A New Approach to Probabilistic Rounding Error Analysis

Higham, Nicholas J. and Mary, Theo

2018

MIMS EPrint: **2018.33**

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

A NEW APPROACH TO PROBABILISTIC ROUNDING ERROR ANALYSIS*

NICHOLAS J. HIGHAM† AND THEO MARY‡

Abstract. Traditional rounding error analysis in numerical linear algebra leads to backward error bounds involving the constant $\gamma_n = nu/(1-nu)$, for a problem size n and unit roundoff u. In the light of large-scale and possibly low-precision computations, such bounds can struggle to provide any useful information. We develop a new probabilistic rounding error analysis that can be applied to a wide range of algorithms. By using a concentration inequality and making probabilistic assumptions about the rounding errors, we show that in several core linear algebra computations γ_n can be replaced by a relaxed constant $\widetilde{\gamma}_n$ proportional to $\sqrt{n \log n}\, u$ with a probability bounded below by a quantity independent of n. The new constant $\widetilde{\gamma}_n$ grows much more slowly with n than γ_n. Our results have three key features: they are backward error bounds; they are exact, not first order; and they are valid for any n, unlike results obtained by applying the central limit theorem, which apply only as $n \to \infty$. We provide numerical experiments that show that for both random and real-life matrices the bounds can be much smaller than the standard deterministic bounds and can have the correct asymptotic growth with n. We also identify two special situations in which the assumptions underlying the analysis are not valid and the bounds do not hold. Our analysis provides, for the first time, a rigorous foundation for the rule of thumb that "one can take the square root of an error constant because of statistical effects in rounding error propagation".

Key words. Rounding error analysis, floating-point arithmetic, numerical linear algebra.

AMS subject classifications. 65G50, 65Fxx

1. Introduction.

Modern rounding error analysis faces a double challenge with the rise of large-scale, mixed-precision computations. On the one hand, increasingly large problems are being solved. For example, the supercomputers currently at the top of the TOP500 list[1] are there by virtue of having solved, in record time, linear systems $Ax = b$ of dimensions of order 10^8 by LU factorization, and future exascale systems will solve problems of even larger size. Traditional error analysis does not guarantee small residuals for such large systems because the error constants are so large—yet a small residual is obtained in practice.

On the other hand, low-precision floating-point arithmetic—in particular half precision (fp16)—is becoming increasingly attractive due to both its higher speed and its lower energy consumption [5], [9], [10]. For half precision arithmetic the unit roundoff is so large that traditional bounds cannot guarantee an accurate inner product even for relatively small problems (of dimensions in the thousands). Yet, machine learning algorithms do successfully run in half, or even lower, precision.

This discrepancy between theory and practice stems from the fact that traditional rounding error bounds are worst-case bounds and so are pessimistic on average. In most practical cases, they do not provide good estimates of the size of the error, and in particular they overestimate the error growth, that is, the asymptotic dependence of the error on the problem size.

*Version of March 18, 2019. **Funding**: This work was supported by Engineering and Physical Sciences Research Council grant EP/P020720/1, The MathWorks, and the Royal Society. The opinions and views expressed in this publication are those of the authors, and not necessarily those of the funding bodies.

†School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (nick.higham@manchester.ac.uk, http://www.maths.manchester.ac.uk/~higham)

‡School of Mathematics, The University of Manchester, Manchester M13 9PL, UK (theo.mary@manchester.ac.uk)

[1] https://www.top500.org/

Since the beginning of the digital computer era many researchers have modeled rounding errors as random variables in an attempt to obtain better estimates of how the error behaves on average. These include (in chronological order) von Neumann and Goldstine [27], Henrici [11], [12], [13], Hull and Swenson [17], Tienari [26], Barlow and Bareiss [1], Chatelin and Brunet [7], and Calvetti [4]. These treatments typically linearize the forward error into a sum of the form $e = \sum_{i=1}^{n} \delta_i t_i$, with $|\delta_i| \leq u$ and where u is the unit roundoff. The key intuition is that it is very unlikely that $|e|$ will attain its worst-case magnitude $u \sum_{i=1}^{n} |t_i|$, which can happen only when each δ_i is of maximal magnitude and the $\delta_i t_i$ have identical signs. In most of these references the δ_i are assumed to be independent random variables with mean zero, and usually also assumed to be from a uniform distribution or a normal distribution. The central limit theorem (e.g., [2, sec. 27]) shows that as $n \to \infty$ the probability distribution of $e/(\sum_{i=1}^{n} t_i^2)^{1/2}$ tends towards a normal distribution of mean zero and standard deviation $\sigma \leq u$; therefore for sufficiently large n, the probability that $|e|$ will not exceed $u(\sum_{i=1}^{n} t_i^2)^{1/2}$ times a small constant λ is very high (e.g., by the "three-sigma rule", it is about 99.7% for $\lambda = 3$). Compared to the worst-case constant $\sum_{i=1}^{n} |t_i|$, $(\sum_{i=1}^{n} t_i^2)^{1/2}$ can be smaller by a factor up to $\sqrt{n}$.

This probabilistic approach to rounding error analysis has led to the well-known rule of thumb, based on informal arguments and assumptions, that constants in rounding error bounds can be replaced by their square roots. For example, Wilkinson applies this rule of thumb in [28, p. 318], [29, pp. 26, 52, 102, 151]. However, a rigorous result along these lines for a wide class of algorithms has not previously been obtained, to our knowledge.

The goal of this work is to obtain rigorous probabilistic rounding error bounds for a wide range of linear algebra algorithms based on a minimal number of assumptions and to test them experimentally.

Previous probabilistic rounding error analyses suffer from four important shortcomings. Our results, based on Theorem 2.4 in the next section, overcome these shortcomings as follows.

Backward rather than forward bounds. Previous work has almost exclusively focused on forward error bounds. We choose instead to analyze backward errors. Backward error bounds have the advantage that they bound perturbations to the data and can be interpreted without the need for condition numbers. Moreover, forward error bounds can be directly derived from backward ones as their product with the condition number of the problem.

Bounds correct to all orders. Rounding error analysis of a sequence of operations leads to products of terms $1+\delta_i$ with $|\delta_i| \leq u$. When these products are linearized and the central limit theorem is applied to the first order term, bounds containing (implicitly or explicitly) a "$+O(u^2)$" term are obtained. In the proof of Theorem 2.4 we overcome this limitation by taking the logarithm of the product, thereby transforming it to a sum. This transformation has the drawback of introducing nonlinearities, but we are able to bound the expected value of a random variable of the form $\log(1 + \delta_i)$ using Taylor expansions.

Fewer assumptions for a more rigorous proof. Some assumptions made in previous probabilistic error analysis are unnecessary. Our analysis requires only two assumptions beyond rounding errors being bounded in modulus by u (see Model 2.1 below): that the rounding errors have mean zero and that they are independent. In particular, in contrast to much existing work we do not assume any specific probability distribution for the rounding errors (e.g., uniform or normal). Such assumptions are usually made so as to bound the standard deviation σ of the rounding errors.

However, since rounding errors are bounded by u, σ is obviously also bounded by u. Additional assumptions on σ are thus mostly unnecessary, as they would only slightly improve the constants in the resulting bounds. Moreover, our assumptions even allow the rounding errors not to be identically distributed, as long as they are independent.

Finally, the major assumption that we drop is the one on the problem size n. To the best of our knowledge, all previous work is based on the crucial assumption that n is "sufficiently large", which is necessary to use the central limit theorem. However, it is difficult to quantify the accuracy of the resulting approximation for a given n. In Theorem 2.4, we overcome this issue by using a concentration inequality (specifically, Hoeffding's inequality [16]) instead of the central limit theorem, and this inequality is valid for all n.

General framework applicable to a wide class of algorithms. Modern rounding error analysis builds on a basic result (Lemma 2.2 below) that bounds the distance from 1 of a product $\prod_{i=1}^{n}(1+\delta_i)$, where $|\delta_i| \le u$, by

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

for $nu < 1$. We provide in Theorem 2.4 a probabilistic analogue of this result with bounds proportional to $\sqrt{n}u$ and with no restriction on n. We also derive corresponding bounds for an inner product combined with a subtraction and a division, matrix–vector products, matrix–matrix products, the solution of triangular systems, and LU and Cholesky factorizations. These are all key computational kernels and so our analysis facilitates the probabilistic error analysis of a wide class of algorithms.

We note that recent work has shown how to relax the condition $nu < 1$ in some traditional analysis, at the cost of stronger assumptions on the arithmetic than (1.2) below and more complicated proofs [20], [25].

In the next section we obtain the main result of this paper: we show that under a certain probabilistic model the constant (1.1) in the basic result described above can be replaced by a constant $\widetilde{\gamma}_n$ proportional to $\sqrt{n}u$ with probability at least a certain value. We apply this result in section 3 to a variety of numerical linear algebra algorithms. In section 4, we perform an extensive set of numerical experiments on both random and real-life matrices. We provide our concluding remarks in section 5.

