Introduction to computer-assisted proofs in nonlinear analysis

Séminaire doctorant du LAMFA

Damien Galant

CERAMATHS/DMATHS
Université Polytechnique Hauts-de-France

Département de Mathématique
Université de Mons
F.R.S.-FNRS Research Fellow

Joint work with Colette De Coster (CERAMATHS/DMATHS) and Christophe Troestler (UMONS)

Wednesday 5 February 2025

A first example

Let us compute $\sin(0)$ and $\sin(\pi)$ using Python.

Image from https://fr.wikipedia.org/wiki/Fichier:Python-logo-notext.svg

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the "scientific notation" on base 2, where both the significand and the exponent are written with a given number of bits.

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the "scientific notation" on base 2, where both the significand and the exponent are written with a given number of bits.

Given a sequence of 64 binary digits

$$b_{63}b_{62}\cdots b_1 b_0,$$

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the “scientific notation” on base 2, where both the significand and the exponent are written with a given number of bits.

Given a sequence of 64 binary digits

$$b_{63}\,b_{62}\cdots b_1\,b_0,$$

one defines the *sign* s by $s := b_{63}$

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the "scientific notation" on base 2, where both the significand and the exponent are written with a given number of bits.

Given a sequence of 64 binary digits

$$b_{63}\, b_{62} \cdots b_1\, b_0,$$

one defines the *sign* s by $s := b_{63}$ and the *biased exponent* e as the integer whose representation in binary is $b_{62} \cdots b_{52}$.

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the "scientific notation" on base 2, where both the significand and the exponent are written with a given number of bits.

Given a sequence of 64 binary digits

$$b_{63}b_{62}\cdots b_1 b_0,$$

one defines the *sign* s by $s := b_{63}$ and the *biased exponent* e as the integer whose representation in binary is $b_{62}\cdots b_{52}$. Then, the floating-point number encoded by b is

$$(-1)^s \cdot (1.b_{51}b_{50}\cdots b_0)_2 \cdot 2^{e-1023}.$$

Floating-point numbers in a nutshell

Rough idea

Floating-point numbers use the “scientific notation” on base 2, where both the significand and the exponent are written with a given number of bits.

Given a sequence of 64 binary digits

$$b_{63} b_{62} \cdots b_1 b_0,$$

one defines the *sign* s by $s := b_{63}$ and the *biased exponent* e as the integer whose representation in binary is $b_{62} \cdots b_{52}$. Then, the floating-point number encoded by b is

$$(-1)^s \cdot (1.b_{51} b_{50} \cdots b_0)_2 \cdot 2^{e-1023}.$$

$\mathbb{F}$: set of finite 64 bit (double precision) floating-point numbers.

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty,$$

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty...$$

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty...$$

yet $0^+ = 0^-$!).

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty...$$

yet $0^+ = 0^-$!). Note: in practice, Python raises a `ZeroDivisionError`.

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty\dots$$

  yet $0^+ = 0^-$!). Note: in practice, Python raises a `ZeroDivisionError`.
- "not-a-number" (`NaN`, e.g. $0/0$);

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty ...$$

  yet $0^+ = 0^-$!). Note: in practice, Python raises a `ZeroDivisionError`.
- "not-a-number" (`NaN`, e.g. $0/0$);
- there is not a well-defined total order!

More about floating-point numbers

Nowadays, most of the floating-point units use the *IEEE 754 standard*.

There are several implementation subtleties:

- representations for $\pm\infty$;
- several representations of 0, depending on the sign (e.g.

$$1/0^+ = +\infty, \qquad 1/0^- = -\infty\ldots$$

  yet $0^+ = 0^-$!). Note: in practice, Python raises a `ZeroDivisionError`.
- "not-a-number" (`NaN`, e.g. $0/0$);
- there is not a well-defined total order!
- etc.

How not to launch a rocket

Figure: An Ariane 5 launcher (click for the video)

Image from https://commons.wikimedia.org/wiki/File:Ariane_5_with_James_Webb_Space_Telescope_Prelaunch_(51773093465).jpg,

video from https://www.youtube.com/watch?v=1qRUFg-Pte0

What happened?

Roughly speaking:

What happened?

Roughly speaking:

- some code that worked perfectly on Ariane 4 was reused in Ariane 5;

What happened?

Roughly speaking:

- some code that worked perfectly on Ariane 4 was reused in Ariane 5;
- a particular quantity was stocked on 16 bits in this code (this is not so much!);

What happened?

