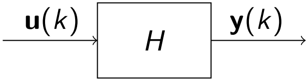
Reliable Evaluation of the Worst-Case Peak Gain Matrix in Multiple Precision

Anastasia Volkova, Thibault Hilaire, Christoph Lauter

Sorbonne Universités, UPMC Univ Paris 06, UMR 7606,
LIP6, F-75005, Paris, France

22nd IEEE Symposium on Computer Arithmetic
June 23, 2015

Digital filters



Digital filters

$$\xrightarrow{\mathbf{u}(k)} \boxed{H} \xrightarrow{\mathbf{y}(k)}$$

Linear Time-Invariant filter in state-space representation:

$$H \begin{cases} \mathbf{x}(k+1) & = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) & = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \end{cases}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{B} \in \mathbb{R}^{n \times q}, \mathbf{C} \in \mathbb{R}^{p \times n}, \mathbf{D} \in \mathbb{R}^{p \times q}$

Digital filters

$$\xrightarrow{\mathbf{u}(k)} \boxed{H} \xrightarrow{\mathbf{y}(k)}$$

Linear Time-Invariant filter in state-space representation:

$$H \begin{cases} \mathbf{x}(k+1) & = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) & = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \end{cases}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{B} \in \mathbb{R}^{n \times q}, \mathbf{C} \in \mathbb{R}^{p \times n}, \mathbf{D} \in \mathbb{R}^{p \times q}$

Bounded-Input Bounded-Output (BIBO) stability:

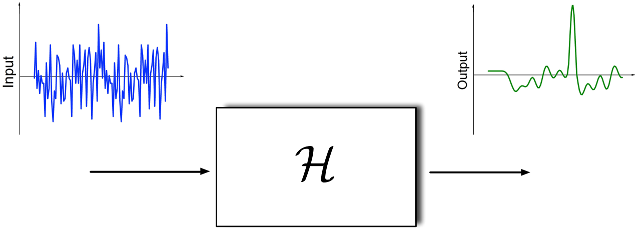
$$\rho(\mathbf{A}) < 1$$

Worst-Case Peak Gain: Definitions

Definition

Worst-case peak gain (WCPG) $\mathbf{W}$ is the largest possible peak value of the output $\mathbf{y}(k)$ over all possible inputs $\mathbf{u}(k)$:

$$\mathbf{W} := |\mathbf{D}| + \sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$



Worst-Case Peak Gain: Motivation

WCPG is required:

- To measure how the computational errors in the implemented filter are propagated to the output
- To measure the magnitude of each variable for implementations in fixed-point arithmetic

Worst-Case Peak Gain: Motivation

WCPG is required:

- To measure how the computational errors in the implemented filter are propagated to the output
- To measure the magnitude of each variable for implementations in fixed-point arithmetic

Goal:

Given a small $\varepsilon > 0$ compute a floating-point approximation $\mathbf{S}$ on the WCPG such that element-by-element

$$|\mathbf{W} - \mathbf{S}| < \varepsilon$$

Outline

Worst-Case Peak Gain

$$\mathbf{W} = |\mathbf{D}| + \sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

Worst-Case Peak Gain

$$\mathbf{W} = |\mathbf{D}| + \sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$

- Cannot sum infinitely $\Longrightarrow$ need to truncate the sum

Worst-Case Peak Gain

$$\mathbf{W} = |\mathbf{D}| + \sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k \mathbf{B} \right|$$

- Cannot sum infinitely $\Longrightarrow$ need to truncate the sum
- 6 sources of errors $\Longrightarrow$ allocate 6 "buckets" ε_i out of the error budget ε

Step 1

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$

Step 1

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right| \longrightarrow \sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$

Step 1

$$\left| \sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| - \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| \right| \leq \varepsilon_1$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

Step 1 Compute an approximate lower bound on truncation order N such that the truncation error is smaller than ε_1.

Step 1

$$\left| \sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| - \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| \right| \leq \varepsilon_1$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| \downarrow \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

Step 1 Compute an approximate lower bound on truncation order N such that the truncation error is smaller than ε_1.