Throughout the paper we denote the expectation and standard deviation of a random variable x by $\mathbb{E}(x)$ and $\sigma(x)$, respectively. We use the following classical model for floating-point arithmetic [14, sec. 2.2]:

$$\mathrm{fl}(a \mathrel{\mathrm{op}} b) = (a \mathrel{\mathrm{op}} b)(1+\delta), \quad |\delta| \le u, \quad \mathrm{op} \in \{+, -, \times, /, \sqrt{}\}. \tag{1.2}$$

This model holds for IEEE arithmetic [19]; indeed, the IEEE standard requires more: that $\mathrm{fl}(a \mathrel{\mathrm{op}} b)$ is the correctly rounded (to nearest) value of $a \mathrel{\mathrm{op}} b$. We will refer to δ as the rounding error in the operation, though the absolute error $a \mathrel{\mathrm{op}} b - \mathrm{fl}(a \mathrel{\mathrm{op}} b)$ is perhaps more commonly so-described.

2. Probabilistic bound for product of rounding errors. To derive our probabilistic error bounds we will use the following model of rounding errors in a given computation.

MODEL 2.1 (probabilistic model of rounding errors). *In the computation of interest, the quantities δ in the model (1.2) associated with every pair of operands are independent random variables of mean zero.*

Note that this model does not require the rounding errors to be identically distributed.

Model 2.1 is clearly not always realistic. For example, in some cases δ is necessarily zero, such as when the operands are (not too large) integers in an addition, subtraction, or multiplication; when the operands in a subtraction differ by at most a factor two and so are subtracted exactly (by Sterbenz's result [14, Thm. 2.5]); or when one of the operands is a power of two in a multiplication or division. Or pairs of operands might be repeated, so that different occurrences of δ are in the fact the same. Indeed non-pathological examples can be found where rounding errors are strongly correlated—notably a rational function example of Kahan [14, sec. 1.17]. More subtly, if an operand comes from an earlier computation it will depend on an earlier δ and so the new δ will depend on the previous one, violating the independence assumption.

We can hope, nevertheless, that conclusions drawn from Model 2.1 remain approximately true. Indeed, as Kahan notes [21] "The fact that rounding errors are neither random nor uncorrelated will not in itself preclude the possibility of modelling them usefully by uncorrelated random variables." In a similar vein, Hull and Swenson [17] point out that "There is no claim that ordinary rounding and chopping are random processes, or that successive errors are independent. The question to be decided is whether or not these particular probabilistic models of the processes will adequately describe what actually happens."

In backward error analysis, products of terms of the form $1+\delta_i$ or its reciprocal, where δ_i is a rounding error satisfying (1.2), are ubiquitous. These products are typically simplified by means of the following result [14, Lem. 3.1], which employs the constant $\gamma_n = nu/(1-nu)$ in (1.1).

LEMMA 2.2 (Deterministic error bound). *If $|\delta_i| \le u$ and $\rho_i = \pm 1$ for $i = 1\colon n$, and $nu < 1$, then*

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

We now derive a probabilistic version of this result containing a constant $\widetilde{\gamma}_n$ that is proportional to $\sqrt{n}$ rather than n. Note that the new constant $\widetilde{\gamma}_n$ does not require that $nu < 1$. To do so, we need to use a concentration inequality, that is, an inequality that bounds the probability that the sum of n independent random variables X_i deviates from its expected value by a given quantity. Several such inequalities exist [3]; we choose to use Hoeffding's inequality [16, Thm. 2], which requires the variables X_i to be bounded.

LEMMA 2.3 (Hoeffding's inequality). *Let $X_1, \dots, X_n$ be independent random variables satisfying*

$$|X_i| \le c_i, \quad i = 1\colon n.$$

Then the sum $S = \sum_{i=1}^{n} X_i$ satisfies

$$\Pr\left(|S - \mathbb{E}(S)| \ge \xi\right) \le 2\exp\left(-\frac{\xi^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

We are now ready to state our main result. Define

$$\widetilde{\gamma}_n(\lambda) = \exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right) - 1. \tag{2.1}$$

THEOREM 2.4 (Probabilistic error bound). *Let $\delta_1, \delta_2, \ldots, \delta_n$ be independent random variables of mean zero bounded in absolute value by the unit roundoff u, and let $\rho_i = \pm 1$, $i = 1\colon n$. Then for any constant $\lambda > 0$ the relation*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \widetilde{\theta}_n$$

holds with $|\widetilde{\theta}_n| \le \widetilde{\gamma}_n(\lambda)$ with probability at least

$$P(\lambda) = 1 - 2\exp\left(-\frac{\lambda^2(1-u)^2}{2}\right). \tag{2.2}$$

Proof. Let $\phi = \prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$. Then $\log\phi = \sum_{i=1}^{n}\rho_i\log(1+\delta_i)$ is a sum of n independent random variables. Since $|\delta_i| \le u < 1$ for all i we can use the Taylor expansion

$$\log(1+\delta_i) = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}\delta_i^k}{k}$$

to obtain the upper and lower bounds

$$\log(1+\delta_i) \le \delta_i + \sum_{k=2}^{\infty}|\delta_i|^k = \delta_i + \frac{\delta_i^2}{1-|\delta_i|},$$
$$\log(1+\delta_i) \ge \delta_i - \sum_{k=2}^{\infty}|\delta_i|^k = \delta_i - \frac{\delta_i^2}{1-|\delta_i|}.$$

We therefore have

$$\delta_i - \frac{u^2}{1-u} \le \log(1+\delta_i) \le \delta_i + \frac{u^2}{1-u}, \tag{2.3}$$

and taking the absolute value we obtain

$$|\rho_i\log(1+\delta_i)| = |\log(1+\delta_i)| \le u + \frac{u^2}{1-u} = \frac{u}{1-u}.$$

We can therefore apply Lemma 2.3 with $X_i = \rho_i\log(1+\delta_i)$ and $c_i = u/(1-u)$. We obtain the bound

$$\Pr\left(|\log\phi - \mathbb{E}(\log\phi)| \ge \xi\right) \le 2\exp\left(\frac{-\xi^2(1-u)^2}{2nu^2}\right).$$

Moreover, by taking the expected value in (2.3) and using $\mathbb{E}(\delta_i) = 0$, we obtain

$$\left|\mathbb{E}(\log(1+\delta_i))\right| \le \frac{u^2}{1-u}, \tag{2.4}$$

and therefore $\mathbb{E}(\log\phi)$ can be bounded by linearity of the expected value:

$$|\,\mathbb{E}(\log\phi)| \le \frac{nu^2}{1-u}.$$

We then have

$$|\log\phi - \mathbb{E}(\log\phi)| \ge |\log\phi| - |\,\mathbb{E}(\log\phi)| \ge |\log\phi| - \frac{nu^2}{1-u},$$

which yields

$$\Pr\left(|\log\phi| - \frac{nu^2}{1-u} \geq \xi\right) \leq \Pr\left(|\log\phi - \mathbb{E}(\log\phi)| \geq \xi\right) \leq 2\exp\left(\frac{-\xi^2(1-u)^2}{2nu^2}\right).$$

Therefore, with probability at least $1 - 2\exp(-\xi^2(1-u)^2/(2nu^2))$, $\log\phi$ lies within the interval with endpoints $\pm(\xi + nu^2/(1-u))$. To make this probability independent of n, we set $\xi = \lambda\sqrt{n}u$, for some constant $\lambda > 0$ independent of n. Then the bound

$$-\lambda\sqrt{n}u - nu^2/(1-u) \leq \log\phi \leq \lambda\sqrt{n}u + nu^2/(1-u)$$

holds with probability at least

$$P(\lambda) = 1 - 2\exp\left(-\frac{\lambda^2(1-u)^2}{2}\right).$$

Taking the exponential, the bound

$$\mathrm{e}^{-t} \leq \phi \leq \mathrm{e}^{t}, \quad t = \lambda\sqrt{n}u + \frac{nu^2}{1-u}$$

also holds with probability at least $P(\lambda)$. Then

$$|\phi - 1| \leq \max\left(|\mathrm{e}^{t} - 1|, |\mathrm{e}^{-t} - 1|\right) = \mathrm{e}^{t} - 1$$

holds with probability at least $P(\lambda)$. The equality $\widetilde{\theta}_n = \phi - 1$ concludes the proof. ∎

Since $\mathrm{e}^t \leq 1 + t/(1-t)$ for $0 \leq t < 1$, the constant $\widetilde{\gamma}_n(\lambda)$ in (2.1) satisfies

$$\widetilde{\gamma}_n(\lambda) \leq \frac{\lambda\sqrt{n}u + \dfrac{nu^2}{1-u}}{1 - \left(\lambda\sqrt{n}u + \dfrac{nu^2}{1-u}\right)} = \lambda\sqrt{n}u + O(u^2), \quad \lambda\sqrt{n}u + \frac{nu^2}{1-u} < 1. \tag{2.5}$$

Note that $\widetilde{\gamma}_n(\lambda)$ is defined for any n, though the bound (2.5) requires $\lambda\sqrt{n}u + nu^2/(1-u) < 1$.

The probability $P(\lambda)$ in (2.2) is independent of n. Moreover, it rapidly approaches 1 as λ increases and is essentially independent of u, as shown in Table 2.1.

It is also important to note that the second order part of the argument of the exponential in (2.1) is innocuous. The argument is $\lambda(\sqrt{n}u) + (\sqrt{n}u)^2/(1-u)$, so for $\lambda \geq 1$ (say) the second order term is smaller than the first order term unless the latter is of order 1, in which case the error bound of Theorem 2.4 provides no useful information.

Many rounding error analyses rely on Lemma 2.2 and can therefore potentially yield probabilistic bounds if they are adapted to make use of Theorem 2.4. In the next section we explore some examples from numerical linear algebra.

