A few observations about summation algorithms

Ilse C.F. Ipsen

Joint work with Eric Hallman

also thanks to Johnathan Rhyne

North Carolina State University
Raleigh, NC, USA

Research supported by NSF DMS

This talk

Forward error when summing n real numbers $x_1, \ldots, x_n$

$$s_n = x_1 + \cdots + x_n$$

in floating point arithmetic

Overview

- The traditional method
- More accurate methods (without higher precision or better hardware)
  - Inspired by computer architecture and formal methods: Shifted summation
  - Kahan's 1965 method: Compensated summation
- Summary

The traditional method

Sequential (recursive) summation

[Higham: Accuracy and Stability of Numerical Algorithms, Chapter 4]

- Exact arithmetic

$$s_1 = x_1, \qquad s_k = s_{k-1} + x_k, \quad 2 \le k \le n$$

- Floating point arithmetic[1]

$$\hat{s}_1 = x_1, \qquad \hat{s}_k = (\hat{s}_{k-1} + x_k)(1 + \delta_k), \quad 2 \le k \le n$$

  where $|\delta_k| \le u$ are $n-1$ roundoffs

- First order forward error

$$\left| \frac{\hat{s}_n - s_n}{s_n} \right| \le n\,u \underbrace{\frac{\sum_{k=1}^{n} |x_k|}{|s_n|}}_{\ge 1} + \mathcal{O}(u^2)$$

[1] Assume: $x_1, \ldots, x_n$ are floating point numbers

Tighter forward error bounds for sequential summation

- Exact expression [Hallman 2020]

$$\hat{s}_n - s_n = \sum_{k=2}^{n} s_k \, \delta_k \prod_{j=k}^{n} (1 + \delta_j)$$

- Probabilistic bounds

  [Higham, Mary 2019, 2020], [Rhyne 2020], [Hallman 2020]

  Treat roundoffs δ_k as zero-mean bounded random variables
  (independent, or mean-independent)
  Error $\approx \mathcal{O}(\sqrt{n}u)$ with high probability

- Next: More accurate algorithms

Shifted summation

Motivation for shifted summation

- Computer architecture, formal methods for program verification:

  [Solovyev et al 2015], [Dahlquist, Salvia, Constatinides 2019], [Lohar, Prokop, Darulova 2019], [Constantinides, Dahlquist, Rakamaric, Salvia 2021]

  Assume $x_1, \ldots, x_n$ drawn from a distribution
  Compute statistics for accumulated errors
  Determine probability of errors in certain interval

- Probabilistic bounds for random data

  [Higham, Mary 2020], [Hallman 2020]

  Idea: Sequential summation accurate if x_k tightly clustered

Sequential summation of random data [Hallman 2020]

Assume:

- Roundoffs δ_k are independent zero-mean random variables, and $|\delta_k| \leq u$
- Summands x_k are random variables with mean μ and 'variance' $\max_{1 \leq k \leq n} |x_k - \mu| \leq \sigma$

Then for any $0 < \delta < 1$ with probability at least $1 - \delta$

$$|\hat{s}_n - s_n| \leq (1 + \gamma)(\lambda\, n^{3/2}\, |\mu| + \lambda^2 n\, \sigma)\, u$$

where $\lambda = \sqrt{2 \ln (6/\delta)}$ and $\gamma = \exp\left(\lambda\left(\sqrt{n}u + nu^2\right)/(1-u)\right) - 1$

Sequential summation accurate if x_k tightly clustered around zero

Application to non-random data [Higham, Mary 2020]

Example: Assume $x_k = 10^4 + y_k$ where $|y_k| \leq 1$

- Sequential summation
  $|\hat{s}_n - s_n| \leq nu \sum_{k=1}^{n} |x_k| + \mathcal{O}(u^2) = n^2 u\,(10^4 + 1) + \mathcal{O}(u^2)$
- 'Center' the data to reduce their 'mean'

  Center: $y_k = x_k - 10^4, \;\; 1 \leq k \leq n$
  Sum centered data
  $t_1 = y_1, \; t_k = t_{k-1} + y_k, \;\; 1 \leq k \leq n$
  Uncenter: $s = t_n + n \cdot 10^4$

$|\hat{t}_n - t_n| \leq nu \sum_{k=1}^{n} |y_k| + \mathcal{O}(u^2) \leq n^2 u + \mathcal{O}(u^2)$

