Probabilistic Roundoff Error Analysis for Fundamental Matrix Computations

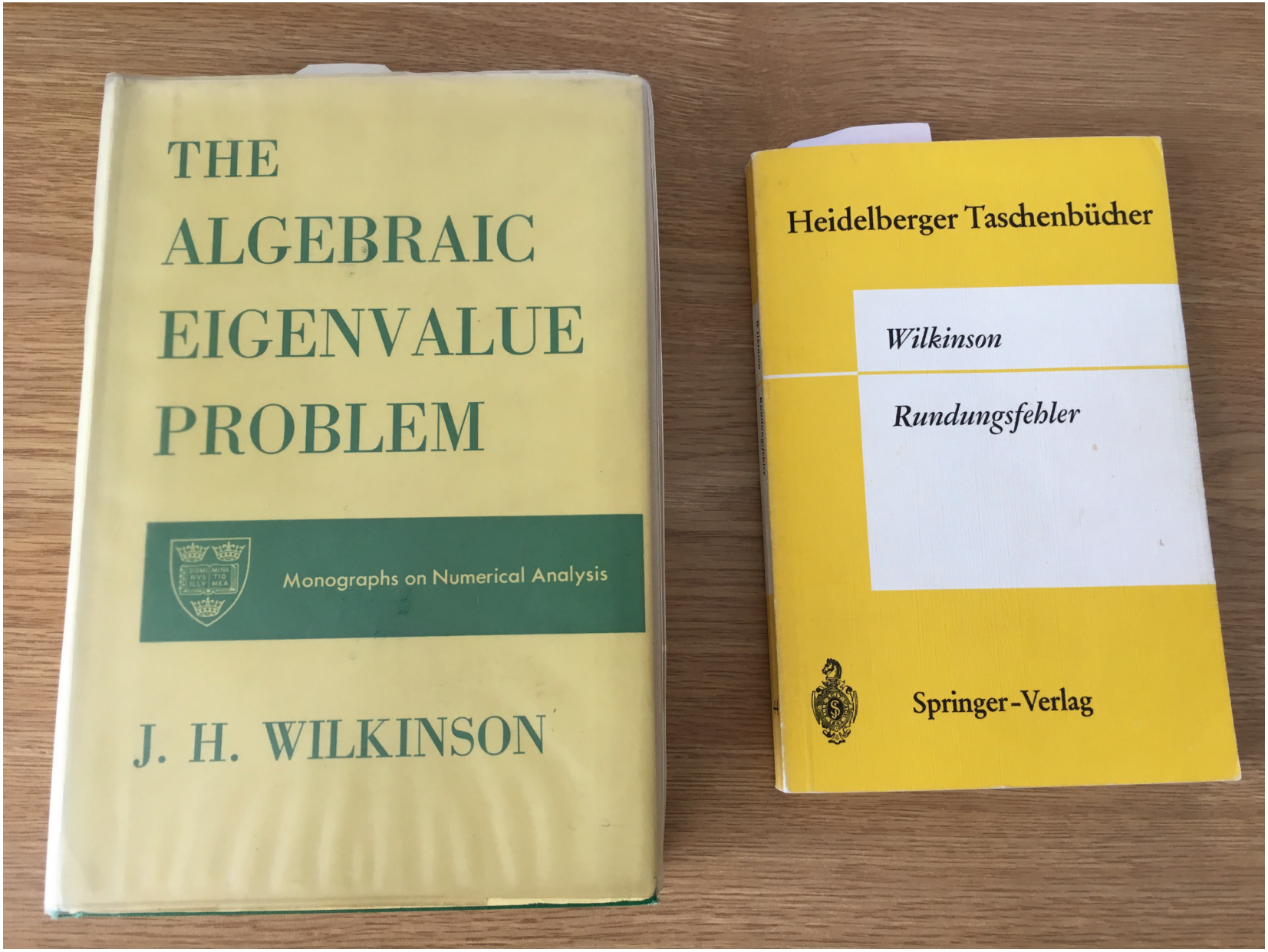
Ilse C.F. Ipsen

Joint work with: Hua Zhou (UCLA, Biostatistics)

North Carolina State University
Raleigh, NC, USA

Research supported in part by NSF DMS FRG, NSF DMS RTG, and EP/P020720/1 Manchester

My Wilkinson Collection



Wilkinson's Take

Definieren wir nun t_2 durch

$$1.06\times 2^{-2t}=2^{-2t_2}, \tag{32.6}$$

oder

$$2t_2=2t-\log_2 1.06, \tag{32.7}$$

so kann (32.2) durch die Ungleichung

$$|\varepsilon|<\tfrac{3}{2}r\times 2^{-2t_2} \tag{32.8}$$

ersetzt werden, wobei sich t_2 nur geringfügig von t unterscheidet.