3. Application to numerical linear algebra. We now apply Theorem 2.4 within the error analysis of a variety of algorithms in numerical linear algebra. We aim to derive probabilistic bounds that have the same form as the original ones but with γ_n replaced by $\widetilde{\gamma}_n(\lambda)$.

Table 2.1: Values of $P(\lambda)$ in (2.2) to four decimal places for half precision (fp16) and double precision (fp64) arithmetic.

λ	fp16	fp64
2	0.7288	0.7293
3	0.9777	0.9778
4	0.9993	0.9993
5	1.0000	1.0000

3.1. Inner products. We first apply Theorem 2.4 to the computation of the inner product of two vectors. Recall that $\widetilde{\gamma}_n(\lambda)$ and $P(\lambda)$ are defined in (2.1) and (2.2), respectively. We define

$$Q(\lambda, n) = 1 - n(1 - P(\lambda)). \tag{3.1}$$

Note that Q can be negative, but $Q(\lambda, n) \in [0, 1]$ for sufficiently large λ. Here, and throughout, inequalities between matrices and vectors hold componentwise: thus $|A| \le |B|$ means that $|a_{ij}| \le |b_{ij}|$ for all i and j.

THEOREM 3.1 (inner products). *Let $y = a^T b$, where $a, b \in \mathbb{R}^n$, be evaluated in floating-point arithmetic. Under Model 2.1, no matter what the order of evaluation the computed $\widehat{y}$ satisfies*

$$\widehat{y} = (a + \Delta a)^T b, \quad |\Delta a| \le \widetilde{\gamma}_n(\lambda)|a|, \tag{3.2}$$

$$= a^T(b + \Delta b), \quad |\Delta b| \le \widetilde{\gamma}_n(\lambda)|b|, \tag{3.3}$$

with probability at least $Q(\lambda, n)$.

Proof. Assume, first, that the sum $s_n = a_1 b_1 + \cdots + a_n b_n$ is evaluated from left to right, i.e., with the recursive relation $s_i = s_{i-1} + a_i b_i$, starting with $s_0 = 0$. The computed intermediate quantities $\widehat{s}_i$ satisfy

$$\widehat{s}_i = \big(\widehat{s}_{i-1} + a_i b_i (1 + \varepsilon_i)\big)(1 + \delta_i), \quad |\varepsilon_i|, |\delta_i| \le u, \tag{3.4}$$

where ε_i and δ_i represent the rounding errors from the products and additions, respectively, and $\delta_1 = 0$. We therefore have

$$\widehat{s}_n = \sum_{i=1}^{n} a_i b_i (1 + \varepsilon_i) \prod_{j=\max(i,2)}^{n} (1 + \delta_j) =: \sum_{i=1}^{n} a_i b_i (1 + \psi_i), \tag{3.5}$$

where by Theorem 2.4 $|\psi_i| \le \widetilde{\gamma}_{n-\max(i,2)+2}(\lambda)$ holds for any particular i with probability at least $P(\lambda)$. For a given i, the latter bound fails to hold with probability at most $1 - P(\lambda)$, therefore by the inclusion–exclusion principle [30, p. 39] the bound fails to hold for at least one i with probability at most $n(1 - P(\lambda))$. It follows that the probability that the bounds hold for all i is at least $Q(\lambda, n)$. Clearly, $|\psi_i| \le \widetilde{\gamma}_n(\lambda)$ holds for all i with at least the same probability, and it is not hard to see that this bound holds for any ordering. ∎

From the backward error bound of Theorem 3.1 we immediately have the forward error bound

$$\frac{|y - \widehat{y}|}{|y|} \le \widetilde{\gamma}_n(\lambda) \frac{|a|^T |b|}{|a^T b|} \tag{3.6}$$

with the same probability bound.

Next we analyze an important kernel that appears in the solution of triangular systems and in LU factorization.

THEOREM 3.2. *Let $y = \bigl(c - \sum_{i=1}^{k-1} a_i b_i\bigr)/b_k$ be evaluated in floating-point arithmetic. Under Model 2.1, no matter what the order of evaluation the computed $\widehat{y}$ satisfies*

$$\widehat{y} b_k \bigl(1 + \mu_0\bigr) = c - \sum_{i=1}^{k-1} a_i b_i (1 + \mu_i), \tag{3.7}$$

where $|\mu_i| \le \widetilde{\gamma}_k(\lambda)$ for all i, with probability at least $Q(\lambda, k)$.

Proof. Assume for the moment that we first form the sum $s_{k-1} = a_1 b_1 + \cdots + a_{k-1} b_{k-1}$ and then subtract the result from c and divide by b_k. Then

$$\widehat{y} = \frac{(c - \widehat{s}_{k-1})}{b_k}(1 + \delta_k)(1 + \delta_{k+1}), \quad |\delta_k|, |\delta_{k+1}| \le u,$$

and so by (3.5)

$$\widehat{y} b_k (1 + \delta_k)^{-1} (1 + \delta_{k+1})^{-1} = c - \sum_{i=1}^{k-1} a_i b_i (1 + \varepsilon_i) \prod_{j=\max(i,2)}^{k-1} (1 + \delta_j).$$

Applying Theorem 2.4 to each of the rounding error terms gives

$$\widehat{y} b_k (1 + \psi_0) = c - \sum_{i=1}^{k-1} a_i b_i (1 + \psi_i), \tag{3.8}$$

where $|\psi_0| \le \widetilde{\gamma}_2(\lambda)$ holds with probability at least $P(\lambda)$ and $|\psi_i| \le \widetilde{\gamma}_{k-\max(i,2)+1}(\lambda)$ holds for any particular i with probability at least $P(\lambda)$. By the same reasoning as in the proof of Theorem 3.1, these inequalities on $|\psi_i|$ hold for $i = 0\colon k-1$ with probability at least $Q(\lambda, k)$.

Different orderings of the evaluation will give different expressions of the form (3.8) with different numbers of $(1 + \delta_i)^{\pm 1}$ terms corresponding to each $1 + \psi_i$, but there can never be more than k such terms. The worst case, which leads to $\widetilde{\gamma}_k(\lambda)$, is when the subtraction with c is done first—see [14, Lem. 8.2]. The theorem as stated therefore holds for any ordering. ∎

Note that Theorem 3.2 is a backward error result showing that the computed $\widehat{y}$ is the exact result for a perturbed set of b_i. Alternatively, the result can be recast to perturb the a_i and c instead of the b_i.

3.2. Matrix–vector and matrix–matrix products. Building on the error analysis for inner products we can obtain results for matrix–vector and matrix–matrix products.

THEOREM 3.3 (Matrix–vector products). *Let $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $y = Ax$. Under Model 2.1, the computed result $\widehat{y}$ satisfies*

$$\widehat{y} = (A + \Delta A)x, \quad |\Delta A| \le \widetilde{\gamma}_n(\lambda)|A| \tag{3.9}$$

with probability at least $Q(\lambda, mn)$.

Proof. The vector y is obtained via m inner products $y_i = a_i^T x$, where a_i is the ith row of A. By Theorem 3.1, we know that

$$\widehat{y}_i = (a_i + \Delta a_i)^T x, \quad |\Delta a_i| \le \widetilde{\gamma}_n(\lambda)|a_i| \tag{3.10}$$

holds with probability at least $Q(\lambda, n)$. Combining the m instances of (3.10) for $i = 1\colon m$, we have that (3.9) holds with probability at least $1 - m(1 - Q(\lambda, n)) = Q(\lambda, mn)$. ∎

THEOREM 3.4 (Matrix–matrix products). *Let $C = AB$ with $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Under Model 2.1, the jth column of the computed $\widehat{C}$ satisfies*

$$\widehat{c}_j = (A + \Delta A_j)b_j, \quad |\Delta A_j| \le \widetilde{\gamma}_n(\lambda)|A|, \quad j = 1\colon n, \tag{3.11}$$

with probability at least $Q(\lambda, mn)$, and hence

$$|C - \widehat{C}| \le \widetilde{\gamma}_n(\lambda)|A||B| \tag{3.12}$$

with probability at least $Q(\lambda, mnp)$.

Proof. Equation (3.11) is simply an application of Theorem 3.3. The bound (3.12) follows by combining the p instances of (3.11) and from the fact that $1 - p(1 - Q(\lambda, mn)) = Q(\lambda, mnp)$. ∎

The two loops used to evaluate a matrix–vector product can be ordered in an "ij" form based on inner products or a "ji" form based on vector operations. While the proof of Theorem 3.3 assumes the use of inner products, the error bound is nevertheless applicable to both orderings. As for standard error analysis [14, sec. 3.5], different orderings result in the same operations being performed in a different order, so the same rounding errors are generated but in a different order and the same error bounds are satisfied. The same is true for Theorem 3.4, with the six possible orderings of the three nested loops underlying a matrix product. In the rest of this paper we will implicitly use this equivalence of different orderings.

3.3. LU factorization and linear systems. In the following three theorems we give probabilistic backward error bounds for triangular systems, LU factorization, and general linear systems.

THEOREM 3.5 (Solution of triangular systems). *Let the triangular system $Tx = b$, where $T \in \mathbb{R}^{n\times n}$ is nonsingular, be solved by substitution. Under Model 2.1, the computed solution $\widehat{x}$ satisfies*

$$(T + \Delta T)\widehat{x} = b, \quad |\Delta T| \le \widetilde{\gamma}_n(\lambda)|T|, \tag{3.13}$$

with probability at least $Q(\lambda, n(n+1)/2)$.

