

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

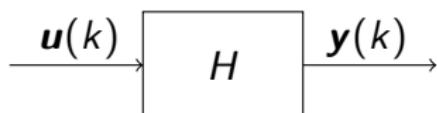
Anastasia Volkova, Thibault Hilaire, Christoph Lauter

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

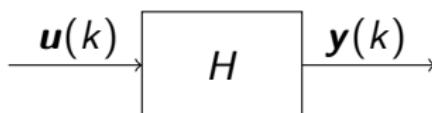
22nd IEEE Symposium on Computer Arithmetic
June 23, 2015



Digital filters



Digital filters

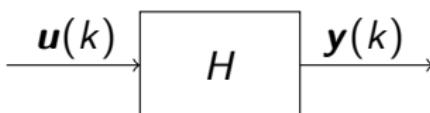


Linear Time-Invariant filter in state-space representation:

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

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



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:

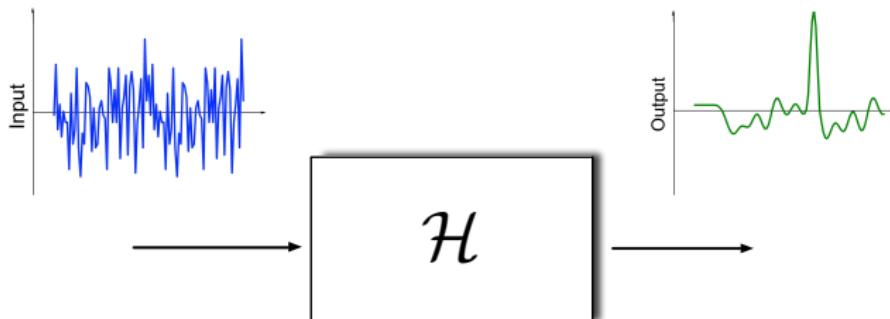
$$\rho(\mathbf{A}) < 1$$

Worst-Case Peak Gain: Definitions

Definition

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

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



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

- 1 Problem statement
- 2 Algorithm of WCPG evaluation
- 3 Basic bricks
- 4 Numerical Examples
- 5 Conclusion

Worst-Case Peak Gain

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

Worst-Case Peak Gain

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

- Cannot sum infinitely \implies need to truncate the sum

Worst-Case Peak Gain

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

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

Step 1

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

Step 1

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

Step 1

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

$$\begin{array}{c} \sum_{k=0}^{\infty} |\mathbf{C}\mathbf{A}^k \mathbf{B}| \\ \downarrow \\ \sum_{k=0}^N |\mathbf{C}\mathbf{A}^k \mathbf{B}| \end{array}$$

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

Step 1

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

$$\begin{array}{c} \sum_{k=0}^{\infty} |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ \downarrow \\ \sum_{k=0}^N |\mathbf{C}\mathbf{A}^k\mathbf{B}| \end{array}$$

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

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

Step 2

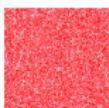
$$\sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}|$$

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$

Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$


$$\times =$$


cancellation

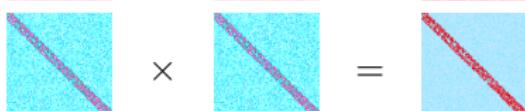
Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$


$$\begin{matrix} \times & = & \end{matrix}$$

cancellation


$$\begin{matrix} \times & = & \end{matrix}$$

less cancellation

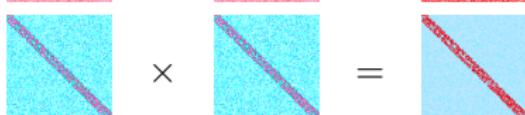
Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$


$$\begin{matrix} \times & = & \end{matrix}$$

cancellation


$$\begin{matrix} \times & = & \end{matrix}$$

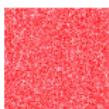
less cancellation

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

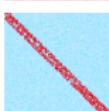
Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$


$$\times =$$


cancellation


$$\times =$$


less cancellation

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

$$\mathbf{V} \approx \mathbf{X} \text{ and } \mathbf{T} \approx \mathbf{E}$$

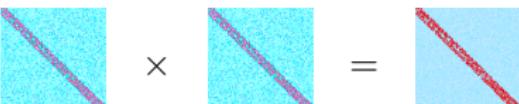
Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \end{aligned}$$



cancellation



less cancellation

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

$$\mathbf{V} \approx \mathbf{X} \text{ and } \mathbf{T} \approx \mathbf{E}$$

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

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

Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \end{aligned}$$

Step 2

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{V} \mathbf{T}^k \mathbf{V}^{-1} \mathbf{B}| \end{aligned}$$