Roughly speaking:

- some code that worked perfectly on Ariane 4 was reused in Ariane 5;
- a particular quantity was stocked on 16 bits in this code (this is not so much!);
- at some point, an *overflow* occurred, basically causing the rocket suddenly believing it was horizontal and not vertical, causing the failure of the launch.

What happened?

Roughly speaking:

- some code that worked perfectly on Ariane 4 was reused in Ariane 5;
- a particular quantity was stocked on 16 bits in this code (this is not so much!);
- at some point, an *overflow* occurred, basically causing the rocket suddenly believing it was horizontal and not vertical, causing the failure of the launch.

The failure came *entirely from the program used!*

What happened?

Roughly speaking:

- some code that worked perfectly on Ariane 4 was reused in Ariane 5;
- a particular quantity was stocked on 16 bits in this code (this is not so much!);
- at some point, an *overflow* occurred, basically causing the rocket suddenly believing it was horizontal and not vertical, causing the failure of the launch.

The failure came *entirely from the program used!*

To summarize:

A first (obvious) limitation of numerical computations

$\mathbb{F}$ is **finite**!

Rounding modes

Since $\mathbb{F}$ is finite, not all real numbers may be represented by floating-point numbers.

Rounding modes

Since $\mathbb{F}$ is finite, not all real numbers may be represented by floating-point numbers.

Perhaps worse, even if a, b are floating-point numbers, $a + b$ may not be such a number.

Rounding modes

Since $\mathbb{F}$ is finite, not all real numbers may be represented by floating-point numbers.

Perhaps worse, even if a, b are floating-point numbers, $a + b$ may not be such a number.

There are thus several *rounding modes*, depending on whether the result is to be rounded up, down, towards zero, etc.

Accumulation of round-off errors

The Vancouver stock index

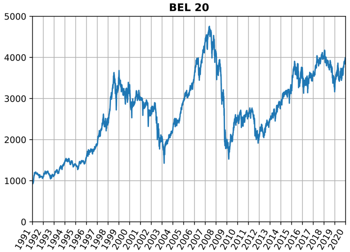


Figure: The BEL20 stock index

Image from https://commons.wikimedia.org/wiki/File:BEL_20.svg

Accumulation of round-off errors

The Vancouver stock index

Between 1982 and 1983, the Vancouver stock index dropped anomalously due to the accumulation of small round-off errors, due to the fact that quantities were always rounded *down* after each computation.

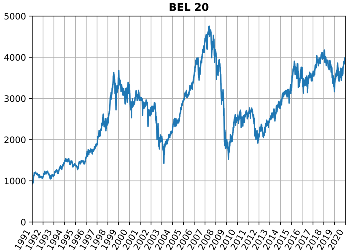


Figure: The BEL20 stock index

Image from https://commons.wikimedia.org/wiki/File:BEL_20.svg

Accumulation of round-off errors

Patriot missiles

Figure: A Patriot missile launch

Image from

https://upload.wikimedia.org/wikipedia/commons/f/f8/Patriot_missile_launch_b.jpg

Accumulation of round-off errors
Patriot missiles

In 1991, American Patriot missiles failed to intercept an incoming Scud missile, killing 28 soldiers and injuring 100 other people, due to a bad computation of internal time due to an accumulation of round-off errors.

Figure: A Patriot missile launch

Image from

https://upload.wikimedia.org/wikipedia/commons/f/f8/Patriot_missile_launch_b.jpg

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, “sources of errors”:

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, “sources of errors”:

1. approximation errors;

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, “sources of errors”:

1. approximation errors;
2. round-off errors.

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, "sources of errors":

1. approximation errors;
2. round-off errors.

Example

Several errors occur when approximating $\int_0^1 f(x)\,\mathrm{d}x$: details on the blackboard!

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, "sources of errors":

1. approximation errors;
2. round-off errors.

Example

Several errors occur when approximating $\int_0^1 f(x)\,\mathrm{d}x$: details on the blackboard!

Approximation errors are typically studied by numerical analysts: rigorous error bounds, convergence results, etc.

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, “sources of errors”:

1. approximation errors;
2. round-off errors.

Example

Several errors occur when approximating $\int_0^1 f(x)\,\mathrm{d}x$: details on the blackboard!

Approximation errors are typically studied by numerical analysts: rigorous error bounds, convergence results, etc.

As for round-off errors, in “practical applications” it is important to **be aware** of them and to **keep them small by design**.

Summary of limitations

When numerical computations are performed, there are typically two, rather distinct, "sources of errors":

1. approximation errors;
2. round-off errors.