Centered sum captures 'tail' bits of $x_k \Longrightarrow$ smaller error

Shifted sequential summation [Higham, Mary 2020]

Input: summands $x_1, \ldots, x_n$, shift c

$t_0 = 0$

for $k = 1 : n$ do

$\hat{y}_k = (x_k - c)(1 + \epsilon_k)$ $\{\epsilon_k$ are centering roundoffs$\}$

$\hat{t}_k = (\hat{t}_{k-1} + \hat{y}_k)(1 + \delta_k)$ $\{\delta_k$ are summation roundoffs$\}$

end for

$\hat{y}_{n+1} = cn\,(1 + \epsilon_{n+1})$

$\hat{s}_n = (\hat{t}_n + \hat{y}_{n+1})(1 + \delta_{n+1})$ $\{$uncentering$\}$

Error [Hallman, II 2021]

$$\hat{s}_n - s_n = \underbrace{\sum_{k=1}^{n+1} t_k \delta_k \prod_{\ell=k+1}^{n+1} (1 + \delta_\ell)}_{\text{Summation}} + \underbrace{\sum_{k=1}^{n+1} y_k \epsilon_k \prod_{\ell=k}^{n+1} (1 + \delta_\ell)}_{\text{Centering}}$$

Error bounds for shifted summation [Hallman, II 2021]

Error depends only on shifted quantities

- Deterministic

$$|\hat{s}_n - s_n| \leq u(1+u)^n \left(\sum_{k=2}^{n} \underbrace{|s_k - kc|}_{|t_k|} + \sum_{k=1}^{n} \underbrace{|x_k - c|}_{|y_k|} + |s| + |nc| \right)$$

- Probabilistic

  Let δ_j and ϵ_j be independent zero-mean random variables
  Then for any $0 < \delta < 1$ with probability at least $1 - \delta$,

$$|\hat{s}_n - s_n| \leq \max_{1 \leq k \leq n+1} (|s_k - kc| + |x_k - c|)\, \beta \sqrt{2 \ln(2/\delta)}$$

  where $\beta = \sqrt{\frac{u}{2}\gamma_{2n}} \leq \sqrt{\frac{(n+2)u^2}{1-2(n+2)u}} \approx \sqrt{n+2}\, u$

  (provided $n \ll 1/u$)

Numerical experiments: Sequential summation with shifting vs no shifting

- Number of summands $n = 10^6$
- Shift $c = (\max_k x_k + \min_k x_k)/2$
- Working precision: Julia Float64 $\quad u = 2^{-53} \approx 1.11 \cdot 10^{-16}$
- 'Exact' computation: Julia Float256
- Plots: relative errors $|\hat{s}_n - s_n|/|s_n|$ versus n
- Probabilistic 'bound'

$$u\sqrt{n+2} \max_{1 \le k \le n+1} \frac{|s_k - kc| + |x_k - c|}{|s_n|} \underbrace{\sqrt{2\ln(2/\delta)}}_{3.26} \qquad \delta = 10^{-2}$$