In dieser Schreibweise entsteht aus (32.1) die Gleichung

$$\begin{aligned} &gl_2(a_1b_1+a_2b_2+\cdots+a_nb_n)-(a_1b_1+a_2b_2+\cdots+a_nb_n)(1+\varepsilon)\\ &\quad\equiv(a_1b_1\varepsilon_1+a_2b_2\varepsilon_2+\cdots+a_nb_n\varepsilon_n)(1+\varepsilon), \end{aligned} \tag{32.9}$$

wobei

$$\left.\begin{aligned} |\varepsilon|&\le 2^{-t}\\ |\varepsilon_1|&<\tfrac{3}{2}n\times 2^{-2t_2}\\ |\varepsilon_r|&<\tfrac{3}{2}(n+2-r)2^{-2t_2}. \end{aligned}\right\} \tag{32.10}$$

Wie man sich überlegt, ist der in (32.4) und (32.5) eingeführte Faktor 1.06 so groß, daß die Wirkung des Faktors $(1+\varepsilon)$ auf der rechten Seite von (32.9) durch die Abschätzung der ε_i mittels t_2 mit abgegolten ist. Daher erhalten wir schließlich

$$\begin{aligned} &gl_2(a_1b_1+a_2b_2+\cdots+a_nb_n)-(a_1b_1+a_2b_2+\cdots+a_nb_n)(1+\varepsilon)\\ &\quad\equiv(a_1b_1\varepsilon_1+a_2b_2\varepsilon_2+\cdots+a_nb_n\varepsilon_n), \end{aligned} \tag{32.11}$$

wobei die Abschätzungen (32.10) nach wie vor gelten. Für $gl_2(a_1+a_2+\ldots a_n)$ kann man die Gleichungen und Abschätzungen in ähnlicher Weise vereinfachen. Die Beziehungen (32.10) und (32.11) benutzen wir an mehreren Stellen des Buchs, ohne ausdrücklich auf die Einschränkung (32.3) zu verweisen.

Statistische Fehlerabschätzungen

33. Alle bisher abgeleiteten Fehlerabschätzungen liefern Schranken für den maximalen Rundungsfehler. Obwohl die meisten dieser Abschätzungen nicht scharf sind, kommen sie bis auf einen Faktor 2 oder 3 an die bestmögliche Abschätzung heran. Nun liegen aber die Rundungsfehler bei jeder einzelnen arithmetischen Operation zwischen dem $-\frac{1}{2}$- und $+\frac{1}{2}$fachen der letzten vorhandenen Stelle, und

man kann im allgemeinen annehmen, daß die Rundungsfehler im Verlauf einer längeren Rechnung sich in irgendeiner Weise über dieses Intervall verteilen.

Betrachten wir etwa unsere Abschätzungen für den Fehler bei einer Gleitpunktmultiplikation. Wir hatten hierfür angegeben

$$\begin{aligned} gl(x_1x_2)&\equiv x_1x_2(1+\varepsilon)\\ |\varepsilon|&\le 2^{-t}. \end{aligned}$$

Damit der Maximalfehler 2^{-t} tatsächlich angenommen wird, müssen nicht nur die weggelassenen Stellen der Mantisse ihren Maximalwert 2^{-1-t} annehmen; darüber hinaus muß die gerundete Mantisse auch noch den kleinsten möglichen Wert, nämlich $\frac{1}{2}$ besitzen. Es ist daher vernünftig anzunehmen, daß

$$\left.\begin{aligned} gl(x_1x_2\ldots x_n)&\equiv x_1x_2\ldots x_n(1+\varepsilon)\\ |\varepsilon|&<n^{1/2}\times 2^{-t} \end{aligned}\right\} \tag{33.1}$$

bei großem n eine Fehlerabschätzung für das mehrfache Produkt liefert, welche mehr den tatsächlichen Gegebenheiten entspricht.

Damit der Gesamtfehler bei einer größeren Summe die in Abschnitt 25 angegebene obere Schranke erreicht, muß nicht nur jeder einzelne Rundungsfehler seinen Maximalwert annehmen; darüber hinaus müssen die einzelnen Summanden $x_1, x_2, \ldots, x_n$ ganz spezielle Werte besitzen. Trotzdem werden wir in diesem Buch keinen Versuch unternehmen, um statistische Fehlerabschätzungen anzugeben, auch dann nicht, wenn wir nur ganz einfache Annahmen über die Verteilung der Fehler machen müßten, um die Abschätzung zu begründen. Gelegentlich werden wir allerdings ganz grobe Überschlagsrechnungen, wie sie auch (33.1) zugrundeliegen, vornehmen, um die Größenordnung des Fehlers zu ermitteln, der in der Praxis tatsächlich zu erwarten ist.

Blockskalierte Vektoren und Matrizen

34. Bei Festpunktrechnung gebraucht man im Zusammenhang mit Vektoren und Matrizen gerne eine [illegible] bis zu einem gewissen [illegible] die Gleitpunktdarstellung [illegible]

Statistical/Probabilistic Error Analysis

- Von Neumann & Goldstine (1947): Matrix inversion
- Hull & Swenson (1966): Matrix addition, multiplication, Runge Kutta
- Henrici (1966): ODEs
- Tienari (1970): Matrix inversion
- Barlow & Bareiss (1985): Gaussian elimination
- Calvetti (1991, 1992): Convolution, FFT
- Chatelin & Brunet (1990): Eigenvalues
- Higham & Mary (2018):

  Backward errors for: Inner products, matvec, matmult, LU, Cholesky

Inner (Dot) Product

- Given: Real vectors of dimension n

$$\boldsymbol{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T \qquad \boldsymbol{y} = \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix}^T$$

- Want: Inner product

$$\boldsymbol{x}^T \boldsymbol{y} = \sum_{j=1}^{n} x_j * y_j$$

- Floating point computation (guard digit model)

$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)\,(1+\delta) \qquad \mathrm{op} \in \{+, -, *, \ \}$$