Step 2

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

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 .

$$\begin{array}{c} \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ \downarrow \\ \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ \downarrow \\ \sum_{k=0}^N |\mathbf{C} \mathbf{V} \mathbf{T}^k \mathbf{V}^{-1} \mathbf{B}| \end{array}$$

Step 3

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}| \end{aligned}$$

Step 3

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}'\mathbf{T}^k\mathbf{B}'| \end{aligned}$$

Step 3

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}\mathbf{V}\mathbf{T}^k\mathbf{V}^{-1}\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}'\mathbf{T}^k\mathbf{B}'| \end{aligned}$$

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

Step 4

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{A}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C} \mathbf{V} \mathbf{T}^k \mathbf{V}^{-1} \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}' \mathbf{T}^k \mathbf{B}'| \\ & \downarrow \end{aligned}$$

Step 4

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

$$\mathbf{P}_0 := \mathbf{I}$$

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

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

Step 4

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

$$\mathbf{P}_0 := \mathbf{I}$$

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

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

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 .

Step 5

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

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

Step 5

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

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

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

Step 5

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

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

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 .

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

Step 6

$$\sum_{k=0}^N |\mathcal{L}_k|$$

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathcal{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}\mathbf{V}\mathcal{T}^k\mathbf{V}^{-1}\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}'\mathcal{T}^k\mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}'\mathcal{P}_k\mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{L}_k| \end{aligned}$$

Step 6

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

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathcal{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}\mathbf{A}^k\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}\mathbf{V}\mathcal{T}^k\mathbf{V}^{-1}\mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}'\mathcal{T}^k\mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{C}'\mathcal{P}_k\mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathcal{L}_k| \\ & \downarrow \\ & \mathcal{S}_N \end{aligned}$$

Step 6

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

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

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

$$\begin{aligned} & \sum_{k=0}^{\infty} |\mathbf{CA}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{CA}^k \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{CV}\mathbf{T}^k \mathbf{V}^{-1} \mathbf{B}| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}'\mathbf{T}^k \mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{C}'\mathbf{P}_k \mathbf{B}'| \\ & \downarrow \\ & \sum_{k=0}^N |\mathbf{L}_k| \\ & \downarrow \\ & \mathbf{S}_N \end{aligned}$$

Taking $\varepsilon_i = \frac{1}{6}\varepsilon$ we obtain that $\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_6 \leq \varepsilon$ 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: $\mathbf{L}_k \leftarrow \mathbf{C}' \otimes (\mathbf{P}_k \otimes \mathbf{B}')$

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

end for

Outline

- 1 Problem statement
- 2 Algorithm of WCPG evaluation
- 3 Basic bricks
- 4 Numerical Examples
- 5 Conclusion

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

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

where the error-matrix Δ is bounded by $|\Delta| < \delta$, 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

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

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

Basic bricks

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

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

where the error matrix Δ is bounded by $|\Delta| < \delta$, 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(A, δ)`: for $A \in \mathbb{C}^{p \times n}$ computes f an upper bound on the Frobenius norm of A such that

$$f = \|A\|_F + \gamma$$

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

Outline

- 1 Problem statement
- 2 Algorithm of WCPG evaluation
- 3 Basic bricks
- 4 Numerical Examples
- 5 Conclusion

Examples

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

Example 2: 12th-order Butterworth filter

	Example 1			Example 2		
	$n = 10$, $p = 11$, $q = 1$	1.39×10^{-2}	3.88×10^1	1.29×10^0	$n = 12$, $p = 1$, $q = 25$	8.65×10^{-3}
sizes n , p and q						
$1 - \rho(\mathbf{A})$						
$\max(\mathcal{S}_N)$						
$\min(\mathcal{S}_N)$						
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
\mathcal{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}		
$\max(\mathcal{S}_N)$	3.88×10^1			5.50×10^9		
$\min(\mathcal{S}_N)$	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
\mathcal{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}	
$\max(\mathcal{S}_N)$		3.88×10^1			5.50×10^9	
$\min(\mathcal{S}_N)$		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}	
$\max(\mathcal{S}_N)$		3.88×10^1			5.50×10^9	
$\min(\mathcal{S}_N)$		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}		
$\max(\mathcal{S}_N)$	3.88×10^1			5.50×10^9		
$\min(\mathcal{S}_N)$	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}		
$\max(\mathcal{S}_N)$	3.88×10^1			5.50×10^9		
$\min(\mathcal{S}_N)$	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
\mathcal{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?