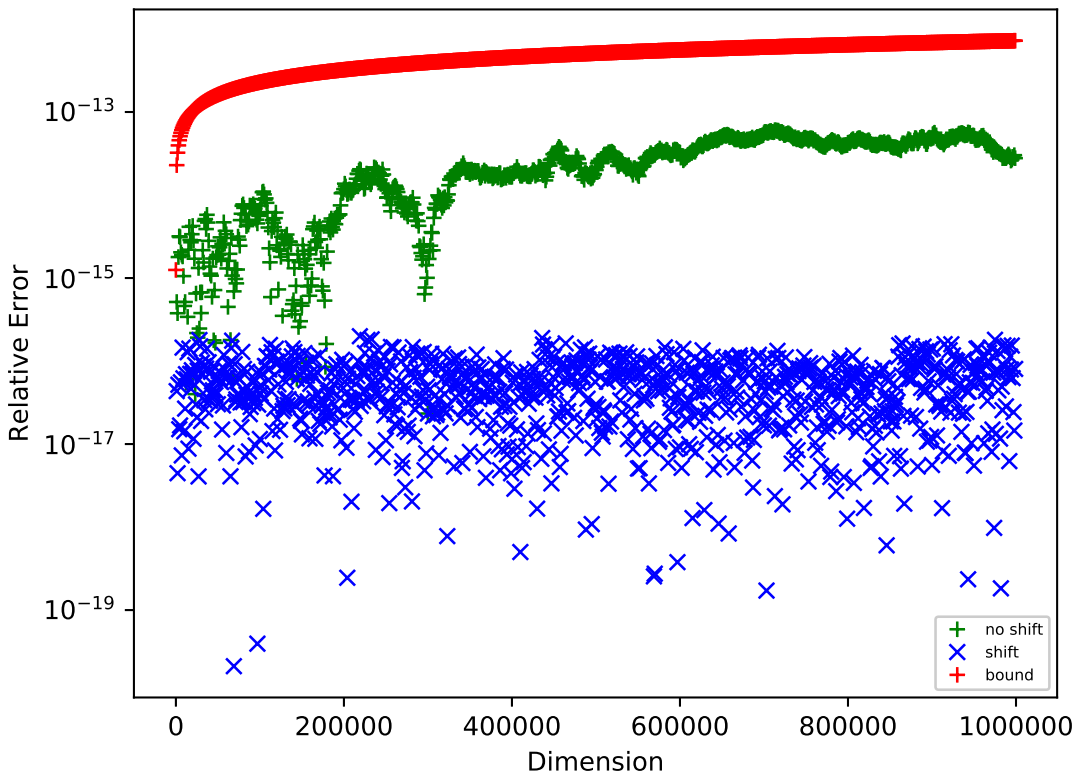
Two different types of summands x_k, $1 \le k \le n$

Figure 1: $x_k = 10^4 + y_k$ where y_k is uniform[0,1]

Figure 2: x_k is normal(0,1)

Summands tightly clustered at 10^4

$x_k = 10^4 + y_k$ where y_k is uniform[0,1]

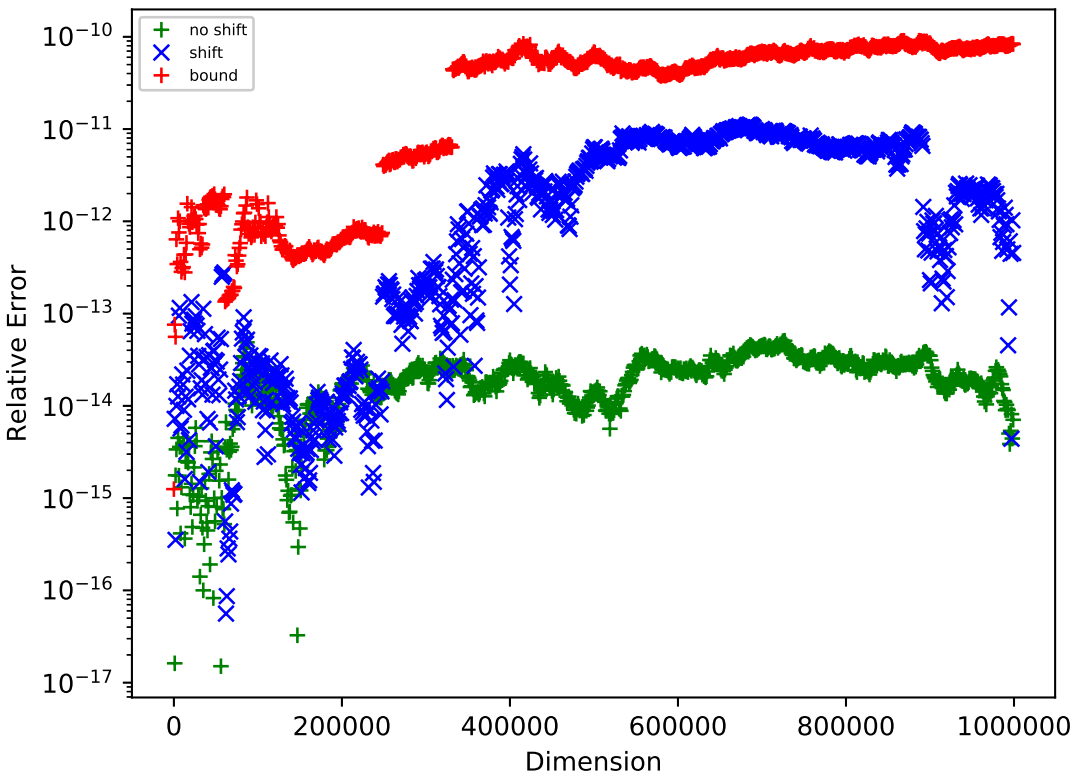


Shifting increases accuracy compared to plain summation

Bound accurate within factor of 100

Summands tightly clustered at zero

x_k is normal(0,1)