Example

Several errors occur when approximating $\int_0^1 f(x)\,\mathrm{d}x$: details on the blackboard!

Approximation errors are typically studied by numerical analysts: rigorous error bounds, convergence results, etc.

As for round-off errors, in "practical applications" it is important to **be aware** of them and to **keep them small by design**. This typically involves a suitable **stability analysis** of the numerical methods.

Where are we now?

For us, an important question remains.

How to obtain **mathematically rigorous** results based on numerical computations?

Where are we now?

For us, an important question remains.

How to obtain **mathematically rigorous** results based on numerical computations?

If only one could ignore round-off errors...

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals**

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals in such a way that the result of an operation belongs to the returned interval**.

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals in such a way that the result of an operation belongs to the returned interval**.

Appealing:

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals in such a way that the result of an operation belongs to the returned interval**.

Appealing:

- to analysts: this is a quantitative version of the ε's and the δ's;

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals in such a way that the result of an operation belongs to the returned interval**.

Appealing:

- to analysts: this is a quantitative version of the ε's and the δ's;
- to physicists: physical measurements are performed up to a finite precision anyway.

A simple solution?

The main idea of interval arithmetic is very simple, yet powerful.

The idea of interval arithmetic

We will **replace numbers by intervals in such a way that the result of an operation belongs to the returned interval**.

Appealing:

- to analysts: this is a quantitative version of the ε's and the δ's;
- to physicists: physical measurements are performed up to a finite precision anyway.

Although this may seem a paradox, all exact science is dominated by the idea of approximation.

— Bertrand Russell, The Scientific Outlook

The class $\mathcal{I}_{\mathbb{R}}$ of intervals

The intervals we will consider are the topologically closed and connected subsets of $\mathbb{R}$ (as specified in the standard IEEE-1788 devoted to interval arithmetic[1]), i.e. they belong to the class $\mathcal{I}_{\mathbb{R}}$ of subsets of $\mathbb{R}$ defined by

$$\begin{aligned}\mathcal{I}_{\mathbb{R}} := \left\{\emptyset\right\} &\cup \left\{[a, b] \mid a, b \in \mathbb{R}, a \leq b\right\}\\ &\cup \left\{[a, +\infty[\mid a \in \mathbb{R}\right\}\\ &\cup \left\{]-\infty, b] \mid b \in \mathbb{R}\right\}\\ &\cup \left\{]-\infty, +\infty[:= \mathbb{R}\right\}.\end{aligned}$$

[1] See https://standards.ieee.org/ieee/1788/4431/.

Operations on intervals

Given two intervals $\mathbf{x}$ and $\mathbf{y}$, their *sum* is given by

$$\mathbf{x} + \mathbf{y} := \left\{ x + y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\},$$

their *difference* by

$$\mathbf{x} - \mathbf{y} := \left\{ x - y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\}$$

and their *product* by

$$\mathbf{x} \cdot \mathbf{y} := \left\{ x \cdot y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\}.$$

Operations on intervals

Given two intervals $\mathbf{x}$ and $\mathbf{y}$, their *sum* is given by

$$\mathbf{x} + \mathbf{y} := \left\{ x + y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\},$$

their *difference* by

$$\mathbf{x} - \mathbf{y} := \left\{ x - y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\}$$

and their *product* by

$$\mathbf{x} \cdot \mathbf{y} := \left\{ x \cdot y \mid x \in \mathbf{x}, y \in \mathbf{y} \right\}.$$

Examples and surprises: on the blackboard!

In general: interval extensions

Definition

Let $D \subseteq \mathbb{R}$ be a set and let $F : D \to \mathbb{R}$ be a map.

An *interval extension* of F is an application $\mathbf{F} : \mathcal{I}_\mathbb{R} \to \mathcal{I}_\mathbb{R}$ which satisfies the *containment property*, namely so that for all $\mathbf{x} \in \mathcal{I}_\mathbb{R}$, the set

$$F(\mathbf{x}) := \left\{ F(x) \mid x \in \mathbf{x} \cap D \right\}$$

is included in $\mathbf{F}(\mathbf{x})$.

In general: interval extensions

Definition

Let $D \subseteq \mathbb{R}$ be a set and let $F : D \to \mathbb{R}$ be a map.

An *interval extension* of F is an application $\mathbf{F} : \mathcal{I}_{\mathbb{R}} \to \mathcal{I}_{\mathbb{R}}$ which satisfies the *containment property*, namely so that for all $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$, the set

$$F(\mathbf{x}) := \left\{ F(x) \mid x \in \mathbf{x} \cap D \right\}$$

is included in $\mathbf{F}(\mathbf{x})$.