Lower bound on truncation order N

$$N \geq \left\lceil \frac{\log \frac{\varepsilon_1}{\|\mathbf{M}\|_{min}}}{\log \rho(\mathbf{A})} \right\rceil \quad \text{with} \quad \mathbf{M} := \sum_{l=1}^{n} \frac{|\mathbf{R}_l|}{1 - |\boldsymbol{\lambda}_l|} \frac{|\boldsymbol{\lambda}_l|}{\rho(\mathbf{A})}$$

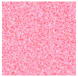
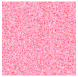
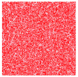
Step 2

$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$

Step 2

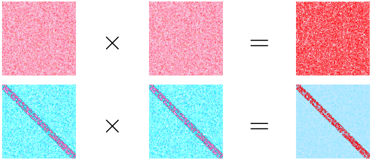
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

 $\times$  $=$  cancellation

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left| \mathbf{C} {\color{red}\mathbf{A}^{k}} \mathbf{B} \right|$$



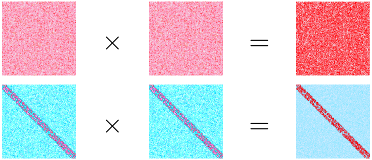
cancellation

less cancellation

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}{\color{red}\mathbf{A}^{k}}\mathbf{B} \right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$



cancellation

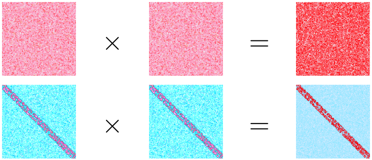
less cancellation

$$\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$



cancellation

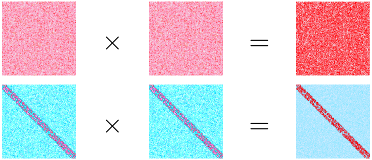
less cancellation

$\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$ $\qquad$ $\mathbf{V} \approx \mathbf{X}$ and $\mathbf{T} \approx \mathbf{E}$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$



cancellation

less cancellation

$$\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1} \qquad \mathbf{V} \approx \mathbf{X} \text{ and } \mathbf{T} \approx \mathbf{E}$$

$$\mathbf{T} \approx \mathbf{V}^{-1} \times \mathbf{A} \times \mathbf{V}$$

$$\downarrow$$

$$\mathbf{A}^k \approx \mathbf{V} \times \mathbf{T}^k \times \mathbf{V}^{-1}$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$

Step 2

$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^{k}\mathbf{B}\right| \longrightarrow \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^{k}\mathbf{V}^{-1}\mathbf{B}\right|$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^{k}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^{k}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^{k}\mathbf{V}^{-1}\mathbf{B}\right|$$

Step 2

$$\left| \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| \quad - \quad \sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right| \right| \leq \varepsilon_2$$

Step 2 Given matrix $\mathbf{V}$ compute $\mathbf{T}$ such that the error of substitution of the product $\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}$ instead of $\mathbf{A}^k$ is less than ε_2.

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$

Step 3

$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^{k}\mathbf{V}^{-1}\mathbf{B} \right|$$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^{k}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^{k}\mathbf{V}^{-1}\mathbf{B} \right|$$
$$\downarrow$$

Step 3

$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right| \longrightarrow \sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$

Step 3

$$\left| \sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B} \right| \quad - \quad \sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{T}^k\mathbf{B}' \right| \right| \leq \varepsilon_3$$

Step 3 Compute the products $\mathbf{CV}$ and $\mathbf{V}^{-1}\mathbf{B}$ such that the propagated error of matrix multiplications is bounded by ε_3.

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{T}^k\mathbf{B}' \right|$$

Step 4

$$\sum_{k=0}^{N} \left| \mathbf{C}' \mathbf{T}^k \mathbf{B}' \right|$$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{T}^k\mathbf{B}' \right|$$
$$\downarrow$$

Step 4

$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right| \longrightarrow \sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$

$$\mathbf{P}_0 := \mathbf{I}$$
$$\mathbf{P}_k := \mathbf{T} \otimes \mathbf{P}_{k-1}$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$

Step 4

$$\left| \sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right| \quad - \quad \sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right| \right| \leq \varepsilon_4$$

$$\mathbf{P}_0 := \mathbf{I}$$
$$\mathbf{P}_k := \mathbf{T} \otimes \mathbf{P}_{k-1}$$

Step 4 Compute the powers $\mathbf{P}_k$ of matrix $\mathbf{T}$ such that the propagated error of matrix multiplications is bounded by ε_4.