Shifting hurts accuracy $(-.2 \leq c \leq 1.2)$ over plain summation

Bound accurate within factor of 10-100

Summary: Shifted sequential summation

- Error bounds depend only on shifted quantities
  and hold to all orders of u
- Probabilistic bound: Error $\approx \mathcal{O}(\sqrt{n}u)$
- Shifting improves accuracy if
  shift c decreases magnitude of partial sums

$$\max_k \left(|s_k - k\,c| + |x_k - c|\right) \ll \max_k \left(|s_k| + |x_k|\right)$$

See advice in [Higham book, Section 4.2]

- Extension to general summation algorithms [Hallman, II 2021]

  Roundoffs treated as (mean-) independent random variables
  Probabilistic bounds depend on $\sqrt{\text{height of computational tree}}$
  Bounds valid to all orders of u

- Warning: Shifting can worsen accuracy
  for naturally centered data, such as $\mathit{normal}(0, 1)$

Kahan's 1965 method: Compensated sequential summation

Compensated sequential summation

[Goldberg 1991], [Higham book], [Kahan 1973]

Input: summands $x_1, \ldots, x_n$

$s_1 = x_1,\ c_1 = 0$

for $k = 2 : n$ do

$y_k = x_k - c_{k-1}$

$s_k = s_{k-1} + y_k$

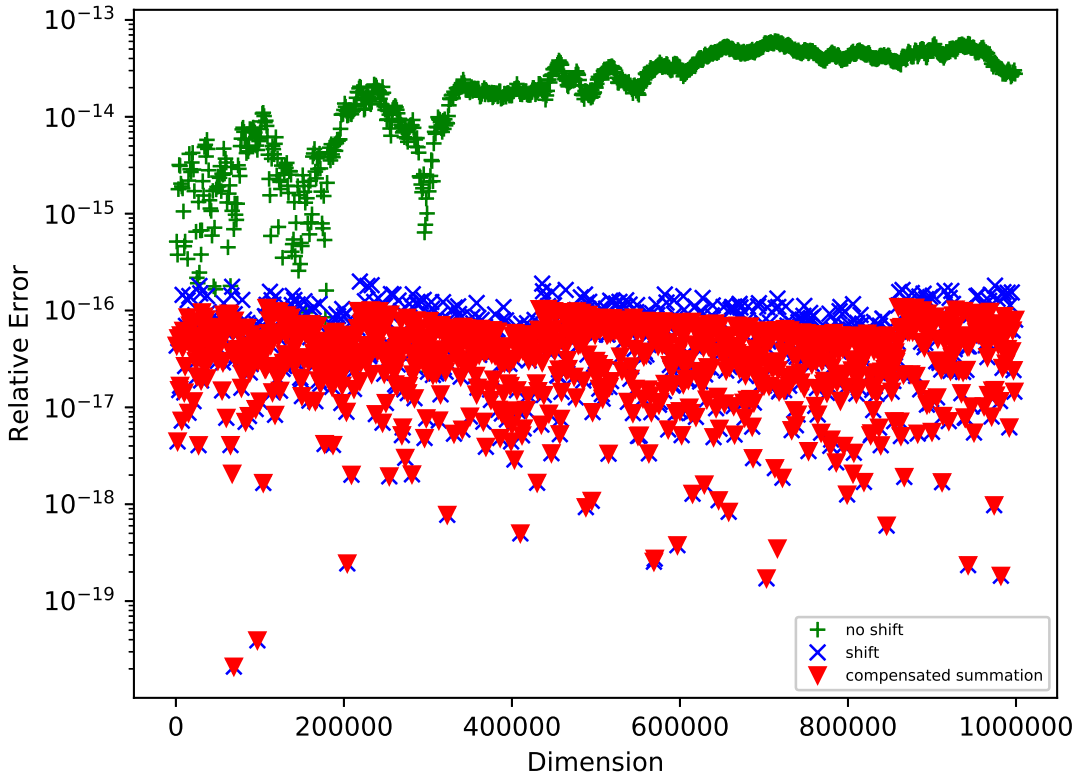
$c_k = (s_k - s_{k-1}) - y_k$

end for

return s_n

Summands tightly clustered at 10^4

$x_k = 10^4 + y_k$ where y_k is uniform[0,1]