Examples on the blackboard!

In general: interval extensions

Definition

Let $D \subseteq \mathbb{R}$ be a set and let $F : D \to \mathbb{R}$ be a map.

An *interval extension* of F is an application $\mathbf{F} : \mathcal{I}_{\mathbb{R}} \to \mathcal{I}_{\mathbb{R}}$ which satisfies the *containment property*, namely so that for all $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$, the set

$$F(\mathbf{x}) := \left\{ F(x) \mid x \in \mathbf{x} \cap D \right\}$$

is included in $\mathbf{F}(\mathbf{x})$.

Examples on the blackboard! *Compare extensions of $F : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$ with the product operation.*

Fundamental theorem of interval arithmetic

Theorem

If interval extensions of real functions $f_1, \dots, f_k$ are composed, the result is an interval extension of the composition $f_1 \circ \cdots \circ f_k$.

Fundamental theorem of interval arithmetic

Theorem

If interval extensions of real functions $f_1, \ldots, f_k$ are composed, the result is an interval extension of the composition $f_1 \circ \cdots \circ f_k$.

Allows to obtain interval extensions of complicated functions by composing interval extensions of its subparts.

In practice

The set $\mathcal{I}_{\mathbb{R}}$ is a mathematical notion.

In practice

The set $\mathcal{I}_{\mathbb{R}}$ is a mathematical notion.
In practice, the implementation will use intervals from the set

$$\mathcal{I}_{\mathbb{F}} := \left\{ \mathbf{x} = [\underline{x}, \overline{x}] \mid \underline{x} \leq \overline{x} \text{ are two floating-point numbers} \right\} \cup \left\{ \emptyset \right\}.$$

Back to the computation of sin(π)

Let us use the "`mpmath`" library[2] in Python3 and ask the value of

`iv.pi`

then

`iv.sin(iv.pi)`.

[2] See in particular the module `iv`, devoted to interval arithmetic at https://www.mpmath.org/doc/1.0.0/contexts.html.

What interval arithmetic can and cannot do

- It may allow to prove that some values are nonzero, but it cannot prove that some values are equal to zero.

What interval arithmetic can and cannot do

- It may allow to prove that some values are nonzero, but it cannot prove that some values are equal to zero.

Example

Let us evaluate `iv.sin(1.)` as well as `iv.sin(iv.pi)` and comment on the result.

What interval arithmetic can and cannot do

- It may allow to prove that some values are nonzero, but it cannot prove that some values are equal to zero.

Example

Let us evaluate `iv.sin(1.)` as well as `iv.sin(iv.pi)` and comment on the result.

- If a returned interval is “too big”, it is valid but useless.

What interval arithmetic can and cannot do

- It may allow to prove that some values are nonzero, but it cannot prove that some values are equal to zero.

Example

Let us evaluate `iv.sin(1.)` as well as `iv.sin(iv.pi)` and comment on the result.

- If a returned interval is "too big", it is valid but useless.
  For instance, `iv.sin(x)` could return `[-1, 1]` regardless of the value of `x`, but this bound is useless.

What interval arithmetic can and cannot do

- It may allow to prove that some values are nonzero, but it cannot prove that some values are equal to zero.

Example

Let us evaluate `iv.sin(1.)` as well as `iv.sin(iv.pi)` and comment on the result.

- If a returned interval is "too big", it is valid but useless.
  For instance, `iv.sin(x)` could return `[-1, 1]` regardless of the value of `x`, but this bound is useless.
- Nevertheless, it is in principle possible to show that given matrices are invertible, positive/negative definite... using interval arithmetic.

Locating roots of a function

Let $F : [0,1] \to \mathbb{R}$. If $\mathbf{F}$ is an interval extension of F and if $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$ is included in $[0,1]$, then

Locating roots of a function

Let $F : [0,1] \to \mathbb{R}$. If $\mathbf{F}$ is an interval extension of F and if $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$ is included in $[0,1]$, then the implication

$$\left(0 \notin \mathbf{F}(\mathbf{x})\right) \implies \left(\mathbf{x} \text{ does not contain any roots of } F\right)$$

holds.

Locating roots of a function

Let $F : [0,1] \to \mathbb{R}$. If $\mathbf{F}$ is an interval extension of F and if $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$ is included in $[0,1]$, then the implication

$$\Big(0 \notin \mathbf{F}(\mathbf{x})\Big) \implies \Big(\mathbf{x} \text{ does not contain any roots of } F\Big)$$

holds.