Proof. Assuming, without loss of generality, that T is lower triangular, we have $x_i = \big(b_i - \sum_{j=1}^{i-1} t_{ij}x_j\big)/t_{ii}$, $i = 1\colon n$. By Theorem 3.2, for any i we have

$$t_{ii}\widehat{x}_i(1 + \mu_i) = b_i - \sum_{j=1}^{i-1} t_{ij}\widehat{x}_j(1 + \mu_j), \qquad |\mu_j| \le \widetilde{\gamma}_i(\lambda), \quad j = 1\colon i \tag{3.14}$$

with probability at least $Q(\lambda, i)$. The probability of (3.14) holding for all i is therefore at least

$$1 - \sum_{i=1}^{n}(1 - Q(\lambda, i)) = 1 - \sum_{i=1}^{n} i(1 - P(\lambda)) = Q(\lambda, n(n+1)/2).$$

The result follows by weakening the bounds on $|\mu_j|$ to $\widetilde{\gamma}_n(\lambda)$. ∎

THEOREM 3.6 (LU factorization). *If Gaussian elimination applied to $A \in \mathbb{R}^{n\times n}$ runs to completion then under Model 2.1 the computed LU factors $\widehat{L}$ and $\widehat{U}$ satisfy*

$$A + \Delta A = \widehat{L}\widehat{U}, \quad |\Delta A| \le \widetilde{\gamma}_n(\lambda)|\widehat{L}||\widehat{U}| \tag{3.15}$$

with probability at least $Q(\lambda, n^3/3 + n^2/2 + n/6)$.

Proof. The Doolittle form of Gaussian elimination [14, secs. 2.2, 2.3] gives the following recurrences for the LU factors:

$$\left.\begin{aligned} u_{kj} &= a_{kj} - \sum_{i=1}^{k-1} l_{ki}u_{ij}, && j = k\colon n \\ l_{ik} &= \Bigl(a_{ik} - \sum_{j=1}^{k-1} l_{ij}u_{jk}\Bigr)/u_{kk}, && i = k+1\colon n \end{aligned}\right\} \quad k = 1\colon n.$$

We apply Theorem 3.2 to each of these n^2 equations. This readily gives (3.15) with probability at least

$$\begin{aligned} &1 - \sum_{k=1}^{n}\bigl[(n-k+1)(1-Q(\lambda,k)) + (n-k)(1-Q(\lambda,k))\bigr] \\ &= 1 - \sum_{k=1}^{n}\bigl[(2n+1-2k)(1-Q(\lambda,k))\bigr] \\ &= 1 - \sum_{k=1}^{n}\bigl[(2n+1)k - 2k^2\bigr](1-P(\lambda)) \\ &= 1 - \bigl[(2n+1)n(n+1)/2 - 2n(n+1)(2n+1)/6\bigr](1-P(\lambda)) \\ &= Q(\lambda, n^3/3 + n^2/2 + n/6). \end{aligned}$$

∎

THEOREM 3.7 (Linear system). *Let $A \in \mathbb{R}^{n\times n}$ and suppose Gaussian elimination produces a computed solution $\widehat{x}$ to $Ax = b$. Under Model 2.1,*

$$(A + \Delta A)\widehat{x} = b, \quad |\Delta A| \le \bigl(3\widetilde{\gamma}_n(\lambda) + \widetilde{\gamma}_n(\lambda)^2\bigr)|\widehat{L}||\widehat{U}|, \tag{3.16}$$

holds with probability at least $Q(\lambda, n^3/3 + 3n^2/2 + 7n/6)$.

Proof. From Theorem 3.6, $\widehat{L}\widehat{U} = A + \Delta A_1$, where $|\Delta A_1| \le \widetilde{\gamma}_n(\lambda)|\widehat{L}||\widehat{U}|$ with probability at least $Q(\lambda, n^3/3 + n^2/2 + n/6)$. By Theorem 3.5, the triangular solves produce $\widehat{y}$ and $\widehat{x}$ satisfying

$$\begin{aligned} (\widehat{L} + \Delta L)\widehat{y} = b, \quad |\Delta L| \le \widetilde{\gamma}_n(\lambda)|\widehat{L}|, \\ (\widehat{U} + \Delta U)\widehat{x} = \widehat{y}, \quad |\Delta U| \le \widetilde{\gamma}_n(\lambda)|\widehat{U}|, \end{aligned}$$

each inequality holding with probability at least $Q(\lambda, n(n+1)/2)$. Thus

$$\begin{aligned} b &= (\widehat{L} + \Delta L)(\widehat{U} + \Delta U)\widehat{x} \\ &= (A + \Delta A_1 + \widehat{L}\Delta U + \Delta L\widehat{U} + \Delta L\Delta U)\widehat{x} \\ &= (A + \Delta A)\widehat{x}, \end{aligned}$$

where

$$|\Delta A| \le \bigl(3\widetilde{\gamma}_n(\lambda) + \widetilde{\gamma}_n(\lambda)^2\bigr)|\widehat{L}||\widehat{U}|$$

holds with probability at least one minus the probability that one of the bounds for ΔA_1, ΔL, and ΔU is violated, namely

$$1 - (1 - Q(\lambda, n^3/3 + n^2/2 + n/6)) - 2(1 - Q(\lambda, n(n+1)/2)),$$

which yields the result. ∎

3.4. Cholesky factorization. We also give a result for Cholesky factorization, because symmetry brings an improvement in the probability compared with LU factorization.

THEOREM 3.8. *If Cholesky factorization applied to the symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ runs to completion then the computed factor $\widehat{R}$ satisfies*

$$\widehat{R}^T\widehat{R} = A + \Delta A, \quad |\Delta A| \le \widetilde{\gamma}_{n+1}(\lambda)|\widehat{R}^T||\widehat{R}| \tag{3.17}$$

with probability at least $Q(\lambda, n^3/6 + n^2/2 + n/3)$.

Proof. The recurrences for R can be written

$$\left.\begin{aligned} r_{ij} &= \Bigl(a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj}\Bigr)/r_{ii}, \quad i = 1\colon j-1 \\ r_{jj} &= \Bigl(a_{jj} - \sum_{k=1}^{j-1} r_{kj}^2\Bigr)^{1/2} \end{aligned}\right\} \; j = 1\colon n.$$

We apply Theorem 3.2 to each of the $n(n-1)/2$ equations for the r_{ij} with $i < j$. For r_{jj} we need an analogue of Theorem 3.2 in which a square root replaces the division (cf. [14, Prob. 10.4]): the constant is $\widetilde{\gamma}_{k+1}(\lambda)$ and the probability remains $Q(\lambda, k)$. We readily find that (3.17) holds with probability at least

$$\begin{aligned} &1 - \sum_{j=1}^{n}\left[\left(\sum_{i=1}^{j-1} 1 - Q(\lambda, i)\right) + 1 - Q(\lambda, j)\right] \\ &= 1 - \sum_{j=1}^{n}(j^2/2 + j/2)(1 - P(\lambda)) \\ &= 1 - \bigl(n(n+1)(2n+1)/12 + n(n+1)/4\bigr)(1 - P(\lambda)) \\ &= Q(\lambda, n^3/6 + n^2/2 + n/3). \end{aligned}$$

∎

The interesting feature of Theorem 3.8 is that the argument $f(n)$ in the probability $Q(\lambda, f(n))$ for Cholesky factorization is about half that for LU factorization. The reason is that there are half as many invocations of Theorem 3.2, which is essentially because there are half as many flops, and hence half as many probabilistic events that need to hold. In fact, for all the algorithms analyzed in this section, the argument $f(n)$ in the probability is approximately half the flop count, that is, the bounds hold with probability approximately $Q(\lambda, \text{flops}/2)$; this is a direct consequence of Theorem 3.1: an inner product requires $2n - 1$ flops and the error bound holds with probability $Q(\lambda, n)$.

3.5. Assessment of constant and probabilities. Two features of Theorems 3.1–3.8 deserve comment. First, the constant $\widetilde{\gamma}_n(\lambda)$ passes through essentially unchanged from Theorem 2.4, maintaining the $\sqrt{n}u$ proportionality of the bounds. By comparison, the corresponding standard deterministic bounds are proportional to nu [14].

Table 3.1: Values of $Q(\lambda, n^3/3)$ for several n and λ, for single precision arithmetic. The results are shown to 5 significant figures.