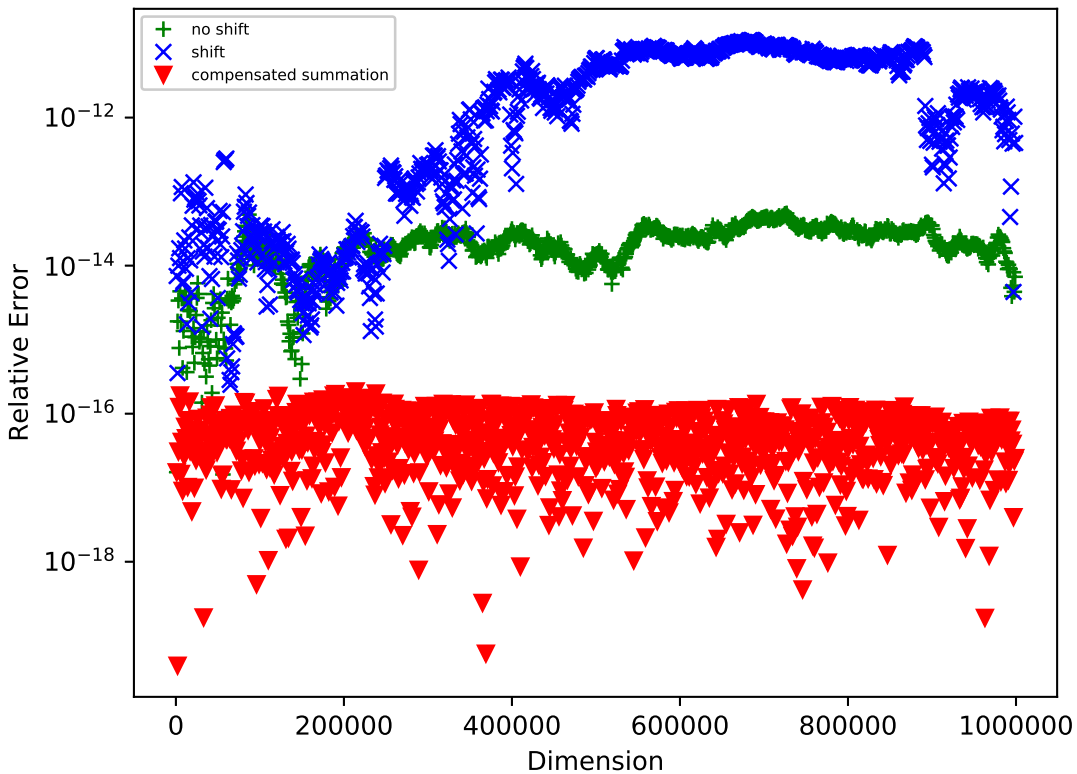


Compensated summation as or more accurate

than shifted summation with $c = (\min_k x_k + \max_k x_k)/2$

Summands clustered at zero

x_k is normal(0,1)



Compensated summation more accurate

than shifted summation with $c = (\min_k x_k + \max_k x_k)/2$

Roundoff in sequential compensated summation

[Goldberg 1991], [Kahan 1973]

$\hat{s}_1 = s_1 = x_1,\ \hat{c}_1 = 0$
for $k = 2 : n$ do
  $\hat{y}_k = (x_k - \hat{c}_{k-1})(1 + \eta_k)$
  $\hat{s}_k = (\hat{s}_{k-1} + \hat{y}_k)(1 + \sigma_k)$
  $\hat{c}_k = ((\hat{s}_k - \hat{s}_{k-1})(1 + \delta_k) - \hat{y}_k)(1 + \beta_k)$
end for

Forward error

$$|\hat{s}_n - s_n| \leq \left(2u + \mathcal{O}(nu^2)\right) \sum_{k=1}^{n} |x_k|$$