$|\delta| \leq u$ where u is unit roundoff

Probabilistic Bound
First Try

Sequential Accumulation (Recursive Summation)

- Exact computation

$$
\begin{aligned}
s_1 &= x_1 y_1 \\
s_{k+1} &= s_k + x_{k+1} y_{k+1} \qquad 2 \leq k < n
\end{aligned}
$$

Output: $s_n = \boldsymbol{x}^T \boldsymbol{y}$

Sequential Accumulation (Recursive Summation)

- Exact computation

$$\begin{aligned} s_1 &= x_1 y_1 \\ s_{k+1} &= s_k + x_{k+1} y_{k+1} \qquad 2 \le k < n \end{aligned}$$

Output: $s_n = \boldsymbol{x}^T \boldsymbol{y}$

- Floating point arithmetic

$$\hat{s}_1 = x_1 y_1 \underbrace{(1+\theta_1)}_{* \text{ error}}$$

$$\hat{s}_{k+1} = \left(\hat{s}_k + x_{k+1} y_{k+1} \underbrace{(1+\theta_{k+1})}_{* \text{ error}} \right) \underbrace{(1+\delta_{k+1})}_{+ \text{ error}} \qquad 2 \le k < n$$

Output: $\hat{s}_n = \mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y})$

Breaking Down the Forward Error

- Computed inner product $\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) = \hat{s}_n = Z_1 + \cdots + Z_n$
- Local backward error

$$Z_k \equiv x_k y_k (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) \qquad 1 \leq k \leq n$$

- Local forward error

$$\begin{aligned} |Z_k - x_k y_k| &= |x_k y_k| \left| (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) - 1 \right| \\ &\leq |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \end{aligned}$$

Breaking Down the Forward Error

- Computed inner product $\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) = \hat{s}_n = Z_1 + \cdots + Z_n$
- Local backward error

$$Z_k \equiv x_k y_k (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) \qquad 1 \le k \le n$$

- Local forward error

$$\begin{aligned} |Z_k - x_k y_k| &= |x_k y_k| \left| (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) - 1 \right| \\ &\le |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \end{aligned}$$

- Total forward error = sum of local forward errors

$$|\,\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \le \underbrace{|Z_1 - x_1 y_1|}_{c_1} + \cdots + \underbrace{|Z_n - x_n y_n|}_{c_n}$$

Deterministic Version of the Error Bound

- Total forward error $\quad |\,\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \leq \sum_{k=1}^{n} c_k$

$$c_k = |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \qquad 1 \leq k \leq n$$

- Apply the Hölder inequality

$$\begin{aligned} \sum_{k=1}^{n} |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) &\leq \sum_{k=1}^{n} |x_k||y_k| \left((1+u)^{n} - 1 \right) \\ &\leq |\boldsymbol{x}|^T |\boldsymbol{y}| \, \frac{nu}{1-nu} \qquad \text{if } nu < 1 \end{aligned}$$

Deterministic Version of the Error Bound

- Total forward error $|\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \leq \sum_{k=1}^{n} c_k$

$$c_k = |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \qquad 1 \leq k \leq n$$

- Apply the Hölder inequality

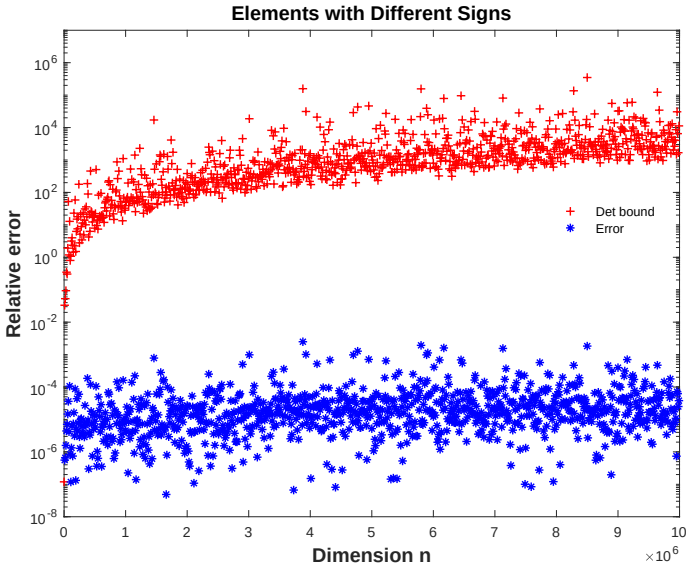
$$\begin{aligned} \sum_{k=1}^{n} |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) &\leq \sum_{k=1}^{n} |x_k||y_k| \left((1+u)^{n} - 1 \right) \\ &\leq |\boldsymbol{x}|^T |\boldsymbol{y}| \, \frac{nu}{1-nu} \qquad \text{if } nu < 1 \end{aligned}$$