We may thus divide $[0,1]$ into many "small" intervals and discard all those for which we are sure that F has no roots, this being determined by evaluating the interval extension $\mathbf{F}$.

Locating roots of a function

Let $F : [0,1] \to \mathbb{R}$. If $\mathbf{F}$ is an interval extension of F and if $\mathbf{x} \in \mathcal{I}_{\mathbb{R}}$ is included in $[0,1]$, then the implication

$$\Big(0 \notin \mathbf{F}(\mathbf{x})\Big) \implies \Big(\mathbf{x} \text{ does not contain any roots of } F\Big)$$

holds.

We may thus divide $[0,1]$ into many "small" intervals and discard all those for which we are sure that F has no roots, this being determined by evaluating the interval extension $\mathbf{F}$. We end up with (possibly many) small intervals such that all potential roots of F belong to one of those.

Application of interval arithmetic to nonlinear analysis

Existence of the Lorenz strange attractor

The system of ODEs

$$\begin{aligned} \partial_t x_1 &= -\sigma x_1 + \sigma x_2 \\ \partial_t x_2 &= \rho x_1 - x_2 - x_1 x_3, \\ \partial_t x_3 &= -\beta x_3 + x_1 x_2 \end{aligned}$$

was introduced by Edward Lorenz in 1963 as a simple model of atmospheric dynamics.

Application of interval arithmetic to nonlinear analysis

Existence of the Lorenz strange attractor

The system of ODEs

$$\begin{aligned} \partial_t x_1 &= -\sigma x_1 + \sigma x_2 \\ \partial_t x_2 &= \rho x_1 - x_2 - x_1 x_3, \\ \partial_t x_3 &= -\beta x_3 + x_1 x_2 \end{aligned}$$

was introduced by Edward Lorenz in 1963 as a simple model of atmospheric dynamics.

Remarkably, this system is **chaotic** (i.e., it is very sensitive to the initial conditions in long time)

Application of interval arithmetic to nonlinear analysis

Existence of the Lorenz strange attractor

The system of ODEs

$$\begin{aligned} \partial_t x_1 &= -\sigma x_1 + \sigma x_2 \\ \partial_t x_2 &= \rho x_1 - x_2 - x_1 x_3, \\ \partial_t x_3 &= -\beta x_3 + x_1 x_2 \end{aligned}$$

was introduced by Edward Lorenz in 1963 as a simple model of atmospheric dynamics.

Remarkably, this system is **chaotic** (i.e., it is very sensitive to the initial conditions in long time) and possesses a strange attractor.

Application of interval arithmetic to nonlinear analysis

Existence of the Lorenz strange attractor

The system of ODEs

$$\begin{aligned} \partial_t x_1 &= -\sigma x_1 + \sigma x_2 \\ \partial_t x_2 &= \rho x_1 - x_2 - x_1 x_3, \\ \partial_t x_3 &= -\beta x_3 + x_1 x_2 \end{aligned}$$

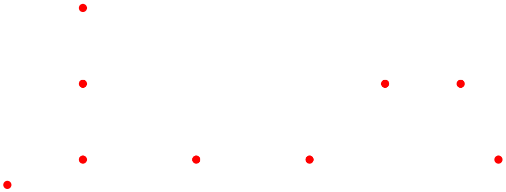
was introduced by Edward Lorenz in 1963 as a simple model of atmospheric dynamics.

Remarkably, this system is **chaotic** (i.e., it is very sensitive to the initial conditions in long time) and possesses a strange attractor.

This fact, though conjectured since the 1960s, was only proved by Warwick Tucker in 1999, using a computer-assisted proof using interval arithmetic.

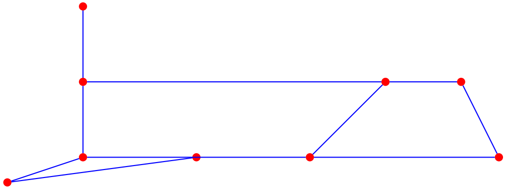
What is a compact metric graph?

A compact metric graph is made of a finite number of vertices



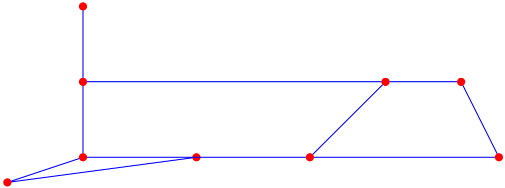
What is a compact metric graph?

