Lab Session 3: Introduction to Interval Arithmetic

Floating-Point Arithmetic and Error Analysis

Alberto Taddei
Giulia Lionetti
Thies Weel
Master 2 CCA, Sorbonne Université

November 2025

1 Exercise 1: Range of a Function

1.1 Objective and Theoretical Background

We evaluate $f(x) = x^2 - 4x$ on $X = [1, 4]$ using three algebraically equivalent formulations to demonstrate the dependency problem in interval arithmetic. By calculus, $f'(x) = 2x - 4 = 0$ at $x = 2$ (minimum), with $f(1) = -3$, $f(2) = -4$, $f(4) = 0$. The **exact range** is $[-4, 0]$.

1.2 Implementation and Results

Using INTLAB, we test:

1. $f_1(x) = x^2 - 4x$ (direct form)
2. $f_2(x) = x(x - 4)$ (factored)
3. $f_3(x) = (x - 2)^2 - 4$ (completed square)

Formulation	**Result**	**Width**	**Accuracy**
$f_1(x) = x^2 - 4x$	$[-15, 12]$	27	Poorest
$f_2(x) = x(x - 4)$	$[-12, 0]$	12	Better
$f_3(x) = (x - 2)^2 - 4$	$[-4, 0]$	4	**Exact**

Table 1: Interval enclosures for $f(X)$. Only f_3 achieves the exact range.

1.3 Analysis

Why different results? When X appears multiple times, interval arithmetic treats each occurrence independently, allowing impossible combinations.

f_1**: Severe overestimation.** Computes $X^2 = [1, 16]$ and $4X = [4, 16]$ independently, yielding $[1 - 16, 16 - 4] = [-15, 12]$. This erroneously allows $x = 1$ for X^2 while $x = 4$ for $4X$.

f_2**: Moderate overestimation.** Computes $[1, 4] \times [-3, 0] = [-12, 0]$. Better because $(X - 4) < 0$ provides constraint, but dependency persists.

f_3**: Optimal.** X appears once: $(X - 2) = [-1, 2]$, so $(X - 2)^2 = [0, 4]$ (squaring interval containing zero gives $[0, \max(1^2, 2^2)]$), thus $(X - 2)^2 - 4 = [-4, 0]$ exactly.

Key insight: As noted in the slides (p.19), "$x - x \neq \{0\}$" in interval arithmetic due to dependency. Minimizing variable occurrences and centering around critical points yields tightest bounds. The completed square eliminates dependency entirely.

2 Exercise 2: Invertibility of a Matrix

2.1 Objective

Prove that for a matrix $A \in \mathbb{R}^{n\times n}$, if there exists $R \in \mathbb{R}^{n\times n}$ such that $\|I - RA\| < 1$, then A is invertible. Implement an algorithm using interval arithmetic to certify invertibility.

2.2 Theoretical Proof

Theorem: If $\|I - RA\| < 1$ for some matrix R, then A is nonsingular.

Proof by contrapositive: Assume A is singular. Then $\exists x \neq 0$ such that $Ax = 0$. This implies:

$$(I - RA)x = x - R(Ax) = x - R \cdot 0 = x$$

Therefore $\|I - RA\| \geq \frac{\|(I-RA)x\|}{\|x\|} = \frac{\|x\|}{\|x\|} = 1$, contradicting $\|I - RA\| < 1$. Hence A must be nonsingular.

2.3 Algorithm

To certify invertibility of A using interval arithmetic:

1. Compute approximate inverse: $R = \text{inv}(A)$ (floating-point)
2. Form interval matrix: $C = I - R \cdot \text{intval}(A)$ (interval arithmetic)
3. Compute norm: $\|C\|_1$ with interval arithmetic
4. If $\sup(\|C\|_1) < 1$, then A is certified invertible

2.4 Implementation and Results

We test on a 3×3 matrix. The interval computation of $I - RA$ yields:

```
intval C =
 [ 1.0000...,  0.0000...,  0.0000...]
 [-0.0000...,  1.0000..., -0.0000...]
 [-0.0000..., -0.0000...,  1.0000...]
```

The matrix is essentially the identity with errors at machine precision level. Computing $\|C\|_1$ in interval arithmetic confirms $\|C\|_1 < 1$, thus **certifying that A is invertible**.

The interval inverse computed by INTLAB is:

```
intval inv(A) =
 [ 1.4642...,  0.2936..., -1.0241...]
 [-0.6886...,  0.9852...,  0.3784...]
 [ 0.6294..., -1.0920...,  0.7641...]
```

Each interval contains the true inverse element, with widths on the order of machine precision, confirming both invertibility and accurate inverse computation.