Recursion for error and correction [Hallman, II 2021]

$$
\begin{aligned}
e_k &= \hat{s}_k - s_k \\
\hat{c}_k &= ((\hat{s}_k - \hat{s}_{k-1})(1+\delta_k) - \underbrace{(x_k - \hat{c}_{k-1})(1+\eta_k)}_{\hat{y}_k})(1+\beta_k)
\end{aligned}
$$

satisfy the recurrence

$$
\begin{bmatrix} e_k \\ \hat{c}_k \end{bmatrix} = P_k \begin{bmatrix} e_{k-1} \\ \hat{c}_{k-1} \end{bmatrix} + P_k \begin{bmatrix} s_{k-1} \\ -x_k \end{bmatrix} + \begin{bmatrix} -s_k \\ 0 \end{bmatrix} \qquad 2 \le k \le n
$$

where

$$
P_k \equiv \begin{bmatrix} 1+\sigma_k & -(1+\eta_k)(1+\sigma_k) \\ \sigma_k(1+\delta_k)(1+\beta_k) & (1+\eta_k)(1-(1+\sigma_k)(1+\delta_k))(1+\beta_k) \end{bmatrix}
$$

Explicit expressions for error and correction [Hallman, II 2021]

$$\begin{bmatrix} e_n \\ \hat{c}_n \end{bmatrix} = (P_n \cdots P_2) \begin{bmatrix} x_1 \\ 0 \end{bmatrix} + \sum_{j=2}^{n} (P_n \cdots P_j) \begin{bmatrix} 0 \\ -x_j \end{bmatrix} + \begin{bmatrix} -s_n \\ 0 \end{bmatrix}$$

A second explicit expression

$$\begin{aligned} e_n &= s_n \sigma_n + \sum_{j=4}^{n} x_j \eta_j \prod_{k=j}^{n} (1 + \sigma_k) \\ &\quad + \sum_{j=2}^{n-1} (s_j \sigma_j - \hat{c}_j (1 + \eta_{j+1})) \prod_{k=j+1}^{n} (1 + \sigma_k) \end{aligned}$$

A third explicit expression [Hallman, II 2021]

$$e_n = \hat{s}_n - s_n = s_n\,\sigma_n + X_n + \underbrace{X_n\sigma_n + E_{n-1}(1+\sigma_n)}_{\mathcal{O}(u^2)}$$

where

$$X_n \equiv x_n\eta_n\,(1+\beta_n) + \underbrace{\sum_{j=2}^{n-1} x_j(\eta_j - \delta_j)}_{\mathcal{O}(u)} \prod_{\ell=j}^{n}(1+\beta_\ell)(1+\eta_{\ell+1})$$

$$E_{n-1} \equiv e_{n-1}\Theta_{n-1} + \sum_{j=2}^{n-1} e_j(\Theta_j + \delta_{j+1}) \prod_{\ell=j+1}^{n}(1+\beta_\ell)(1+\eta_{\ell+1})$$

$$\Theta_k = 1 - (1+\delta_k)(1+\beta_k)(1+\eta_{k+1}) \qquad 2 \le k \le n-1$$

First order error [Hallman, II 2021]

- The previous expressions imply

$$\hat{s}_n - s_n = s_n \sigma_n + x_n \eta_n + \sum_{k=2}^{n-1} x_k(\eta_k - \delta_k) + \mathcal{O}(u^2)$$

Error dominated by correction and final summation roundoffs

$\hat{y}_k = (x_k - \hat{c}_{k-1})(1+\eta_k)$, $\hat{c}_k = ((\hat{s}_k - \hat{s}_{k-1})(1+\delta_k) - \hat{y}_k)(1+\beta_k)$

$\hat{s}_n = (\hat{s}_{n-1} + \hat{y}_n)(1+\sigma_n)$

- First order bound $|\hat{s}_n - s_n| \leq 3u \sum_{k=1}^{n} |x_k| + \mathcal{O}(u^2)$

inconsistent with

previous bounds $|\hat{s}_n - s_n| \leq 2u \sum_{k=1}^{n} |x_k| + \mathcal{O}(u^2)$

Second order error [Hallman, II 2021]

Let $\mu_k \equiv \eta_k - \delta_k$, $1 \leq k \leq n-1$, $\quad \mu_n = \eta_n$

$$\hat{s}_n - s_n = s_n\sigma_n + \sum_{k=2}^{n} x_k\mu_k\,(1+\sigma_n)$$
$$- \sum_{k=2}^{n-1} \left(s_k\sigma_k(\mu_{k+1}+\beta_k+\delta_k) + x_k\delta_k(\mu_{k+1}+\beta_k+\eta_k)\right) + \mathcal{O}(u^3)$$