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$

Step 5

$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$
$$\downarrow$$

Step 5

$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right| \longrightarrow \sum_{k=0}^{N} \left|\mathbf{L}_k\right|$$

$$\mathbf{L}_k := \mathbf{C}' \otimes (\mathbf{P}_k \otimes \mathbf{B}')$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{L}_k\right|$$

Step 5

$$\left| \sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{P}_k\mathbf{B}' \right| \quad - \quad \sum_{k=0}^{N} \left| \mathbf{L}_k \right| \right| \leq \varepsilon_5$$

$$\mathbf{L}_k := \mathbf{C}' \otimes (\mathbf{P}_k \otimes \mathbf{B}')$$

Step 5 Compute on each step the matrix product $\mathbf{C}'\mathbf{T}^k\mathbf{B}'$ such the overall error of these multiplications on each step is bounded by ε_5.

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{T}^k\mathbf{B}' \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{P}_k\mathbf{B}' \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{L}_k \right|$$

Step 6

$$\sum_{k=0}^{N} \left| \mathbf{L}_k \right|$$

$$\sum_{k=0}^{\infty} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B} \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{T}^k\mathbf{B}' \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{C}'\mathbf{P}_k\mathbf{B}' \right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left| \mathbf{L}_k \right|$$
$$\downarrow$$

Step 6

$$\sum_{k=0}^{N} \left|\mathbf{L}_k\right| \longrightarrow \mathbf{S}_N$$

$$\mathbf{S}_k := \mathbf{S}_{k-1} \oplus \left|\mathbf{L}_k\right|$$

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{L}_k\right|$$
$$\downarrow$$
$$\mathbf{S}_N$$

Step 6

$$\left| \sum_{k=0}^{N} |\mathbf{L}_k| \quad - \quad \mathbf{S}_N \right| \leq \varepsilon_6$$

$$\mathbf{S}_k := \mathbf{S}_{k-1} \oplus |\mathbf{L}_k|$$

Step 6 Compute the absolute value of matrix and accumulate it in the result such that the error is bounded by ε_6.

$$\sum_{k=0}^{\infty} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{T}^k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} \left|\mathbf{C}'\mathbf{P}_k\mathbf{B}'\right|$$
$$\downarrow$$
$$\sum_{k=0}^{N} |\mathbf{L}_k|$$
$$\downarrow$$
$$\mathbf{S}_N$$

Taking $\varepsilon_i = \frac{1}{6}\varepsilon$ we obtain that $\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_6 \leq \varepsilon$ hence the overall error bound is satisfied.

A floating-point evaluation of the WCPG:

Step 1: Compute N

Step 2: Compute $\mathbf{V}$
$\mathbf{T} \leftarrow \text{inv}(\mathbf{V}) \otimes (\mathbf{A} \otimes \mathbf{V})$

Step 3: $\mathbf{B}' \leftarrow \text{inv}(\mathbf{V}) \otimes \mathbf{B}$
$\mathbf{C}' \leftarrow \mathbf{C} \otimes \mathbf{V}$
$\mathbf{S}_{-1} \leftarrow |\mathbf{D}|$, $\mathbf{P}_{-1} \leftarrow \mathbf{I}_n$
for k **from** 0 **to** N **do:**

Step 4: $\mathbf{P}_k \leftarrow \mathbf{T} \otimes \mathbf{P}_{k-1}$

Step 5: $\mathbf{L}_k \leftarrow \mathbf{C}' \otimes \left(\mathbf{P}_k \otimes \mathbf{B}'\right)$

Step 6: $\mathbf{S}_k \leftarrow \mathbf{S}_{k-1} \oplus \text{abs}(\mathbf{L}_k)$

end for

Outline

Basic bricks

Requirement:

Provide matrix operations which satisfy an element-by-element absolute error bound δ given in the argument.