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

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

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

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

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

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

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

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

Suppose $\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$, where $\mathbf{E} = \text{diag}(\lambda_1, \dots, \lambda_n)$ 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

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

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

Step 1. Bound on truncation error

Truncation error is the tail of the infinite sum:

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

Suppose $\mathbf{A} = \mathbf{X}\mathbf{E}\mathbf{X}^{-1}$, where $\mathbf{E} = \text{diag}(\lambda_1, \dots, \lambda_n)$ 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

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

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

Step 1. Bound on truncation order

Lower bound on truncation order

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

$$\mathbf{M} := \sum_{I=1}^n \frac{|\mathbf{R}_I|}{1 - |\lambda_I|} \frac{|\lambda_I|}{\rho(\mathbf{A})}$$

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

Step 1. Bound on truncation order

Lower bound on truncation order

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

$$\mathbf{M} := \sum_{I=1}^n \frac{|\mathbf{R}_I|}{1 - |\lambda_I|} \frac{|\lambda_I|}{\rho(\mathbf{A})}$$

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 **A**

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

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

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

- \mathbf{V} is some approximation on \mathbf{X}
- Δ_2 represents the element-by-element errors due to the two matrix multiplications and the inversion of matrix \mathbf{V}

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

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

- \mathbf{V} is some approximation on \mathbf{X}
- Δ_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}

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

$$\mathbf{A}^k = \mathbf{V} (\mathbf{T} + \Delta_2)^k \mathbf{V}^{-1}$$

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

$$\sqrt{n}(N+1)(N+2) \|\Delta_2\|_F \|\mathbf{C}\mathbf{V}\|_F \|\mathbf{V}^{-1}\mathbf{B}\|_F$$

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

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

$$\mathbf{A}^k = \mathbf{V} (\mathbf{T} + \Delta_2)^k \mathbf{V}^{-1}$$

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

$$\sqrt{n}(N+1)(N+2) \|\Delta_2\|_F \|\mathbf{C}\mathbf{V}\|_F \|\mathbf{V}^{-1}\mathbf{B}\|_F \stackrel{!}{\leq} \varepsilon_2$$

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

$$\mathbf{T} := \mathbf{V}^{-1} \mathbf{A} \mathbf{V} - \Delta_2$$

$$\mathbf{A}^k = \mathbf{V} (\mathbf{T} + \Delta_2)^k \mathbf{V}^{-1}$$

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

$$\sqrt{n}(N+1)(N+2) \|\Delta_2\|_F \|\mathbf{C}\mathbf{V}\|_F \|\mathbf{V}^{-1}\mathbf{B}\|_F \stackrel{!}{\leq} \varepsilon_2$$

A condition on the error-matrix Δ_2 :

$$\|\Delta_2\|_F \leq \frac{1}{\sqrt{n}(N+1)(N+2)} \frac{\varepsilon_2}{\|\mathbf{C}\mathbf{V}\|_F \|\mathbf{V}^{-1}\mathbf{B}\|_F}$$

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

$$\mathbf{C}' := \mathbf{C}\mathbf{V} + \Delta_{3_C}$$

$$\mathbf{B}' := \mathbf{V}^{-1}\mathbf{B} + \Delta_{3_B}$$

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

Bound on the multiplication errors Δ_{3_C} and Δ_{3_B} :

$$\|\Delta_{3_C}\|_F \leq \frac{1}{3\sqrt{n}} \cdot \frac{1}{N+1} \frac{\varepsilon_3}{\|\mathbf{C}'\|_F}$$

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

Step 4. Powering \mathbf{T}

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

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

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

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

Bound on the error-matrix Γ_k

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

Step 5. Computing L_k

$$L_k := C' P_k B' + \Delta_{5_k},$$

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

Bound on the error-matrix Δ_{5_k}

$$|\Delta_{5_k}| \leq \frac{1}{N+1} \cdot \varepsilon_5.$$

Step 6. Summation

$$S_N = |\mathbf{D}| + \sum_{l=0}^N |\mathbf{L}_l| + \Delta_6,$$

where the error-matrix $\Delta_6 \in \mathbb{C}^{p \times q}$ represents the error of $N+1$ absolute value accumulations.

Bound on the error matrix Δ_{6_k}

$$\Delta_{6_k} \leq \frac{1}{N} \varepsilon_6, \quad k = 1 \dots N$$