[5] Karen Braman, Ralph Byers, and Roy Mathias. The multishift QR algorithm. Part I: Maintaining well-focused shifts and level 3 performance. *SIAM J. Matrix Anal. Appl.*, 23(4): 929–947, 2002.

[6] Karen Braman, Ralph Byers, and Roy Mathias. The multishift QR algorithm. Part II: Aggressive early deflation. *SIAM J. Matrix Anal. Appl.*, 23(4):948–973, 2002.

[7] Michael P. Connolly, Nicholas J. Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM J. Sci. Comput.*, 43(1):A566–A585, 2021.

[8] Matteo Croci, Massimiliano Fasi, Nicholas J. Higham, Theo Mary, and Mantas Mikaitis. Stochastic rounding: Implementation, error analysis and applications. *Roy. Soc. Open Sci.*, 9 (3):1–25, 2022.

[9] Timothy A. Davis. SuiteSparse Matrix Collection. https://sparse.tamu.edu/.

[10] Timothy A. Davis and Yifan Hu. The University of Florida Sparse Matrix Collection. *ACM Trans. Math. Software*, 38(1):1:1–1:25, 2011.

[11] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Fourth edition, Johns Hopkins University Press, Baltimore, MD, USA, 2013. xxi+756 pp. ISBN 978-1-4214-0794-4.

[12] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. xxx+680 pp. ISBN 0-89871-521-0.

[13] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. xx+425 pp. ISBN 978-0-898716-46-7.

[14] Nicholas J. Higham. Numerical stability of algorithms at extreme scale and low precisions. MIMS EPrint 2021.14, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, September 2021. 21 pp. To appear in Proc. Int. Cong. Math.

[15] Nicholas J. Higham and Theo Mary. A new approach to probabilistic rounding error analysis. *SIAM J. Sci. Comput.*, 41(5):A2815–A2835, 2019.

[16] Nicholas J. Higham and Theo Mary. Sharper probabilistic backward error analysis for basic linear algebra kernels with random data. *SIAM J. Sci. Comput.*, 42(5):A3427–A3446, 2020.

[17] Michael Mitzenmacher and Eli Upfal. *Probability and Computing. Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, Cambridge, UK, 2017. xx+467 pp. ISBN 978-1-107-15488-9.

[18] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[19] J. H. Wilkinson. Error analysis of direct methods of matrix inversion. *J. ACM*, 8:281–330, 1961.

[20] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, Oxford, UK, 1965. xviii+662 pp. ISBN 0-19-853403-5 (hardback), 0-19-853418-3 (paperback).

[21] David Williams. *Weighing the Odds. A Course in Probability and Statistics*. Cambridge University Press, Cambridge, UK, 2001. xvii+547 pp. ISBN 0-521-00618-X.

[22] L. Minah Yang, Alyson Fox, and Geoffrey Sanders. Rounding error analysis of mixed precision block Householder QR algorithms. *SIAM J. Sci. Comput.*, 43(3):A1723–A1753, 2021.