Basic bricks

Requirement:

Provide matrix operations which satisfy an element-by-element absolute error bound δ given in the argument.

Problem:

In fixed-precision FP arithmetic such absolute bound is not generally possible.

Basic bricks

Requirement:

Provide matrix operations which satisfy an element-by-element absolute error bound δ given in the argument.

Problem:

In fixed-precision FP arithmetic such absolute bound is not generally possible.

Solution:

Use multiple-precision FP arithmetic and dynamically adapt precision of the result variables.

Basic bricks

- multiplyAndAdd($\mathbf{A}, \mathbf{B}, \mathbf{C}, \delta$): for $\mathbf{A} \in \mathbb{C}^{p\times n}$, $\mathbf{B} \in \mathbb{C}^{n\times q}$, $\mathbf{C} \in \mathbb{C}^{p\times q}$, computes a matrix $\mathbf{D} \in \mathbb{C}^{p\times q}$ such that

$$\mathbf{D} = \mathbf{A} \cdot \mathbf{B} + \mathbf{C} + \boldsymbol{\Delta},$$

  where the error-matrix $\boldsymbol{\Delta}$ is bounded by $|\boldsymbol{\Delta}| < \delta$, for a certain scalar absolute error bound δ, given in argument to the algorithm.

The algorithm performs an error-free scalar multiplication and uses a modified software-implemented Kulisch-like accumulator.

Basic bricks

- sumAbs($\mathbf{A}, \mathbf{B}, \delta$): for $\mathbf{A} \in \mathbb{R}^{p \times n}$, $\mathbf{B} \in \mathbb{C}^{p \times n}$, computes a matrix $\mathbf{C} \in \mathbb{R}^{p \times n}$ such that

$$\mathbf{C} = \mathbf{A} + |\mathbf{B}| + \mathbf{\Delta},$$

where the error matrix $\mathbf{\Delta}$ is bounded by $|\mathbf{\Delta}| < \delta$, for a certain scalar absolute error bound δ, given in argument to the algorithm.

Basic bricks

- inv(V, δ): for a complex square matrix $\mathbf{V} \in \mathbb{C}^{n \times n}$, computes a matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ such that

$$\mathbf{U} = \mathbf{V}^{-1} + \mathbf{\Delta},$$

where the error matrix $\mathbf{\Delta}$ is bounded by $|\mathbf{\Delta}| < \delta$, for a certain scalar absolute error bound δ, given in argument to the algorithm.

The algorithm is based on Newton-Raphson matrix iteration, requires a seed matrix in argument and works on certain conditions, easily verified in our case.

Basic bricks

- frobeniusNormUpperBound($\mathbf{A}, \delta$): for $\mathbf{A} \in \mathbb{C}^{p \times n}$ computes f an upper bound on the Frobenius norm of $\mathbf{A}$ such that

$$f = \|\mathbf{A}\|_F + \gamma$$

where $0 \leq \gamma < \delta$, for a certain scalar absolute error bound δ, given in argument to the algorithm.