Deterministic bound

$$|\hat{s}_n - s_n| \leq (3u + 4n\,u^2) \sum_{k=1}^{n} |x_k| + \mathcal{O}(u^3)$$

Probabilistic bounds [Hallman, II 2021]

Vector of summands $\mathsf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$
Independent zero-mean random variables δ_j and ϵ_j
Then for any $0 < \delta < 1$ with probability at least δ

Second order bound

$$\begin{aligned} |\hat{s}_n - s_n| &\leq u\left((2+6u)\,\|\mathsf{x}\|_2 + \sqrt{|s_n|^2 + 16u^2\sum_{k=1}^{n-1}|s_k|^2}\right)\sqrt{2\ln(2/\delta)} + \mathcal{O}(u^3) \\ &\leq u\left((2+6u)\|\mathsf{x}\|_2 + \sqrt{1+16(n-2)u^2}\|\mathsf{x}\|_1\right)\sqrt{2\ln(2/\delta)} + \mathcal{O}(u^3) \end{aligned}$$

First order bound

$$|\hat{s}_n - s_n| \leq u(2\|\mathsf{x}\|_2 + |s_n|)\sqrt{2\ln(2/\delta)} + \mathcal{O}(u^2)$$

Numerical experiments: Compensated summation in half precision

- Working precision: Julia Float16 $\quad u = 2^{-11} \approx 4.88 \cdot 10^{-4}$
- 'Exact' computation: Julia Float64
- Plots: relative errors $|\hat{s}_n - s_n|/|s_n|$ versus n
- First order probabilistic bound

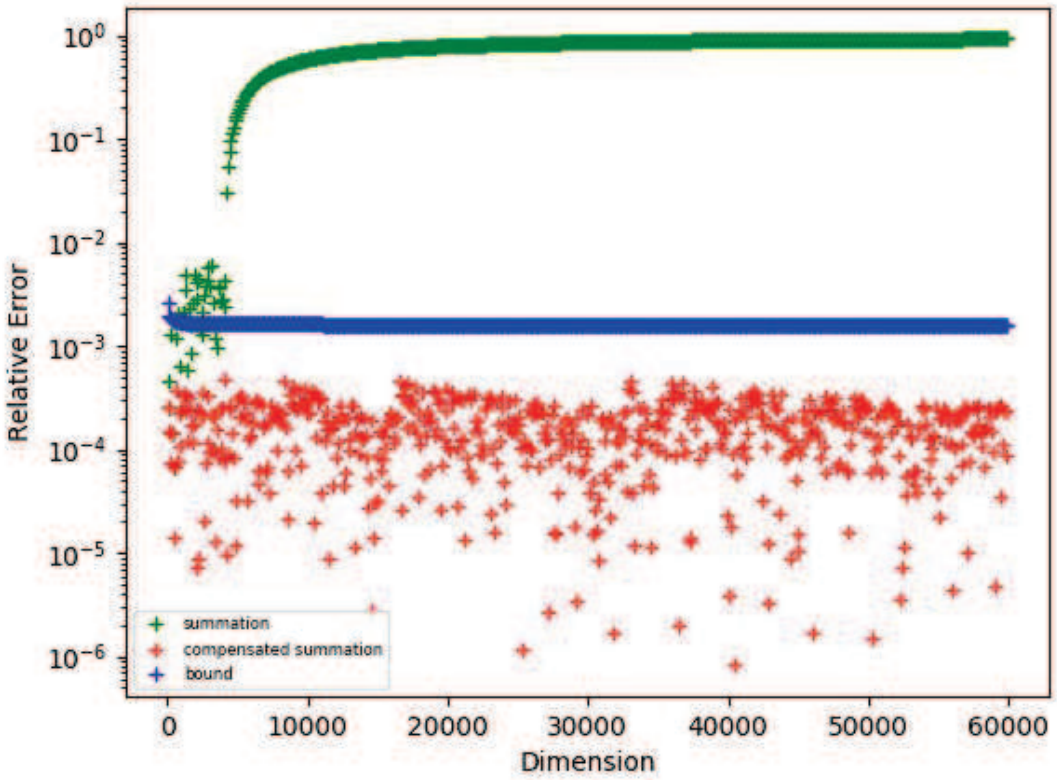
$$u \, \frac{2\|x\|_2 + |s_n|}{|s_n|} \, \underbrace{\sqrt{2\ln(2/\delta)}}_{3.26} \qquad \text{where} \quad \delta = 10^{-2}$$