λ	10^6	10^7	10^8	10^9	10^{10}
10.0	9.9987e−01	8.7142e−01	−1.2758e+02	−1.2858e+05	−1.2858e+08
10.5	1.0000e+00	9.9924e−01	2.3542e−01	−7.6358e+02	−7.6458e+05
11.0	1.0000e+00	1.0000e+00	9.9646e−01	−2.5407e+00	−3.5397e+03
11.5	1.0000e+00	1.0000e+00	9.9999e−01	9.8723e−01	−1.1770e+01
12.0	1.0000e+00	1.0000e+00	1.0000e+00	9.9996e−01	9.6413e−01
12.5	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00	9.9992e−01
13.0	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00

The second important feature is that the overall probability $Q(\lambda, f(n))$ depends on the problem dimensions, where $f(n)$ is as large as $n^3/3 + 3n^2/2 + 7n/6$ for the solution of $Ax = b$ by LU factorization. An important question is thus how fast λ must increase in order that $Q(\lambda, cn^3)$, where c is a constant as in our bounds, stays independent of n. Write $P(\lambda)$ as $P(\lambda) = 1 - \varepsilon$, where $\varepsilon = 2\exp(-\lambda^2(1-u)^2/2)$. Then, recalling the definition (3.1), we have $Q(\lambda, cn^3) = 1 - cn^3\varepsilon$. Thus $\varepsilon = O(1/n^3)$ is enough to ensure that $Q(\lambda, cn^3)$ is bounded below independent of n. From the expression of ε, it is easy to check that λ must therefore increase proportionally to $\sqrt{\log n}$, which represents an extremely slow increase. The dependence on n of $Q(\lambda, f(n))$ therefore does not cause any serious deterioration in the probabilities as long as we increase λ a little.

Table 3.1 shows values of $Q(\lambda, n^3/3)$ for double precision arithmetic, a range of n up to 10^{10}, and $\lambda = 10\colon 13$. In order to avoid cancellation affecting the results the values are computed at 100-digit precision using the Multiprecision Computing Toolbox [23] and then rounded to the accuracy shown. For $\lambda = 13$, we have a probability of 1.0000 for $n \le 10^{10}$ (the same is true for half precision, which is not shown in the table); this value of λ therefore suffices for the largest dense linear systems that will be solved on exascale computers. Furthermore, as we show in the next section, the probability $Q(\lambda, f(n))$ is actually very pessimistic and in practice the bounds hold with much smaller values of λ.

4. Numerical experiments. We now present a set of numerical experiments designed to give insight into our probabilistic bounds. The experiments have three main aims: to test the sharpness of the bounds and the probabilities; to compare the rate of growth of the error with n with that of the probabilistic ($\gamma_n(\lambda) \approx \sqrt{n}u$) and deterministic ($\gamma_n \approx nu$) bounds; and to check whether the probabilistic bounds are applicable at all, since they were derived under Model 2.1.

These experiments are carried out with MATLAB R2018b. Computations are performed in single precision, except in section 4.1.2, where half (fp16) and quarter (fp8) precisions are used, and section 4.5, where double precision is used. Half precision (corresponding to the IEEE standard) and quarter precision (not standard, and as suggested in [22]) are simulated with the rounding function `chop.m` from [15]. The "exact" quantities appearing in the backward error formulas for inner products and matrix-vector products are computed in double precision.

In sections 4.1–4.4 we use randomly generated matrices and vectors. We compare different types of distributions, in particular

- random uniform $[0, 1]$: `rand(m,n)`;

- random uniform $[-1, 1]$: `(rand(m,n)-0.5)*2`;
- random constant: `rand*ones(m,n)`.

In order to make the experiments reproducible, we use `rng(1)` to seed the random number generator at the beginning of each script generating a figure of this section. We have made these scripts available online[2]. For each size of problem n, we run the same experiment N_{test} times and plot the maximum and mean backward errors, denoted by ε_{bwd}^{max} and ε_{bwd}^{mean}, respectively. We have set $N_{test} = 100$ for Figures 4.1, 4.2, and 4.4a, and $N_{test} = 10$ for Figures 4.6 and 4.7. We compare these backward errors with their deterministic and probabilistic bounds γ_n and $\widetilde{\gamma}_n(\lambda)$. Throughout this section, the probabilistic bounds are plotted taking $\lambda = 1$. In Section 4.5 we report numerical results on a large set of real-life matrices coming from a variety of applications, taken from the SuiteSparse collection [8].

The inner product and matrix–vector product computations were implemented in MATLAB using loops such as, for $x = a^T b$,

```
x = 0;
for i=1:n
  x = x + a(i)*b(i);
end
```

This code corresponds to our analysis, where every floating-point operation involves a rounding, so that the model (1.2) of floating-point arithmetic is applicable throughout. If we compute the inner product as `x = a'*b` then, depending on the details of how the underlying BLAS operation is coded and optimized, the sum could be accumulated using fan-in (a binary tree) or with extra precision for intermediate quantities, both of which can make the constants in our analysis and the traditional deterministic bounds pessimistic by a factor up to n [6].

4.1. Inner products. We first report numerical experiments with inner products $y = a^T b$. We record the backward error of the computed $\widehat{y}$,

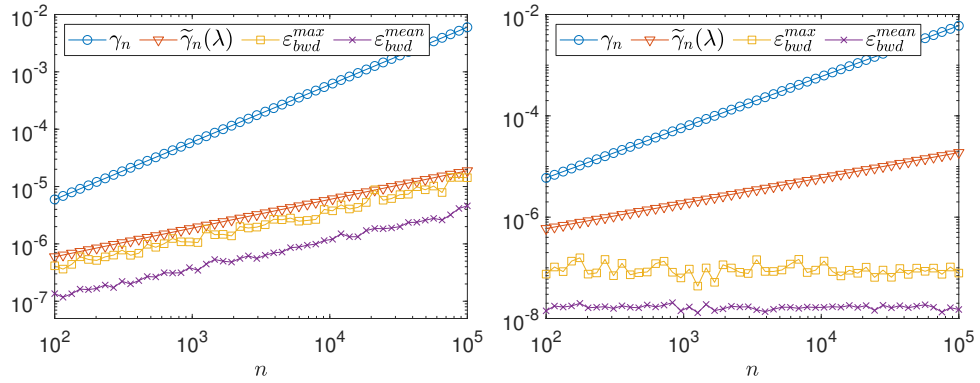
$$\varepsilon_{bwd} = \min\{\, \varepsilon \geq 0 : \widehat{y} = (a + \Delta a)^T b,\ |\Delta a| \leq \varepsilon |a| \,\} = \frac{|\widehat{y} - y|}{|a|^T |b|}, \tag{4.1}$$

the latter equality being a special case of (4.3) below. We compare the backward error with our probabilistic bound $\widetilde{\gamma}_n(\lambda)$ from Theorem 3.1 and the deterministic bound γ_n [14, Eq. (3.4)].

4.1.1. Random uniform vectors. We consider the case where the vectors a and b have random entries from the uniform $[0, 1]$ or uniform $[-1, 1]$ distributions. The results, plotted in Figure 4.1, show that the probabilistic bound, plotted with $\lambda = 1$, is in much better agreement with the actual backward error than the deterministic one and is sharp for the $[0, 1]$ data. In theory, the probabilistic bound can only be guaranteed to hold with high probability for $\lambda \gtrsim 6$. Nevertheless, out of 5000 runs (50 different problem sizes times $N_{test} = 100$), the bound holds with $\lambda = 1$ in all cases except one (which would require a slightly higher value $\lambda = 1.02$) in Figure 4.1a and in all cases in Figure 4.1b. Therefore, the probability $Q(\lambda, n)$ is extremely pessimistic.

However, and perhaps most importantly, in the case of vectors with entries on $[0, 1]$ (Figure 4.1a) the probabilistic bound successfully captures the asymptotic behavior of the error growth, which follows $\sqrt{n}$ rather than n. Interestingly, this is not the case for vectors with entries on $[-1, 1]$ (Figure 4.1b). Since our theory does not assume the data to follow any specific random distribution, and since the probabilistic bound

[2] https://gitlab.com/theo.andreas.mary/proberranalysis



(a) Random uniform $[0, 1]$ vectors.

(b) Random uniform $[-1, 1]$ vectors.

Fig. 4.1: Backward error and its bounds for the computation in single precision of the inner product $y = a^T b$, for vectors a and b with random uniform entries. Here, $N_{test} = 100$ and $\lambda = 1$.

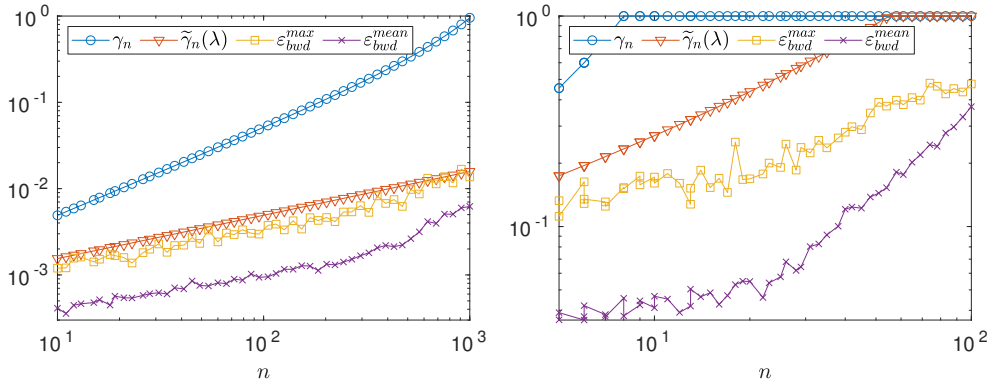
is sharp for the $[0, 1]$ data, we conclude that it cannot be further improved without additional assumptions.

4.1.2. Lower precision arithmetics. Now we repeat the experiment from section 4.1.1 with uniform $[0, 1]$ vectors a and b but using precisions lower than single to compute the inner product $y = a^T b$. We report the results using fp16 ($u = 2^{-11}$) and fp8 ($u = 2^{-4}$) arithmetics in Figures 4.2a and 4.2b, respectively. The results lead to the same conclusions as when using single precision. Importantly, the deterministic bound is unable to guarantee even a single digit of accuracy when $n \geq 10^3$ in fp16, and yet the error is only of order 10^{-2}. Our probabilistic bound is able to successfully explain and predict this behavior. This effect is even clearer with fp8 arithmetic.