Outline

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n = 10$,	$p = 11$,	$q = 1$	$n = 12$,	$p = 1$,	$q = 25$
$1 - \rho(\mathbf{A})$		1.39×10^{-2}			8.65×10^{-3}	
$\max(\mathbf{S}_N)$		3.88×10^{1}			5.50×10^{9}	
$\min(\mathbf{S}_N)$		1.29×10^{0}			1.0×10^{0}	
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n = 10$,	$p = 11$,	$q = 1$	$n = 12$,	$p = 1$,	$q = 25$
$1 - \rho(\mathbf{A})$		1.39×10^{-2}			8.65×10^{-3}	
$\max(\mathbf{S}_N)$		3.88×10^{1}			5.50×10^{9}	
$\min(\mathbf{S}_N)$		1.29×10^{0}			1.0×10^{0}	
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n=10$,	$p=11$,	$q=1$	$n=12$,	$p=1$,	$q=25$
$1-\rho(\mathbf{A})$		1.39×10^{-2}			8.65×10^{-3}	
$\max(\mathbf{S}_N)$		3.88×10^{1}			5.50×10^{9}	
$\min(\mathbf{S}_N)$		1.29×10^{0}			1.0×10^{0}	
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n = 10$,	$p = 11$,	$q = 1$	$n = 12$,	$p = 1$,	$q = 25$
$1 - \rho(\mathbf{A})$		1.39×10^{-2}			8.65×10^{-3}	
$\max(\mathbf{S}_N)$		3.88×10^{1}			5.50×10^{9}	
$\min(\mathbf{S}_N)$		1.29×10^{0}			1.0×10^{0}	
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n=10$,	$p=11$,	$q=1$	$n=12$,	$p=1$,	$q=25$
$1-\rho(\mathbf{A})$		1.39×10^{-2}			8.65×10^{-3}	
$\max(\mathbf{S}_N)$		3.88×10^{1}			5.50×10^{9}	
$\min(\mathbf{S}_N)$		1.29×10^{0}			1.0×10^{0}	
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Examples

Example 1: comes from Control Theory, describes a controller of vehicle longitudinal oscillation

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
sizes n, p and q	$n=10$,	$p=11$,	$q=1$	$n=12$,	$p=1$,	$q=25$
$1-\rho(\mathbf{A})$	1.39×10^{-2}			8.65×10^{-3}		
$\max(\mathbf{S}_N)$	3.88×10^{1}			5.50×10^{9}		
$\min(\mathbf{S}_N)$	1.29×10^{0}			1.0×10^{0}		
ε	2^{-5}	2^{-53}	2^{-600}	2^{-5}	2^{-53}	2^{-600}
N	220	2153	29182	308	4141	47811
Inversion iterations	0	2	4	2	3	5
overall max precision (bits)	212	293	1401	254	355	1459
$\mathbf{V}^{-1}$ max precision (bits)	106	173	727	148	204	756
$\mathbf{P}_N$ max precision (bits)	64	84	639	64	86	640
$\mathbf{S}_N$ max precision (bits)	64	79	630	64	107	658
Overall execution time (sec)	0.11	1.53	60.06	0.85	11.54	473.20

Conclusion and Perspectives

Conclusion

- Rigorous evaluation of the WCPG matrix
- Direct formula for truncation order determination
- Implementation of a library in C

Perspectives

- Use a multiprecision eigensolver
- Formalize proofs in a Formal Proof Checker
- Other measures for filter analysis

Thank you!

Questions?

L_2-norm evaluation

Another related problem is the reliable evaluation of the L_2-norm. If $\mathbf{H}$ is a transfer function, then its L_2-norm is defined by

$$\|\mathbf{H}\|_2 \triangleq \sqrt{\frac{1}{2\pi}\int_0^{2\pi} \|\mathbf{H}(e^{j\omega})\|_F^2 \, d\omega}$$

Parseval's theorem gives another expression when $\mathbf{H}$ is described with state-space matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$:

$$\begin{aligned} \|\mathbf{H}\|_2 &= \sqrt{tr(\mathbf{C}\mathbf{W}_c\mathbf{C}^\top + \mathbf{D}\mathbf{D}^\top)} \\ &= \sqrt{tr(\mathbf{B}^\top\mathbf{W}_o\mathbf{B} + \mathbf{D}^\top\mathbf{D})} \end{aligned}$$

where $\mathbf{W}_c$ and $\mathbf{W}_o$ are the controllability and observability Gramians of the system.