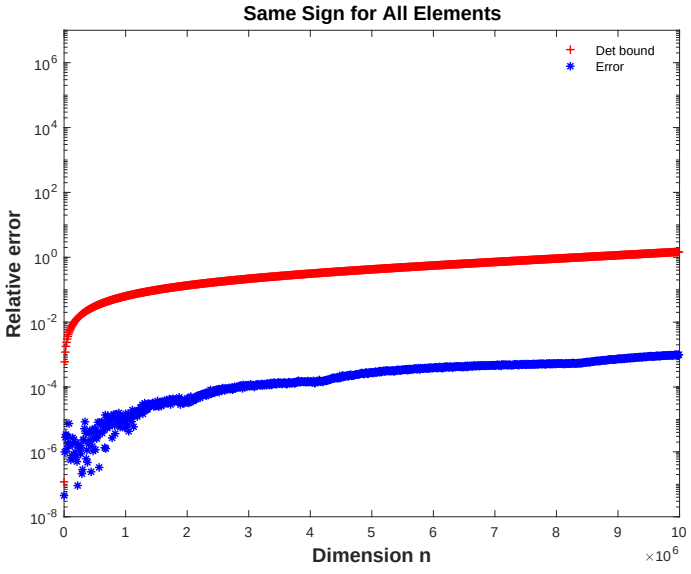
- This gives the traditional bound [Higham 2002]

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{y}} \right| \leq \frac{|\boldsymbol{x}|^T|\boldsymbol{y}|}{|\boldsymbol{x}^T\boldsymbol{y}|} \, \frac{nu}{1-nu} \qquad \text{if } nu < 1$$

Traditional Forward Error Bound is Pessimistic

$\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y})$ in single (binary32) $u = 2^{-24} \approx 5.96 \cdot 10^{-8}$
$\boldsymbol{x}^T\boldsymbol{y}$ in double (binary64) $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$





Errors $\leq 10^{-3}$ for dimensions $n \leq 10^7$

Bound several orders of magnitude larger than error

Bound not informative when elements have different signs

A Second Deterministic Version of the Error Bound

- Total forward error $\quad |\,\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}| \leq \sum_{k=1}^{n} c_k$

$$c_k = |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \qquad 1 \leq k \leq n$$

- Apply the Hölder inequality

$$\sum_{k=1}^{n} c_k \leq \sqrt{n} \sqrt{\sum_{k=1}^{n} c_k^2}$$

- The probabilistic bound is motivated by

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T\boldsymbol{y}) - \boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{y}} \right| \leq \sqrt{n} \, \frac{\sqrt{\sum_{k=1}^{n} c_k^2}}{|\boldsymbol{x}^T\boldsymbol{y}|}$$

Probabilistic Model for Roundoff Errors

1. Bounded random variables with zero mean

$$\mathrm{fl}(x \,\mathrm{op}\, y) \quad = \quad (x \,\mathrm{op}\, y)\,(1+\delta) \qquad |\delta| \le u, \quad \mathbb{E}[\delta] = 0$$

Mean of FP operation $=$ exact operation

$$\mathbb{E}\left[\mathrm{fl}(x \,\mathrm{op}\, y)\right] = (x \,\mathrm{op}\, y)\,(1+\mathbb{E}[\delta]) = x \,\mathrm{op}\, y$$

Probabilistic Model for Roundoff Errors

1. Bounded random variables with zero mean

$$\mathrm{fl}(x \operatorname{op} y) \quad = \quad (x \operatorname{op} y)\,(1+\delta) \qquad |\delta| \le u, \quad \mathbb{E}[\delta] = 0$$

Mean of FP operation = exact operation

$$\mathbb{E}\left[\mathrm{fl}(x \operatorname{op} y)\right] = (x \operatorname{op} y)\,(1+\mathbb{E}[\delta]) = x \operatorname{op} y$$

2. Independent

$$\begin{aligned} \mathrm{fl}(s+xy) &= (s + xy\,(1+\delta))\,(1+\theta) \\ &= s\,(1+\theta) + xy\,(1+\delta)\,(1+\theta) \end{aligned}$$

Mean of FP algorithm = exact algorithm

$$\begin{aligned} \mathbb{E}[\mathrm{fl}(s+xy)] &= s\,(1+\mathbb{E}[\theta]) + xy\,(1+\mathbb{E}[\delta])(1+\mathbb{E}[\theta]) \\ &= s + xy \end{aligned}$$

Probabilistic Interpretation of Roundoff

- Local backward error $\triangleq$ Unbiased random variable

$$Z_k = x_k y_k (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) \qquad \text{with} \quad \mathbb{E}[Z_k] = x_k y_k$$

- Local forward error $\triangleq$ Deviation of Z_k from its mean

$$|Z_k - \mathbb{E}[Z_k]| \leq \underbrace{|x_k y_k| \left((1+u)^{n-k+2} - 1 \right)}_{c_k} \qquad 1 \leq k \leq n$$

Probabilistic Interpretation of Roundoff

- Local backward error $\triangleq$ Unbiased random variable

$$Z_k = x_k y_k (1+\theta_k) \prod_{j=k}^{n} (1+\delta_j) \qquad \text{with} \quad \mathbb{E}[Z_k] = x_k y_k$$

- Local forward error $\triangleq$ Deviation of Z_k from its mean

$$|Z_k - \mathbb{E}[Z_k]| \leq \underbrace{|x_k y_k| \left((1+u)^{n-k+2} - 1 \right)}_{c_k} \qquad 1 \leq k \leq n$$