4.2. Two cases where Model 2.1 is invalid. In this section, we present two cases where Model 2.1 is invalid and the probabilistic bound does not hold. In the first case the rounding errors have nonzero mean and in the second case they are dependent.

4.2.1. Very large nonnegative vectors: rounding errors have nonzero mean. We consider the inner product of two vectors a and b of very large size $n = 10^8$. Their entries are sampled from the uniform $[0, 1]$ distribution and are thus nonnegative. In Figure 4.3a, we plot the backward error and its bounds at each step i of the algorithm. As expected and previously analyzed (see Figure 4.1a), the error is in good agreement with the probabilistic bound (here, with $\lambda = 1$) for moderate values of i. However, for large values of i (starting at around 10^6), the error starts increasing rapidly and violates the probabilistic bound. Increasing λ is not sufficient in this case, because for $i \geq 10^6$ the error increases at a faster rate than the bound. Model 2.1 is thus clearly invalid in this case.

The explanation is that for large nonnegative vectors the value of the inner product $y = a^T b$ is large, whereas each increment $a_i b_i$ is (potentially much) smaller and certainly bounded by 1. Let y_i be the partial sum $\sum_{k=1}^{i-1} a_k b_k$, so that $y_{i+1} = y_i + a_i b_i$.



(a) fp16 (half precision) arithmetic.

(b) fp8 (quarter precision) arithmetic.

Fig. 4.2: Backward error and its bounds for the computation of the inner product $a^T b$ for random uniform $[0,1]$ vectors in lower precision arithmetic. Here, $N_{test} = 100$ and $\lambda = 1$. γ_n and $\widetilde{\gamma}_n$ are plotted as 1 when their value exceeds 1.

The computed $\widehat{y}_{i+1}$ satisfies, by (1.2),

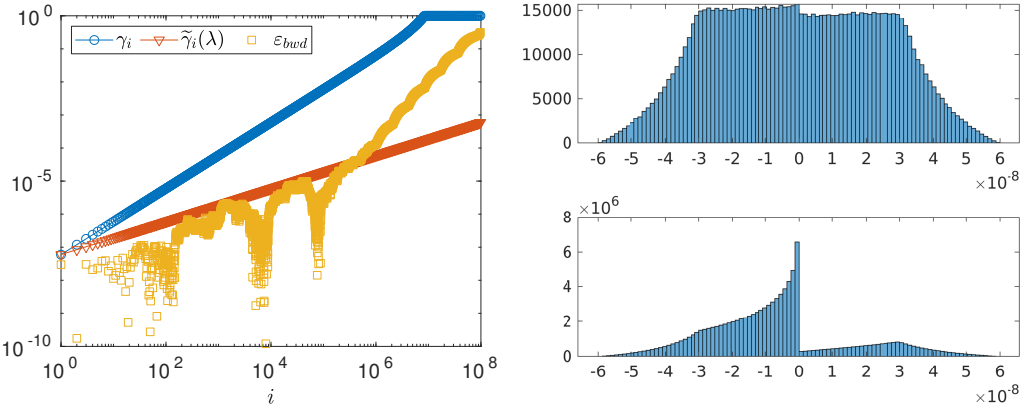
$$\widehat{y}_{i+1} = \mathrm{fl}(\widehat{y}_i + \mathrm{fl}(a_i b_i)) = (\widehat{y}_i + \mathrm{fl}(a_i b_i))(1+\delta_i). \tag{4.2}$$

If n is very large, then at some point y_i will become so large that incrementing it by $\mathrm{fl}(a_i b_i)$ will not change its computed value, that is, $\widehat{y}_{i+1} = \widehat{y}_i$. Specifically, let q be the integer such that $2^{q-1} \le y_i < 2^q$; the spacing between each floating point number in this interval is $2^q u$. Therefore, if $\mathrm{fl}(a_i b_i) < 2^{q-1}u$, we have $\widehat{y}_{i+1} = \widehat{y}_i$ and so by (4.2) $\widehat{y}_i \delta_i + \mathrm{fl}(a_i b_i)(1+\delta_i) = 0$, that is, $\delta_i = -\mathrm{fl}(a_i b_i)/(\widehat{y}_i + \mathrm{fl}(a_i b_i)) < 0$. As i and thus $\widehat{y}_i$ and q increase, the probability that $\mathrm{fl}(a_i b_i) < 2^{q-1}u$ also increases. It is thus clear that as i increases the mean of the errors δ_i will deviate from zero, which violates the assumption made by Model 2.1. This is illustrated in Figure 4.3b, which shows a histogram of the values of δ_i. For $i \le 10^6$ the δ_i have a distribution of mean approximately zero, but for $10^6 \le i \le 10^8$ the mean is significantly smaller than zero.

We note that in half precision this issue arises as soon as $n \approx 10^4$ (cf. Figure 4.2a).

4.2.2. Constant vectors: rounding errors not independent. We now perform the same experiment as in section 4.1.1 but with vectors a and b for which $a_i = \alpha$, $b_i = \beta$, $i = 1\colon n$, where α and β are from the uniform $[0,1]$ distribution. This experiment leads to a very different result. Indeed, as shown in Figure 4.4a, the probabilistic bound does not bound the error. Indeed the probabilistic bound has a slower asymptotic growth than the error, which is unaffected by increasing λ (which has only a constant effect on a logarithmic scale), so clearly Model 2.1 is not valid for this constant data.

The explanation is that, in this case, the δ_i in the proof of Theorem 3.1 are not independent and thus Model 2.1 is violated. In fact, all computed quantities $\widehat{s}_i$ that lie between the same consecutive powers of two produce the same rounding error. As illustrated in Figure 4.5, since the spacing between floating-point numbers is constant between consecutive powers of two, and since the increments $a_i b_i = \alpha\beta$ are also constant, the errors in the sums $\widehat{s}_i = fl(\widehat{s}_{i-1} + fl(\alpha\beta))$ must be constant as well.



(a) Backward error and its bounds at step i.

(b) Distribution of δ_i for $i = 1 : 10^6$ (top) and $i = 10^6 : 10^8$ (bottom).

Fig. 4.3: Computation of the inner product of two vectors a and b of size $n = 10^8$ from the uniform $[0, 1]$ distribution. Here, $\lambda = 1$ and γ_i is plotted as 1 when $iu > 1$.

This phenomenon is illustrated in Figure 4.4b, for $n = 10^4$. Each time $\widehat{s}_i$ crosses a power of two, the rounding error being accumulated changes. If it remains of the same sign, the overall error keeps increasing (this is what happens, e.g., around $i \approx 850$ and $i \approx 3400$), whereas if it switches sign, the overall error first decreases until it crosses the exact result and starts increasing again (this is what happens, e.g., around $i \approx 1700$ and $i \approx 6800$).

We mention that the phenomenon of successive rounding errors reinforcing rather than cancelling was observed and analyzed by Huskey and Hartree [18] in the numerical solution of differential equations on the ENIAC .

4.3. Matrix–vector products. Next, we consider matrix–vector products $y = Ax$, where $A \in \mathbb{R}^{n\times n}$. We compute the backward error

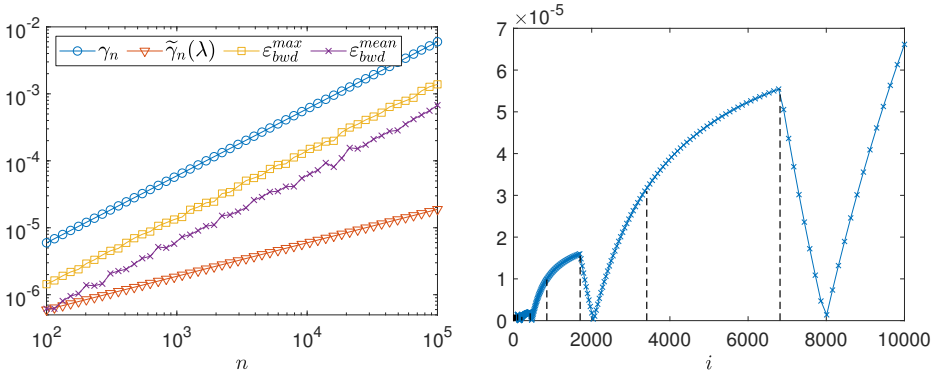
$$\varepsilon_{bwd} = \min\{\, \varepsilon \geq 0 : \widehat{y} = (A + \Delta A)x,\ |\Delta A| \leq \varepsilon |A| \,\} = \max_i \frac{|\widehat{y} - y|_i}{(|A||x|)_i}, \tag{4.3}$$

where the latter formula follows from the Oettli–Prager theorem [14, Thm. 7.3], [24].

In Figure 4.6 we compare the backward error with the probabilistic bound given by Theorem 3.3, with constant $\widetilde{\gamma}_n(\lambda)$ and the deterministic bound with constant γ_n [14, sec. 3.5]. Similarly to the case for inner products, the probabilistic bound can be guaranteed to hold with high probability only for $\lambda \gtrsim 7$, yet $\lambda = 1$ leads to only 63 cases out of 500 violating the bound and for $\lambda = 1.24$ the bound holds in every case. Even though the probability is pessimistic, the bound $\widetilde{\gamma}_n(\lambda)$ itself is quite sharp in the $[0, 1]$ case and successfully predicts the error growing proportionally to $\sqrt{n}$.