Gramians

- $\mathbf{W}_c$ is the controllability Gramian of the system.

$$\mathbf{W}_c \triangleq \sum_{k=0}^{\infty} (\mathbf{A}^k \mathbf{B})(\mathbf{A}^k \mathbf{B})^\top$$

  $\mathbf{W}_c$ is the solution of the discrete-time Lyapunov equation

$$\mathbf{W}_c = \mathbf{A}\mathbf{W}_c\mathbf{A}^\top + \mathbf{B}\mathbf{B}^\top$$

- $\mathbf{W}_o$ is the observability Gramian of the system.

$$\mathbf{W}_o \triangleq \sum_{k=0}^{\infty} (\mathbf{C}\mathbf{A}^k)^\top (\mathbf{C}\mathbf{A}^k)$$

  $\mathbf{W}_o$ is the solution of the discrete-time Lyapunov equation

$$\mathbf{W}_o = \mathbf{A}^\top\mathbf{W}_o\mathbf{A} + \mathbf{C}^\top\mathbf{C}$$

Computation of the Gramians

The Gramians are usually computed by solving the discrete-time Lyapunov equation $\mathbf{X} = \mathbf{A}\mathbf{X}\mathbf{A}^\top + \mathbf{Q}$

The following methods can be used:

- solve $(\mathbf{I} - \mathbf{A} \otimes \mathbf{A})\mathbf{x} = \mathbf{q}$
  where $\mathbf{x} = Vec(\mathbf{X})$ and $\mathbf{q} = Vec(\mathbf{Q})$
  $\rightarrow$ *numerically inefficient*
- use infinite sum $\displaystyle\sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{Q} \mathbf{A}^{k\top}$
  $\rightarrow$ *may required a lot of computation*
- use Hammarling's method, based on Schur decomposition of matrix $\mathbf{A}$
  $\rightarrow$ *efficient, but required a deep analysis of the computational errors of the algorithm*

see "*Computational methods for linear matrix equations*", V. Simoncini

Reliable computation of the L_2-norm

Questions

- How to have a reliable evaluation of the L_2-norm in multiple precision
- How to proceed when $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are interval matrices (small radii, containing previously computed errors)

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

$$\sum_{k>N} \left| \mathbf{CA}^k \mathbf{B} \right|$$

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

$$\sum_{k>N} \left| \mathbf{CA}^k\mathbf{B} \right|$$

Suppose $\mathbf{A} = \mathbf{XEX}^{-1}$, where $\mathbf{E} = diag(\lambda_1, \ldots, \lambda_n)$ is the eigenvalue matrix and $\mathbf{X}$ is the eigenvector matrix. Then,

$$\mathbf{CA}^k\mathbf{B} = \mathbf{CXE}^k\mathbf{X}^{-1}\mathbf{B} = \sum_{l=1}^{n} \mathbf{R}_l \boldsymbol{\lambda}_l^k$$

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

$$\sum_{k>N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right|$$

Suppose $\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$, where $\mathbf{E} = diag(\lambda_1, \ldots, \lambda_n)$ is the eigenvalue matrix and $\mathbf{X}$ is the eigenvector matrix. Then,

$$\mathbf{C}\mathbf{A}^k\mathbf{B} = \mathbf{C}\mathbf{X}\mathbf{E}^k\mathbf{X}^{-1}\mathbf{B} = \sum_{l=1}^{n} \mathbf{R}_l \lambda_l^k$$

Bound on truncation error

$$\sum_{k>N} \left|\mathbf{C}\mathbf{A}^k\mathbf{B}\right| \leq \rho(\mathbf{A})^{N+1}\mathbf{M}$$

$$\mathbf{M} := \sum_{l=1}^{n} \frac{|\mathbf{R}_l|}{1 - |\boldsymbol{\lambda}_l|} \frac{|\boldsymbol{\lambda}_l|}{\rho(\mathbf{A})}$$

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

$$\sum_{k>N} \left| \mathbf{C}\mathbf{A}^k\mathbf{B} \right|$$

Suppose $\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$, where $\mathbf{E} = diag(\lambda_1, \ldots, \lambda_n)$ is the eigenvalue matrix and $\mathbf{X}$ is the eigenvector matrix. Then,

$$\mathbf{C}\mathbf{A}^k\mathbf{B} = \mathbf{C}\mathbf{X}\mathbf{E}^k\mathbf{X}^{-1}\mathbf{B} = \sum_{l=1}^{n} \mathbf{R}_l \lambda_l^k$$