Two different types of summands x_k, $1 \le k \le n$

Figure 1: x_k is uniform[0,1], $\quad n = 6 \cdot 10^4$

Figure 2: x_k is normal(0,1), $\quad n = 10^6$

Summands have same sign: x_k is uniform[0,1]

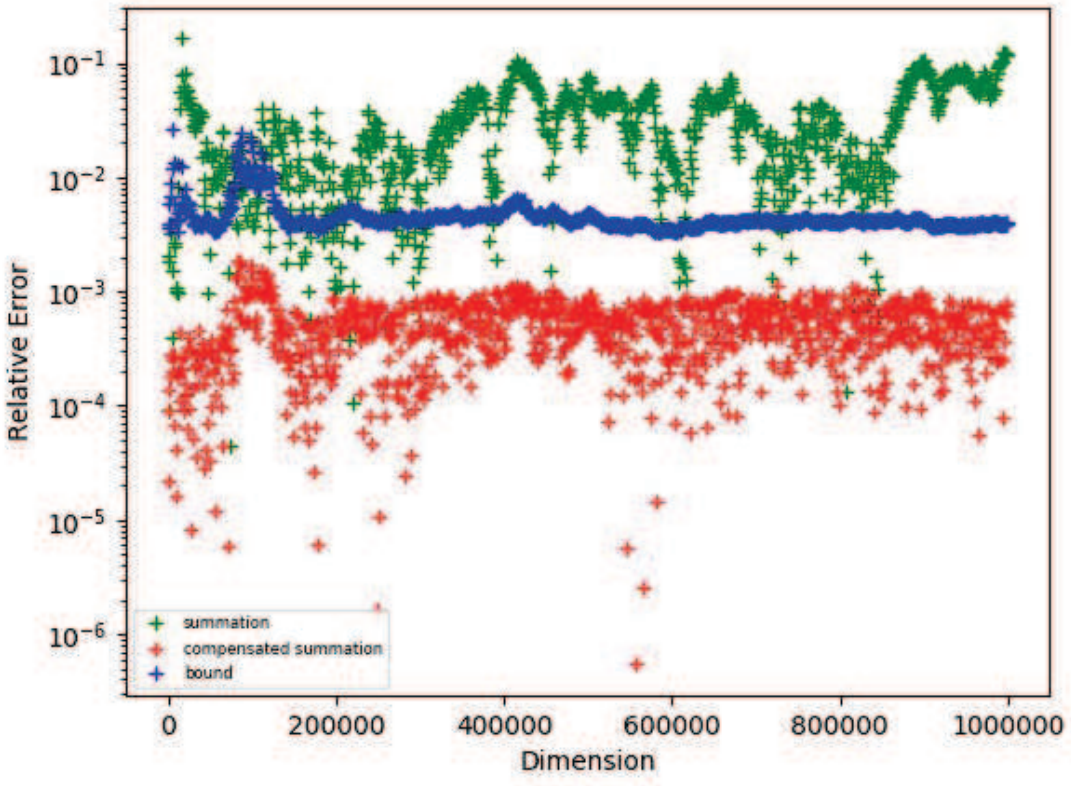


Compensated summation accurate to machine precision

Ordinary summation not accurate (well-conditioned problem)

Bound accurate within factor of 10

Summands have different signs: x_k is normal(0,1)



Bound accurate within factor of 10

Summary

Forward error in summation of n real numbers
without recourse to higher precision or better arithmetic

- Shifted sequential summation
  Explicit expression and probabilistic bounds valid to all orders
  Extension to general summation algorithms
  Centering more accurate if it decreases magnitude of partial sums
  But centering can hurt accuracy
  Mixed precision versions do not appear to be effective
- Compensated sequential summation
  Three explicit expressions for error
  First and second order deterministic and probabilistic error bounds
  First order bound differs by u from existing bounds
  First order probabilistic bound accurate within factor of 10
  Accurate but more expensive
  Compromise: FABsum [Blanchard, Higham, Mary 2020]

E. Hallman and I. C. F. Ipsen:
Deterministic and Probabilistic Error Bounds for Floating Point Summation Algorithms, arXiv:2107.01604