- Computed inner product $\triangleq$
  Sum of independent random variables

$$\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) = Z_1 + \cdots + Z_n$$

with bounded deviations from their means $|Z_k - \mathbb{E}[Z_k]| \leq c_k$

Azuma's Inequality

Given a sum

$$Z = Z_1 + \cdots + Z_n$$

of independent random variables $Z_1, \ldots, Z_n$
with bounded deviations from their means,

$$|Z_k - \mathbb{E}[Z_k]| \leq c_k \qquad 1 \leq k \leq n$$

For any $0 < \delta < 1$ with probability at least $1 - \delta$,
the deviation of the sum from its mean is

$$|Z - E[Z]| \leq \underbrace{\sqrt{\sum_{k=1}^{n} c_k^2}}_{\approx \text{ Variance}} \sqrt{2 \ln (2/\delta)}$$

Probabilistic Forward Error Bound

Assume:
Roundoffs are independent, zero-mean random variables in $[-u, u]$

For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \underbrace{\sqrt{2 \ln (2/\delta)}}_{\text{Probabilistic}}$$

where

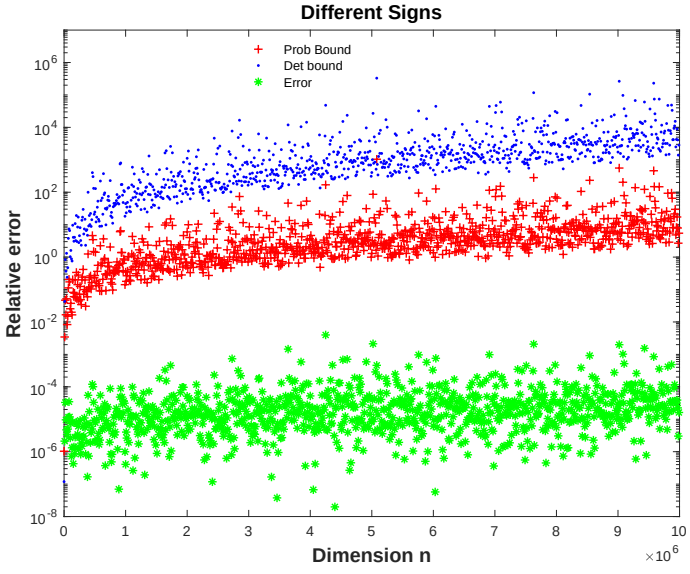
$$c_k \equiv |x_k y_k| \left((1+u)^{n-k+1} - 1 \right) \qquad 1 \leq k \leq n$$

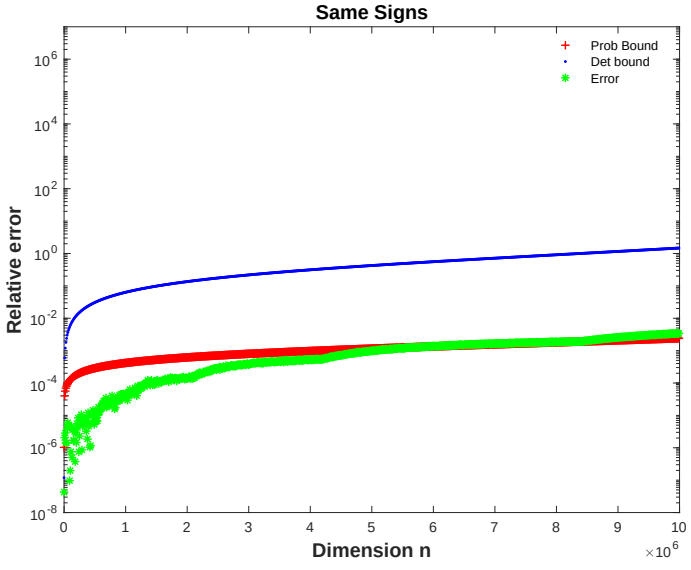
Probabilistic factor is small, even for tiny failure probability:

$$\sqrt{2 \ln (2/\delta)} \leq 9 \qquad \text{for} \quad \delta = 10^{-16}$$

Probabilistic and Traditional Bounds versus Error

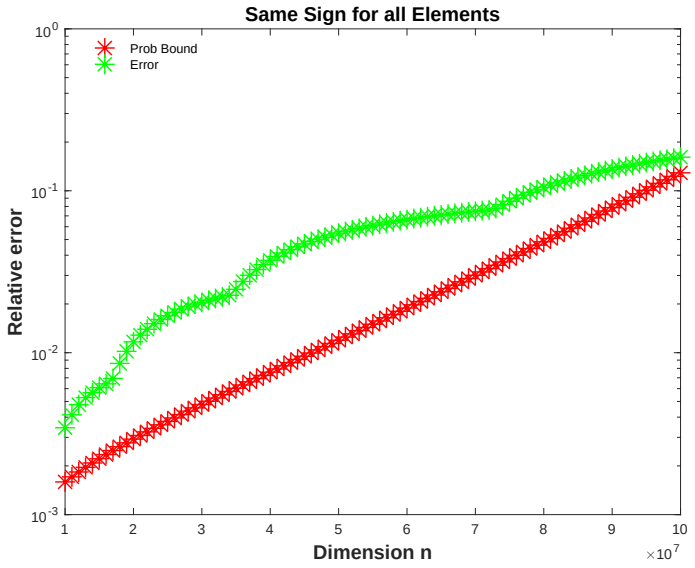
Failure probability $\delta = 10^{-16}$





Probabilistic bound tighter than traditional bound

Probabilistic Bound Stops Being a Bound



Crucial assumption: Roundoffs are independent

Henrici ends his 1965 paper with:

> *While this assumption seems to yield realistic results in many cases, some situations are known, [...], where local errors definitely cannot be considered to be independent. To elucidate the conditions under which local errors act like independent variables would seem to be a fascinating if difficult problem.*

Second Try
No Independence Assumptions on Roundoff

Avoiding Independence of Roundoffs

Think in terms of partial sums
Distinguish every roundoff

n multiplications and $n-1$ additions $\Longrightarrow$ $2n$ distinct roundoffs

(guard digit model)

	Floating Point Arithmetic	Exact Computation
Mult	$\hat{s}_1 = x_1 y_1 (1+\delta_1)$	$s_1 = x_1 y_1$
	$\hat{s}_2 = \hat{s}_1 (1+\delta_2)$	$s_2 = s_1$
Mult	$\hat{s}_{2k+1} = \hat{s}_{2k} + x_{k+1} y_{k+1} (1+\delta_{2k+1})$	$s_{2k+1} = s_{2k} + x_{k+1} y_{k+1}$
Add	$\hat{s}_{2k+2} = \hat{s}_{2k+1} (1+\delta_{2k+2})$	$s_{2k+2} = s_{2k+1}$
	$\hat{s}_{2n} = \mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y})$	$s_{2n} = \boldsymbol{x}^T \boldsymbol{y}$

Unravel the Total Forward Error

- Total forward error $Z_{2n} \equiv \hat{s}_{2n} - s_{2n} = \mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}$
- Forward errors of partial sums are recursive

$$Z_{2k-1} = \hat{s}_{2k-1} - s_{2k-1} = \underbrace{\hat{s}_{2k-2} - s_{2k-2}}_{Z_{2k-2}} + x_k y_k \, \delta_{2k-1}$$

$$Z_{2k} = \hat{s}_{2k} - s_{2k} = \underbrace{\hat{s}_{2k-1} - s_{2k-1}}_{Z_{2k-1}} + \hat{s}_{2k-1} \delta_{2k}$$

- Difference of partial sum errors bounded by incremental errors

$$|Z_{2k-1} - Z_{2k-2}| = |x_k y_k \, \delta_{2k-1}| \leq \underbrace{|x_k y_k|}_{\leq c_{2k-1}} u$$

$$|Z_{2k} - Z_{2k-1}| = |\hat{s}_{2k-1} \, \delta_{2k}| \leq \underbrace{|\hat{s}_{2k-1}|}_{\leq c_{2k}} u$$

Deterministic Version of Error Bound

- Total error is telescoping sum of incremental errors

$$
\begin{aligned}
|Z_{2n}| &\leq \underbrace{|Z_{2n} - Z_{2n-1}|}_{\leq c_{2n}\, u} + \underbrace{|Z_{2n-1} - Z_{2n-2}|}_{\leq c_{2n-1}\, u} + \cdots + \underbrace{|Z_1|}_{\leq c_1\, u} \\
&\leq \sum_{j=1}^{2n} c_j\, u \;\leq\; \sqrt{2n} \sqrt{\sum_{j=1}^{2n} c_j^2}\, u
\end{aligned}
$$

Deterministic Version of Error Bound

- Total error is telescoping sum of incremental errors

$$|Z_{2n}| \leq \underbrace{|Z_{2n} - Z_{2n-1}|}_{\leq c_{2n}\, u} + \underbrace{|Z_{2n-1} - Z_{2n-2}|}_{\leq c_{2n-1}\, u} + \cdots + \underbrace{|Z_1|}_{\leq c_1\, u}$$

$$\leq \sum_{j=1}^{2n} c_j\, u \leq \sqrt{2n} \sqrt{\sum_{j=1}^{2n} c_j^2}\, u$$

- The probabilistic bound is motivated by

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \sqrt{2n}\, \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|}\, u$$

with

$$c_{2k-1} = |x_k y_k| \qquad c_{2k} = \sum_{j=1}^{k} |x_j y_j| (1+u)^{k-j+1}$$

Probabilistic Version of Error Bound

Assume: Roundoffs are zero-mean random variables: $\mathbb{E}[\delta_j] = 0$

1. Forward errors in partial sums are random variables

$$Z_{2k-1} \quad = \quad Z_{2k-2} + x_k y_k \, \delta_{2k-1}$$

2. Conditioned on previous roundoffs,

$$\mathbb{E}\left[Z_{2k-1} | \delta_1, \ldots, \delta_{2k-2}\right] = Z_{2k-2}$$

   mean of current error equals value of previous error

3. Difference of partial sum errors bounded by incremental errors

$$|Z_{2k-1} - Z_{2k-2}| \leq c_{2k-1} \, u$$

Probabilistic Version of Error Bound

Assume: Roundoffs are zero-mean random variables: $\mathbb{E}[\delta_j] = 0$

1. Forward errors in partial sums are random variables

$$Z_{2k-1} \quad = \quad Z_{2k-2} + x_k y_k \, \delta_{2k-1}$$

2. Conditioned on previous roundoffs,

$$\mathbb{E}\left[Z_{2k-1} | \delta_1, \ldots, \delta_{2k-2}\right] = Z_{2k-2}$$

   mean of current error equals value of previous error

3. Difference of partial sum errors bounded by incremental errors

$$|Z_{2k-1} - Z_{2k-2}| \leq c_{2k-1} \, u$$

Partial sum errors $Z_1, Z_2, \ldots$ form Martingale
with respect to roundoffs $\delta_1, \delta_2, \ldots$