Bound on truncation error

$$\rho(\mathbf{A})^{N+1}\mathbf{M} \overset{!}{\leq} \varepsilon_1$$

$$\mathbf{M} := \sum_{l=1}^{n} \frac{|\mathbf{R}_l|}{1 - |\boldsymbol{\lambda}_l|} \frac{|\boldsymbol{\lambda}_l|}{\rho(\mathbf{A})}$$

Step 1. Bound on truncation order

Lower bound on truncation order

$$N \geq \left\lceil \frac{\log \frac{\varepsilon_1}{m}}{\log \rho(\mathbf{A})} \right\rceil$$

$$\mathbf{M} := \sum_{l=1}^{n} \frac{|\mathbf{R}_l|}{1 - |\boldsymbol{\lambda}_l|} \frac{|\boldsymbol{\lambda}_l|}{\rho(\mathbf{A})}$$

where m is defined as $m := \min_{i,j} |\mathbf{M}_{i,j}|$.

Step 1. Bound on truncation order

Lower bound on truncation order

$$N \geq \left\lceil \frac{\log \frac{\varepsilon_1}{m}}{\log \rho(\mathbf{A})} \right\rceil$$

$$\mathbf{M} := \sum_{l=1}^{n} \frac{|\mathbf{R}_l|}{1 - |\boldsymbol{\lambda}_l|} \frac{|\boldsymbol{\lambda}_l|}{\rho(\mathbf{A})}$$

where m is defined as $m := \min_{i,j} |\mathbf{M}_{i,j}|$.

Reliable evaluation

Interval Arithmetic and Rump's Theory of Verified Inclusions are used to determine a rigorous bound of N.

Step 2. "Diagonalization" of matrix $\mathbf{A}$

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \mathbf{\Delta}_2$$

Step 2. "Diagonalization" of matrix **A**

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \boldsymbol{\Delta}_2$$

- **V** is some approximation on **X**
- $\boldsymbol{\Delta}_2$ represents the element-by-element errors due to the two matrix multiplications and the inversion of matrix **V**

Step 2. "Diagonalization" of matrix $\mathbf{A}$

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \mathbf{\Delta}_2$$

- $\mathbf{V}$ is some approximation on $\mathbf{X}$
- $\mathbf{\Delta}_2$ represents the element-by-element errors due to the two matrix multiplications and the inversion of matrix $\mathbf{V}$
- $\mathbf{T}$ diagonal in dominant with very small other elements
- $\|\mathbf{T}\|_2 \leq 1$

Step 2. "Diagonalization" of matrix $\mathbf{A}$

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \boldsymbol{\Delta}_2$$
$$\mathbf{A}^k = \mathbf{V}(\mathbf{T} + \boldsymbol{\Delta}_2)^k \mathbf{V}^{-1}$$

The error of substitution of $\mathbf{A}$ by $\mathbf{V}\mathbf{T}\mathbf{V}^{-1}$:

$$\sqrt{n}(N+1)(N+2)\left\|\boldsymbol{\Delta}_2\right\|_F \left\|\mathbf{C}\mathbf{V}\right\|_F \left\|\mathbf{V}^{-1}\mathbf{B}\right\|_F$$

Step 2. "Diagonalization" of matrix $\mathbf{A}$

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \boldsymbol{\Delta}_2$$

$$\mathbf{A}^k = \mathbf{V}(\mathbf{T} + \boldsymbol{\Delta}_2)^k\mathbf{V}^{-1}$$

The error of substitution of $\mathbf{A}$ by $\mathbf{V}\mathbf{T}\mathbf{V}^{-1}$:

$$\sqrt{n}(N+1)(N+2)\left\|\boldsymbol{\Delta}_2\right\|_F \left\|\mathbf{C}\mathbf{V}\right\|_F \left\|\mathbf{V}^{-1}\mathbf{B}\right\|_F \quad \overset{!}{\leq} \quad \varepsilon_2$$

Step 2. "Diagonalization" of matrix $\mathbf{A}$

$$\mathbf{T} := \mathbf{V}^{-1}\mathbf{A}\mathbf{V} - \mathbf{\Delta}_2$$
$$\mathbf{A}^k = \mathbf{V}(\mathbf{T} + \mathbf{\Delta}_2)^k\mathbf{V}^{-1}$$