4.4. Linear systems. We now consider the solution of linear systems $Ax = b$ via LU factorization. We compute the backward error

$$\begin{aligned} \varepsilon_{bwd} &= \min\{\, \varepsilon \geq 0 : (A + \Delta A)\widehat{x} = b + \Delta b,\ |\Delta A| \leq \varepsilon |\widehat{L}||\widehat{U}| \,\} \\ &= \max_i \frac{|A\widehat{x} - b|_i}{(|\widehat{L}||\widehat{U}||\widehat{x}|)_i}, \end{aligned} \tag{4.4}$$



(a) (b)

Fig. 4.4: (a): backward error and its bounds for the inner product $s_n = a^T b$ of two random vectors a and b with constant entries from the uniform $[0, 1]$ distribution. Here, $N_{test} = 100$ and $\lambda = 1$. (b): Backward error at step i of the computation of s_n with $n = 10^4$. Vertical dashed lines indicate the values of i for which $\widehat{s}_i$ crosses a power of two.

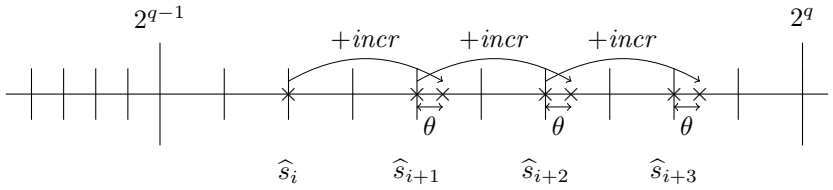
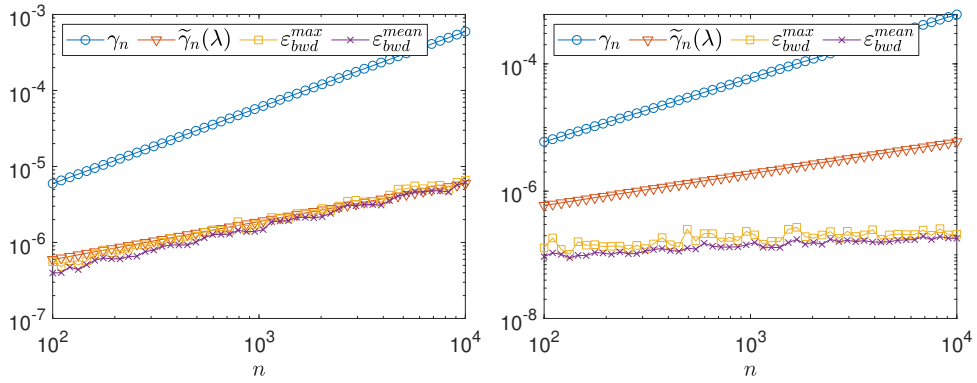
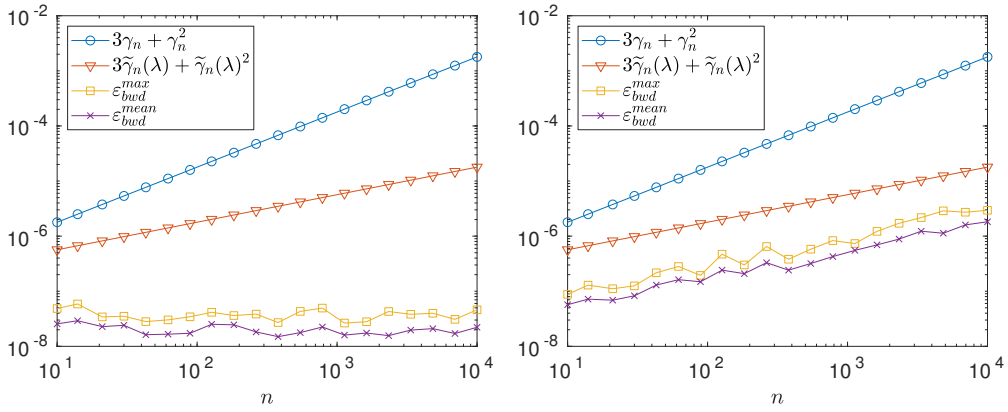


Fig. 4.5: Illustration of the fact that the error $\theta_i = s_i + incr - \widehat{s}_{i+1}$ is constant between consecutive powers of two when the increment $\mathrm{fl}(a_i b_i) = incr$ is constant (here, $\widehat{s}_i$ is the first instance of the sum in the interval $[2^{q-1}, 2^q]$).



(a) Random uniform A and x from $[0, 1]$. (b) Random uniform A and x from $[-1, 1]$.

Fig. 4.6: Backward error and its bounds for the matrix–vector product $y = Ax$, with a matrix A and vector x with random uniform entries. Here, $N_{test} = 10$ and $\lambda = 1$.



(a) Random uniform A and b from $[0, 1]$.

(b) Random uniform A and x from $[0, 1]$.

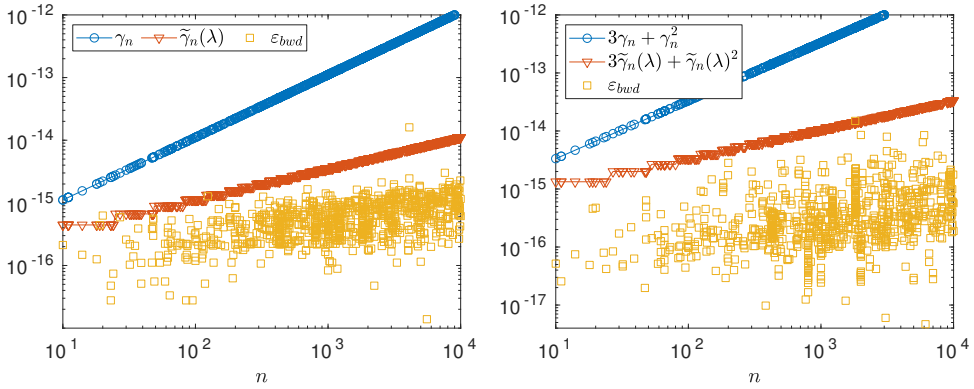
Fig. 4.7: Backward error and its bounds for the solution of linear systems $Ax = b$, where A and x have with random uniform entries. Here, $N_{test} = 10$ and $\lambda = 1$.

where the latter formula is obtained from the Oettli–Prager theorem [14, Thm. 7.3], [24]. Here we are measuring the backward error relative to $|\widehat{L}||\widehat{U}|$ instead of $|A|$, consistently with Theorem 3.7, but also in order to avoid any instability (necessarily corresponding to large element growth) affecting the interpretation of the results.

In Figure 4.7 we compare the backward error with its bound from Theorem 3.7. The case of linear systems is quite different from the other kernels analyzed in the previous sections because the LU factors of a nonnegative matrix may have entries of both positive and negative signs, and thus even if A and b have $[0, 1]$ entries x usually does not. For this reason, for A and b randomly generated with uniform entries, the error is similar regardless of whether the entries are in $[0, 1]$ (see Figure 4.7a) or $[-1, 1]$ (not shown), and does not grow with n. In this case, the probabilistic bound is pessimistic, albeit substantially smaller than the deterministic bound.

If instead we generate the solution vector x to have random uniform entries in $[0, 1]$, then the error grows much more rapidly and the probabilistic bound becomes almost sharp (see Figure 4.7b). Therefore, even for linear systems, there exists some data sets for which the probabilistic bound is almost sharp under the assumptions of Model 2.1.

4.5. Numerical experiments on real-life matrices. In real-life applications matrices are not usually randomly generated. It is therefore important to check whether the probabilistic bounds that we have shown in the previous subsections can give good predictions for random matrices can also do so for matrices coming from various applications. In this section we perform numerical experiments with linear systems with matrices A from the SuiteSparse collection [8]. We selected all the square matrices in the collection with $10 \leq n \leq 10^4$; this corresponds to a set of 1164 matrices. The right-hand side b is also provided for 309 of these matrices; for the remaining cases, we generate it randomly with $[0, 1]$ uniform entries. Since many of these matrices are ill conditioned we perform computations in double rather than single precision, so as to minimize the number of matrices that are singular to the working precision.



(a) LU factorization (1139 matrices).

(b) Solution of $Ax = b$ (944 matrices).

Fig. 4.8: Backward error and its bounds for LU factorization and the solution of a linear system $Ax = b$ for a set of matrices from the SuiteSparse collection sorted by size. Here, $\lambda = 1$.

We first compute the LU factorization of A and measure the backward error

$$\varepsilon_{bwd} = \max_i \frac{|A - \widehat{L}\widehat{U}|_i}{(|\widehat{L}||\widehat{U}|)_i}.$$

For 25 of these real-life matrices, the backward error violates the deterministic, worst-case bound γ_n due to a large amount of underflow, which is not included in our standard floating-point model (1.2). We filter these matrices out and plot the backward errors for the LU factorization for the remaining 1139 matrices in Figure 4.8a. Then we use the computed LU factors to solve the system $Ax = b$. For an additional 195 matrices, we are unable to compute a solution because the U factor is singular. We are left with 944 matrices, for which we plot the linear system backward errors in Figure 4.8b.