2.5 Key Insight

This technique provides *rigorous certification* of invertibility, not just numerical evidence. Unlike standard floating-point checks (e.g., testing if $\det(A) \approx 0$), interval arithmetic proves that $\|I - RA\| < 1$ mathematically, guaranteeing nonsingularity regardless of rounding errors. This is essential for applications requiring mathematical certainty (verified computing, formal proofs).

3 Exercise 3: Numerical Solutions of Linear Systems

3.1 Overview and Test Case

We solve $Ax = b$ using interval arithmetic to obtain *guaranteed enclosures* rather than approximate solutions. All methods are tested on the 4×4 Hilbert matrix $H_{ij} = \frac{1}{i+j-1}$ with $b = A \cdot \mathbf{1}$, giving exact solution $x^* = (1, 1, 1, 1)^T$. The Hilbert matrix is notoriously ill-conditioned with $\kappa_1(H_4) \approx 1.55 \times 10^4$. To model input uncertainty, we perturb by $\epsilon = 10^{-14}$:

$$A_{\text{int}} = [A - \epsilon, A + \epsilon], \quad b_{\text{int}} = [b - \epsilon, b + \epsilon]$$

3.2 Exercise 3.1: Interval Gaussian Elimination

3.2.1 Method

Standard GE with partial pivoting (selecting largest midpoint) adapted for interval arithmetic:

1. **Forward elimination:** At step k, pivot to row maximizing $|\text{mid}(A_{pk})|$, then eliminate:

$$m_i = \frac{A_{ik}}{A_{kk}}, \quad A_{i,:} \leftarrow A_{i,:} - m_i A_{k,:}, \quad b_i \leftarrow b_i - m_i b_k$$

2. **Back-substitution:** Solve upper-triangular system from bottom to top

All operations use interval arithmetic, rigorously propagating input uncertainties.

3.2.2 Results

```
xint  [1.0000000000____, 1.0000000000____,
       1.0000000000____, 1.0000000000____]^T
```

Interval widths $\approx 10^{-10}$.

Verification:

- Each interval contains 1: `true`
- $A_{\text{int}} x_{\text{int}} \subseteq b_{\text{int}}$: `true`

3.2.3 Analysis

Despite $\kappa \approx 10^4$, the interval widths remain narrow ($\approx 10^{-10}$), giving error amplification:

$$\frac{\text{output width}}{\text{input width}} = \frac{10^{-10}}{10^{-14}} = 10^4 \approx \kappa$$

This matches theoretical error bounds, confirming that interval arithmetic correctly quantifies ill-conditioning effects. The self-consistency check verifies that our computed intervals are mathematically valid, though it doesn't explicitly prove existence or uniqueness.

3.3 Exercise 3.2: Krawczyk Operator

3.3.1 Theoretical Foundation

Theorem (Krawczyk, 1969): Let $R \in \mathbb{R}^{n \times n}$, I be the identity, and X an interval vector. Define the Krawczyk operator:

$$K(X) := Rb + (I - RA)X$$

If $K(X) \subseteq \text{int}(X)$ (strict interior containment), then:

1. A and R are nonsingular
2. The system $Ax = b$ has a unique solution $x^* \in K(X) \subseteq X$

Why this works: The condition $K(X) \subseteq \text{int}(X)$ means K is a contraction mapping. By Brouwer's fixed-point theorem, there exists a unique $x^* \in X$ where $K(x^*) = x^*$. Expanding this fixed-point condition:

$$x^* = Rb + (I - RA)x^* \quad \Rightarrow \quad RAx^* = Rb \quad \Rightarrow \quad Ax^* = b$$

The fixed point of K is precisely the solution we seek.

3.3.2 Implementation and Results

Using $R = \text{inv}(\text{mid}(A_{\text{int}}))$ and the interval solution X from Exercise 3.1:

```
K(X) = Rb + (I - RA)X
  [1.0000000000____, 1.000000000_____,
   1.000000000_____, 1.000000000_____]^T
```

Verification: $K(X) \subset \text{int}(X)$: `true`
Numerically: $\inf(K(X)_i) > \inf(X_i)$ and $\sup(K(X)_i) < \sup(X_i)$ for all i.

3.3.3 Interpretation

This single verification simultaneously proves:

- **Existence:** A solution exists
- **Uniqueness:** Exactly one solution in X
- **Invertibility:** A is nonsingular
- **Tight bounds:** Solution is enclosed by $K(X) \subseteq X$

Unlike residual-based error estimates requiring separate condition number analysis, Krawczyk provides a *computer-assisted proof* of correctness. Even with rounding errors and ill-conditioning ($\kappa \approx 10^4$), the proof remains rigorous.

3.4 Exercise 3.3: Determinant Inclusion

3.4.1 Method

Gaussian elimination transforms A into upper-triangular form U. For such matrices:

$$\det(A) = (-1)^s \prod_{i=1}^{n} U_{ii}$$

where s counts row swaps. Computing this product in interval arithmetic yields a rigorous determinant enclosure.