Azuma-Hoeffding Martingale

Sequence of random variables $Z_0, Z_1 \ldots$ is Martingale with respect to sequence $\delta_1, \delta_2 \ldots$ if for $k \geq 0$

1. Z_k is function of $\delta_1, \ldots, \delta_k$
2. $\mathbb{E}[|Z_k|] < \infty$,
3. $\mathbb{E}\left[Z_{k+1} | \delta_1, \ldots, \delta_k\right] = Z_k$

If also

$$|Z_{k+1} - Z_k| \leq c_k \qquad 0 \leq k < m$$

Then: For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$|Z_m - Z_0| \leq \sqrt{\sum_{k=1}^{m} c_k^2} \sqrt{2 \ln (2/\delta)}$$

Probabilistic Forward Error Bound, v2

Assume: Roundoffs are zero-mean random variables in $[-u, u]$

- For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{2 \ln(2/\delta)}\, u$$

where

$$c_{2k-1} = |x_k y_k| \qquad c_{2k} = \sum_{j=1}^{k} |x_j y_j| (1+u)^{k-j+1}$$

Probabilistic Forward Error Bound, v2

Assume: Roundoffs are zero-mean random variables in $[-u, u]$

- For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{2 \ln(2/\delta)}\, u$$

  where

$$c_{2k-1} = |x_k y_k| \qquad c_{2k} = \sum_{j=1}^{k} |x_j y_j| (1+u)^{k-j+1}$$

- In comparison: Deterministic version of this bound is

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{2n}\, u$$

Probabilistic Forward Error Bound, v2

For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{2 \ln(2/\delta)}\, u$$

where $c_{2k-1} = |x_k y_k|$ and $c_{2k} = \sum_{j=1}^{k} |x_j y_j| (1+u)^{k-j+1}$

This is really complicated ☹

Probabilistic Forward Error Bound, v2

For any $0 < \delta < 1$, with probability at least $1 - \delta$

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{\sqrt{\sum_{k=1}^{2n} c_k^2}}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{2 \ln(2/\delta)}\, u$$

where $c_{2k-1} = |x_k y_k|$ and $c_{2k} = \sum_{j=1}^{k} |x_j y_j| (1+u)^{k-j+1}$

This is really complicated ☹

Find an upper bound:

$$\begin{aligned} \sqrt{\sum_{k=1}^{2n} c_k^2} &\leq \frac{|\boldsymbol{x}|^T |\boldsymbol{y}|}{|\boldsymbol{x}^T \boldsymbol{y}|} \sqrt{\tfrac{u}{2}\left((1+u)^{2n+2} - 1\right)} \\ &\leq \frac{|\boldsymbol{x}|^T |\boldsymbol{y}|}{|\boldsymbol{x}^T \boldsymbol{y}|} \frac{\sqrt{n+1}\, u}{1 - (2n+2)u} \quad \textit{if } (2n+2)u < 1 \end{aligned}$$

Comparison of Forward Error Bounds: Traditional vs Probabilistic

$$\left| \frac{\mathrm{fl}(\boldsymbol{x}^T \boldsymbol{y}) - \boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{y}} \right| \leq \frac{|\boldsymbol{x}|^T |\boldsymbol{y}|}{|\boldsymbol{x}^T \boldsymbol{y}|} \, \Delta \, u$$

Assume: $\delta = 10^{-16}$, $u \approx 6 \cdot 10^{-8}$, $n \leq 10^7$

- Traditional bound:

$$\Delta = \frac{n}{1 - nu} \leq 2.5\, n$$

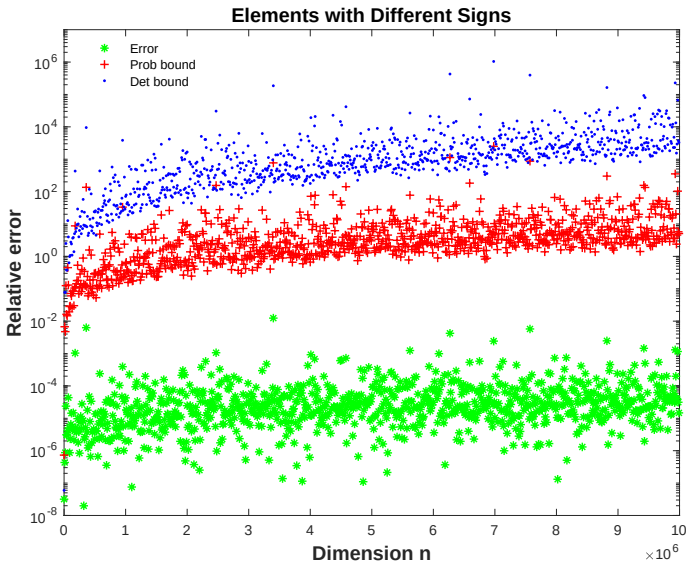
- Probabilistic bound:

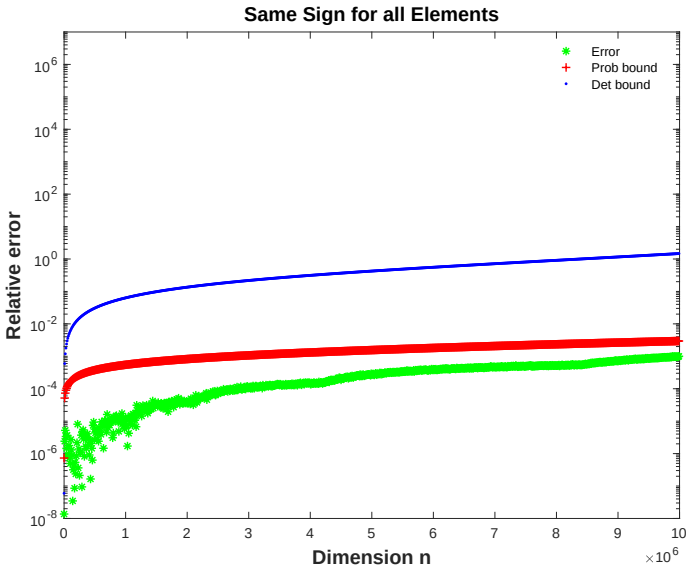
$$\Delta = \sqrt{\tfrac{u}{2}\left((1+u)^{2n+2} - 1\right)} \sqrt{2 \ln\left(2/\delta\right)} \leq 12.1 \sqrt{n+1}$$

Traditional bound $\sim n$
Probabilistic bound $\sim \sqrt{n}$

Probabilistic and Traditional Bounds versus Error Dimension $1 \leq n \leq 10^7$

Failure probability $\delta = 10^{-16}$

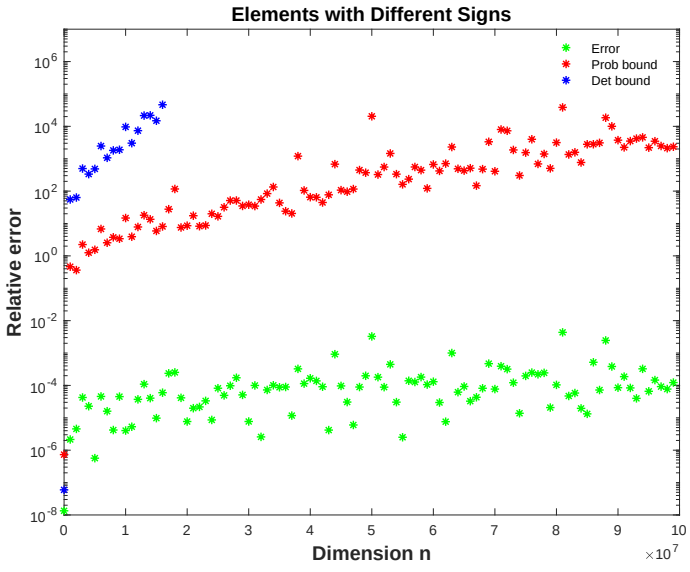


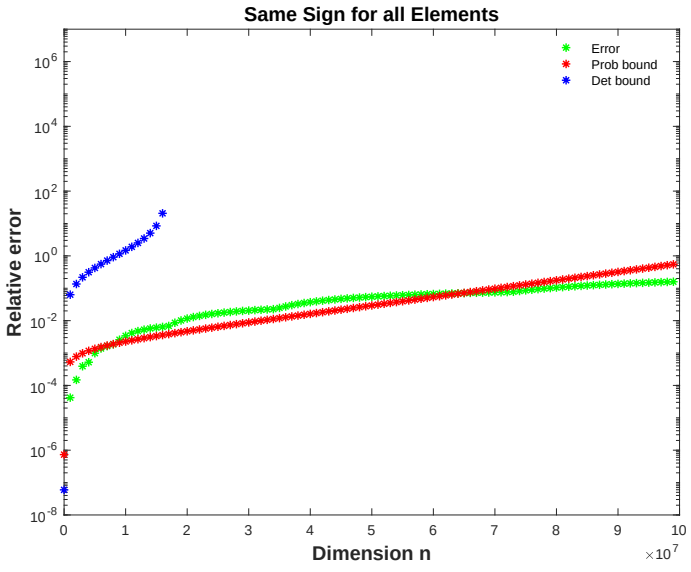


Probabilistic bound tighter than traditional bound

Probabilistic and Traditional Bounds versus Error Dimension $10^6 \leq n \leq 10^8$

Failure probability $\delta = 10^{-16}$





Traditional bound not valid anymore

Probabilistic bound

Too pessimistic for elements with different signs

Stops being an upper bound for elements with same sign

???????

A cheap fix:

- Increasing the failure probability δ makes the probabilistic bound less pessimistic
  for vector elements with different signs
- Decreasing the failure probability δ makes the "probabilistic bound" an upper bound again for large n
  when all vector elements have the same sign

However:

- The concentration inequalities do not explicitly depend on the number n of random variables in a sum
- Forward errors do depend on the number n of summands

Should the failure probability δ depend on the dimension n
and if so, how (systematically)?

Summary

Probabilistic roundoff error analysis for inner products

- New forward error bounds

  (from Martingale concentration inequalities)
- Explicit non-asymptotic bounds, with minimal assumptions

  (no limit on dimension n, independence of roundoffs)
- Extremely stringent success probabilities ($\delta = 10^{-16}$)
- Forward error $\sim \sqrt{n}$ instead of n

Next

- Computations in arbitrary precision (Julia BigFloat)
  Roundoff error model without guard digits
- Allow biased roundoff with non-zero mean
- Analysis of matrix decompositions,
  with matrix concentration inequalities (Kyng, Tropp)