The probabilistic bound holds with $\lambda = 1$ for all but two matrices for LU factorization and one matrix for the solution of $Ax = b$. Increasing λ to 2.26 makes it hold for every matrix. The probabilistic bound often exceeds the actual error by still a few orders of magnitude but is closer to the actual backward error than the deterministic bound by several orders of magnitude.

4.6. Discussion. From the numerical experiments presented in this section, we can draw the following overall conclusions.

- The values of $Q(\lambda, f(n))$ that provide a lower bound on the probability of our bounds holding are extremely pessimistic. The probabilistic bounds hold in all cases with a value of λ smaller than 2.26. While our analysis in section 3.5 suggests that λ should increase asymptotically like $\sqrt{\log n}$, in order to keep the probability bound independent of n, we have not detected any such effect in our experiments.
- The constant $\widetilde{\gamma}_n(\lambda)$ in the probabilistic bounds can be sharp, both in the sense of yielding sharp bounds (as in Figure 4.6a) and, more importantly, in correctly predicting that the error grows like $\sqrt{n}$. Therefore without any as-

sumption on the matrices and vectors, or further assumptions on the rounding errors, the probabilistic bounds cannot be improved upon.

- Overall, the experiments show that our probabilistic Model 2.1 can give useful predictions of the backward error for both random matrices and matrices from real-life applications. The examples in sections 4.2.1 and 4.2.2, however, reveal two situations in which the assumptions in the model do not hold and the probabilistic bound can then be violated.
- For matrices and vectors with elements from the uniform $[-1, 1]$ distribution we have observed the backward errors to be much smaller than for the uniform $[0, 1]$ case, and to grow little, or not at all, with n. Even the probabilistic bounds are pessimistic in this case. Explaining the difference between the $[0, 1]$ and $[-1, 1]$ cases is an interesting question for future research.

5. Conclusions. We have shown that under the assumption that rounding errors are independent and of zero mean, probabilistic backward error bounds for several key matrix and vector operations can be obtained that have exactly the same form as classic deterministic bounds, but are of order $\lambda\sqrt{n}u$ instead of nu, for a constant λ. Strictly, λ should be proportional to $\sqrt{\log n}$ in order to keep the probabilities from decaying with n, but this term is less than 5 for dense linear systems that can currently be solved, that is, for $n \leq 10^8$.

The new bounds hold with a probability bounded below by a quantity $Q(\lambda, f(n))$, and $\lambda = 12$ suffices to give a probability within 10^{-7} of 1 for $n \leq 10^8$ for matrix factorizations. Even this value is pessimistic, as throughout our tests the probabilistic bounds held with $\lambda = 2.26$, except in two tests where the model is not applicable.

Our analysis therefore provides the first rigorous justification of the rule of thumb that one can take the square root of the constant in a deterministic error bound to obtain a more realistic bound that takes account of statistical effects in rounding error propagation. Moreover, our results apply for any n, not just for sufficiently large n, as would be the case for results based on the central limit theorem.

A key question underlying any probabilistic analysis is whether the results it produces are useful for error estimation and for predicting the asymptotic rate of error growth with problem dimension. Our experiments show that the probabilistic bounds are indeed useful for both random matrices and real-life matrices from the SuiteSparse collection. However, we identified two situations involving inner products in which the underlying assumptions are not valid and the bounds do not apply: large nonnegative vectors, for which the rounding errors eventually have nonzero mean, and constant vectors for which the rounding errors are dependent. Clearly, care is required in applying and interpreting the probabilistic error bounds.

In future work we will explore further applications of our probabilistic analysis. We will also modify our analysis to allow for inner products and matrix–vector and matrix-matrix multiplications being implemented in optimized forms that use extra precision internally, as is often the case in modern processors.

Acknowledgements. We thank Pierre Blanchard, Mike Giles, Des Higham, and Ilse Ipsen for helpful discussions and the referees for insightful suggestions.

REFERENCES

[1] J. L. Barlow and E. H. Bareiss, *Probabilistic error analysis of Gaussian elimination in floating point and logarithmic arithmetic*, Computing, 34 (1985), pp. 349–364, https://doi.org/10.1007/BF02251834.

[2] P. Billingsley, *Probability and Measure*, Wiley, New York, third ed., 1995.
[3] S. Boucheron, G. Lugosi, and O. Bousquet, *Concentration inequalities*, in Advanced Lectures on Machine Learning, O. Bousquet, U. von Luxburg, and G. Rätsch, eds., Springer-Verlag, Berlin, 2004, pp. 208–240, https://doi.org/10.1007/978-3-540-28650-9_9.
[4] D. Calvetti, *A stochastic roundoff error analysis for the fast Fourier transform*, Math. Comp., 56 (1991), pp. 755–755, https://doi.org/10.1090/s0025-5718-1991-1068824-0.
[5] E. Carson and N. J. Higham, *Accelerating the solution of linear systems by iterative refinement in three precisions*, SIAM J. Sci. Comput., 40 (2018), pp. A817–A847, https://doi.org/10.1137/17M1140819.
[6] A. M. Castaldo, R. C. Whaley, and A. T. Chronopoulos, *Reducing floating point error in dot product using the superblock family of algorithms*, SIAM J. Sci. Comput., 31 (2008), pp. 1156–1174, https://doi.org/10.1137/070679946.
[7] F. Chatelin and M.-C. Brunet, *A probabilistic round-off error propagation model. Application to the eigenvalue problem*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, 1990, pp. 139–160.
[8] T. A. Davis and Y. Hu, *The University of Florida Sparse Matrix Collection*, ACM Trans. Math. Software, 38 (2011), pp. 1:1–1:25, https://doi.org/10.1145/2049662.2049663, http://doi.acm.org/10.1145/2049662.2049663.
[9] A. Haidar, A. Abdelfattah, M. Zounon, P. Wu, S. Pranesh, S. Tomov, and J. Dongarra, *The design of fast and energy-efficient linear solvers: On the potential of half-precision arithmetic and iterative refinement techniques*, in Computational Science—ICCS 2018, Y. Shi, H. Fu, Y. Tian, V. V. Krzhizhanovskaya, M. H. Lees, J. Dongarra, and P. M. A. Sloot, eds., Springer International Publishing, Cham, 2018, pp. 586–600, https://doi.org/10.1007/978-3-319-93698-7_45.
[10] A. Haidar, S. Tomov, J. Dongarra, and N. J. Higham, *Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC '18 (Dallas, TX), Piscataway, NJ, USA, 2018, IEEE Press, pp. 47:1–47:11, http://dl.acm.org/citation.cfm?id=3291656.3291719.
[11] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.
[12] P. Henrici, *Elements of Numerical Analysis*, Wiley, New York, 1964.
[13] P. Henrici, *Test of probabilistic models for the propagation of roundoff errors*, Comm. ACM, 9 (1966), pp. 409–410, https://doi.org/10.1145/365696.365698.
[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second ed., 2002, https://doi.org/10.1137/1.9780898718027.
[15] N. J. Higham and S. Pranesh, *Simulating low precision floating-point arithmetic*, MIMS EPrint 2019.xx, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2019. In preparation.
[16] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30, https://doi.org/10.1080/01621459.1963.10500830.
[17] T. E. Hull and J. R. Swenson, *Tests of probabilistic models for propagation of roundoff errors*, Comm. ACM, 9 (1966), pp. 108–113, https://doi.org/10.1145/365170.365212.
[18] H. D. Huskey, *On the precision of a certain procedure of numerical integration*, J. Res. Nat. Bur. Standards, 42 (1949), pp. 57–62. With an appendix by Douglas R. Hartree.
[19] *IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2008 (revision of IEEE Std 754-1985)*, IEEE Computer Society, New York, 2008, https://doi.org/10.1109/IEEESTD.2008.4610935.
[20] C.-P. Jeannerod and S. M. Rump, *Improved error bounds for inner products in floating-point arithmetic*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 338–344, https://doi.org/10.1137/120894488.
[21] W. Kahan, *The improbability of probabilistic error analyses for numerical computations*. Manuscript, Mar. 1996, https://people.eecs.berkeley.edu/~wkahan/improber.pdf.
[22] C. B. Moler, *"Half precision" 16-bit floating point arithmetic*. http://blogs.mathworks.com/cleve/2017/05/08/half-precision-16-bit-floating-point-arithmetic/, May 2017.
[23] *Multiprecision Computing Toolbox*. Advanpix, Tokyo. http://www.advanpix.com.
[24] W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409, https://doi.org/10.1007/BF01386090.
[25] S. M. Rump and C.-P. Jeannerod, *Improved backward error bounds for LU and Cholesky factorizations*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 684–698, https://doi.org/10.

1137/130927231.

[26] M. Tienari, *A statistical model of roundoff error for varying length floating-point arithmetic*, BIT, 10 (1970), pp. 355–365, https://doi.org/10.1007/BF01934204.

[27] J. von Neumann and H. H. Goldstine, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099, https://doi.org/10.1090/S0002-9904-1947-08909-6.

[28] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330, https://doi.org/10.1145/321075.321076.

[29] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. Also published by Prentice-Hall, Englewood Cliffs, NJ, USA. Reprinted by Dover, New York, 1994.

[30] D. Williams, *Weighing the Odds. A Course in Probability and Statistics*, Cambridge University Press, Cambridge, UK, 2001.