3.4.2 Results

Upper-triangular matrix after GE:

```
U  [1.000    0.500    0.333    0.250  ]
    [  ~0     0.0833   0.0889   0.0833 ]
    [  ~0       ~0    -0.00556 -0.00833]
    [  ~0       ~0       ~0     0.000357]
```

Determinant enclosure:

$$\begin{aligned} \det(A) &\in [1.6534 \times 10^{-7}, 1.6534 \times 10^{-7}] \\ \text{Midpoint:} &\quad 1.6534 \times 10^{-7} \\ \text{Width:} &\quad 2.80 \times 10^{-16} \end{aligned}$$

Certification: Since $0 \notin \det(A)$, the matrix is **proven nonsingular**.

3.4.3 Analysis

The tiny determinant ($\approx 10^{-7}$) confirms severe ill-conditioning, yet the narrow interval width ($\approx 10^{-16}$, near machine precision) demonstrates robustness against underflow and error accumulation.

Connection to Exercise 3.1: This determinant test fills the theoretical gap left by Exercise 3.1. Together:

- Ex 3.1: Computes solution interval (if solution exists)
- Ex 3.3: Proves A is invertible (solution does exist)
- Combined: Solution provably exists and lies in computed interval

The upper-triangular form makes invertibility checking trivial: all diagonal entries must exclude zero. For upper-triangular matrices, this is equivalent to $\det \neq 0$ since the determinant is simply the product of diagonal elements.

3.5 Exercise 3.4: Gershgorin-Guided Pivoting

3.5.1 Motivation

Standard partial pivoting (maximum magnitude) ignores diagonal dominance. In ill-conditioned systems, this can cause excessive interval widening. Gershgorin circles provide a stability criterion.

Gershgorin Circle Theorem: For matrix A, define discs:

$$D_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}$$

All eigenvalues satisfy $\lambda \in \bigcup_i D_i$. If $0 \in D_i$ (i.e., $|a_{ii}| < \sum_{j \neq i} |a_{ij}|$), row i lacks diagonal dominance—a warning of potential instability.

3.5.2 Enhanced Algorithm

At GE step k, compute *Gershgorin margins* for submatrix rows:

$$\mu_i = |a_{ii}| - \sum_{j \neq i} |a_{ij}|$$

Pivot to row with maximum μ_i (most diagonally dominant). If $\mu_i \leq 0$ for all rows, weak dominance is detected and corrective pivoting is applied.

3.5.3 Results on Hilbert Matrix

```
Step 1: Discs include 0 → swap for dominance
Step 2: Discs include 0 → swapped rows 3 and 4

Upper-triangular U:
  [1.000   0.500   0.333   0.250 ]
  [  ~0    0.0833  0.0833  0.0750]
  [  ~0      ~0    0.00833 0.0129]
  [  ~0      ~0      ~0   -0.000238]
```

Determinant: $[1.6534 \times 10^{-7}, 1.6534 \times 10^{-7}]$, width 4.46×10^{-16}

3.5.4 Comparison and Trade-offs

Pivoting Strategy	**Width** ($\times 10^{-16}$)	**Robustness**
Standard (Ex 3.3)	2.80	Good for structured systems
Gershgorin (Ex 3.4)	4.46	Better for badly scaled systems

Table 2: Comparison of pivoting strategies on H_4.

For the well-structured Hilbert matrix, both methods succeed. The slightly wider Gershgorin interval reflects conservative pivoting choices. However, in badly scaled or near-singular systems, Gershgorin pivoting prevents catastrophic interval blow-up by avoiding weak pivots, trading slight over-conservatism for guaranteed numerical stability.

3.6 Comparative Summary

Method	**Guarantees**	**Complexity**	**Best For**
Interval GE (3.1)	Solution bounds	$O(n^3)$	Basic enclosure
Krawczyk (3.2)	Exist. + uniqueness	$O(n^3)$	Full certification
Determinant (3.3)	Invertibility	$O(n^3)$	Nonsingularity test
Gershgorin (3.4)	Robust bounds	$O(n^4)$	Ill-conditioned cases

Table 3: Summary of interval methods. All provide rigorous guarantees.

Key insights:

- **Paradigm shift:** From "$\hat{x} \approx x^*$ with unknown error" to "$x^* \in X$ provably"
- **Complementary approaches:** Ex 3.1 + 3.3 together achieve what Ex 3.2 does in one theorem
- **Robustness:** Despite $\kappa \approx 10^4$, all methods produce tight enclosures (widths 10^{-10} to 10^{-16})
- **Computational cost:** Interval operations are $\sim$5-10$\times$ slower than floating-point, but provide mathematical certainty impossible otherwise