The error of substitution of $\mathbf{A}$ by $\mathbf{V}\mathbf{T}\mathbf{V}^{-1}$:

$$\sqrt{n}(N+1)(N+2)\left\|\mathbf{\Delta}_2\right\|_F \left\|\mathbf{C}\mathbf{V}\right\|_F \left\|\mathbf{V}^{-1}\mathbf{B}\right\|_F \quad \overset{!}{\leq} \quad \varepsilon_2$$

A condition on the error-matrix $\mathbf{\Delta}_2$:

$$\left\|\mathbf{\Delta}_2\right\|_F \leq \frac{1}{\sqrt{n}(N+1)(N+2)} \frac{\varepsilon_2}{\left\|\mathbf{C}\mathbf{V}\right\|_F \left\|\mathbf{V}^{-1}\mathbf{B}\right\|_F}$$

Step 3. Computing products $\mathbf{C}'$ and $\mathbf{B}'$

$$\mathbf{C}' := \mathbf{C}\mathbf{V} + \mathbf{\Delta}_{3_C}$$
$$\mathbf{B}' := \mathbf{V}^{-1}\mathbf{B} + \mathbf{\Delta}_{3_B}$$

where $\mathbf{\Delta}_{3_C} \in \mathbb{C}^{p\times n}$ and $\mathbf{\Delta}_{3_B} \in \mathbb{C}^{n\times q}$ are error-matrices.

Bound on the multiplication errors $\mathbf{\Delta}_{3_C}$ and $\mathbf{\Delta}_{3_B}$:

$$\|\mathbf{\Delta}_{3_C}\|_F \leq \frac{1}{3\sqrt{n}} \cdot \frac{1}{N+1} \frac{\varepsilon_3}{\|\mathbf{C}'\|_F}$$
$$\|\mathbf{\Delta}_{3_B}\|_F \leq \frac{1}{3\sqrt{n}} \cdot \frac{1}{N+1} \frac{\varepsilon_3}{\|\mathbf{B}'\|_F}.$$

Step 4. Powering $\mathbf{T}$

$$\mathbf{P}_k := \mathbf{T}^k - \mathbf{\Delta}_{4_k}$$

$\mathbf{\Delta}_{4_k} \in \mathbb{C}^{n\times n}$ error-matrix on matrix powers, including error propagation from the first to the last power.

$$\mathbf{P}_k = \mathbf{T}\mathbf{P}_{k-1} + \mathbf{\Gamma}_k,$$

where $\mathbf{\Gamma}_k \in \mathbb{C}^{n\times n}$ is the error-matrix on the error of the matrix multiplication at step k.

Bound on the error-matrix $\mathbf{\Gamma}_k$

$$\|\mathbf{\Gamma}_k\|_F \leq \frac{1}{\sqrt{n}} \cdot \frac{1}{N-1} \cdot \frac{1}{N+1} \cdot \frac{\varepsilon_4}{\|\mathbf{C}'\|_F \, \|\mathbf{B}'\|_F}$$

Step 5. Computing L_k

$$\mathbf{L}_k := \mathbf{C}'\mathbf{P}_k\mathbf{B}' + \mathbf{\Delta}_{5_k},$$

where $\mathbf{\Delta}_{5_k} \in \mathbb{C}^{p\times q}$ is the matrix of element-by-element errors for the two matrix multiplications.

Bound on the error-matrix $\mathbf{\Delta}_{5_k}$

$$|\mathbf{\Delta}_{5_k}| \leq \frac{1}{N+1} \cdot \varepsilon_5.$$

Step 6. Summation

$$S_N = |\mathbf{D}| + \sum_{l=0}^{N} |\mathbf{L}_l| + \mathbf{\Delta}_6,$$

where the error-matrix $\mathbf{\Delta}_6 \in \mathbb{C}^{p \times q}$ represents the error of $N+1$ absolute value accumulations.

Bound on the error matrix $\mathbf{\Delta}_{6_k}$

$$\mathbf{\Delta}_{6_k} \leq \frac{1}{N}\varepsilon_6, \qquad k = 1 \ldots N$$