A compact metric graph is made of a finite number of vertices and of edges joining the vertices.



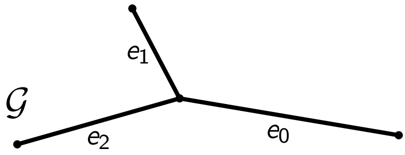
What is a compact metric graph?

A compact metric graph is made of a finite number of vertices and of edges joining the vertices.



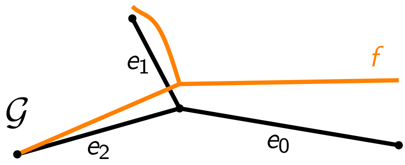
Metric graphs: the lengths of edges are important.

Functions defined on metric graphs



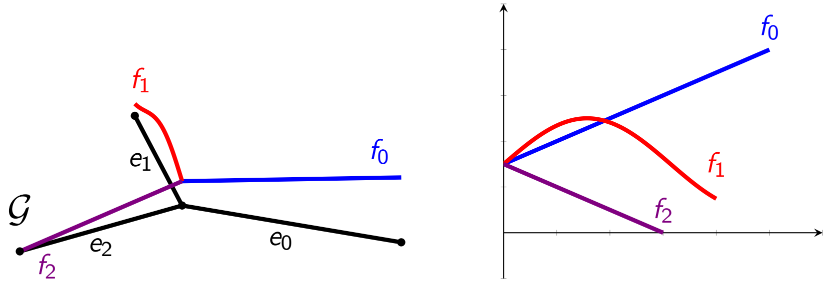
A compact metric graph $\mathcal{G}$ with three edges e_0 (length 5), e_1 (length 4) and e_2 (length 3)

Functions defined on metric graphs



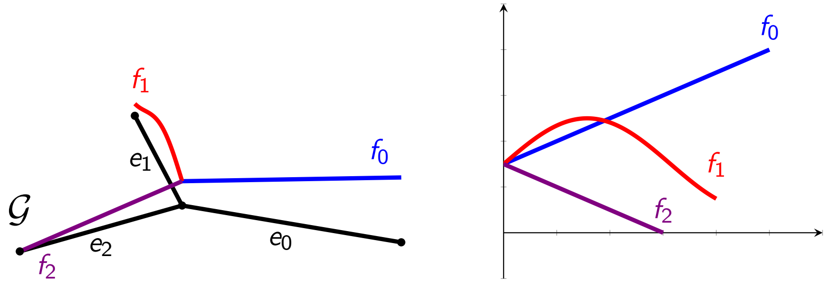
A compact metric graph $\mathcal{G}$ with three edges e_0 (length 5), e_1 (length 4) and e_2 (length 3), a function $f : \mathcal{G} \to \mathbb{R}$

Functions defined on metric graphs



A compact metric graph $\mathcal{G}$ with three edges e_0 (length 5), e_1 (length 4) and e_2 (length 3), a function $f : \mathcal{G} \to \mathbb{R}$, and the three associated real functions.

Functions defined on metric graphs



A compact metric graph $\mathcal{G}$ with three edges e_0 (length 5), e_1 (length 4) and e_2 (length 3), a function $f : \mathcal{G} \to \mathbb{R}$, and the three associated real functions.

$$\int_{\mathcal{G}} f \,\mathrm{d}x := \int_0^5 f_0(x)\,\mathrm{d}x + \int_0^4 f_1(x)\,\mathrm{d}x + \int_0^3 f_2(x)\,\mathrm{d}x$$

The spectral problem on metric graphs

We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

The spectral problem on metric graphs

We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

$$\begin{cases} -u'' = \gamma u & \text{on each edge } e \text{ of } \mathcal{G}, \\ \\ \\ \end{cases}$$

The spectral problem on metric graphs

We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

$$\begin{cases} -u'' = \gamma u & \text{on each edge } e \text{ of } \mathcal{G}, \\ u \text{ is continuous} & \text{for every vertex } \mathrm{V} \text{ of } \mathcal{G}, \\ & \end{cases}$$

The spectral problem on metric graphs

We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

$$\begin{cases} -u'' = \gamma u & \text{on each edge } e \text{ of } \mathcal{G}, \\ u \text{ is continuous} & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \\ \displaystyle\sum_{e \succ \mathrm{v}} \frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v}) = 0 & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \end{cases}$$

The spectral problem on metric graphs

We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

$$\begin{cases} -u'' = \gamma u & \text{on each edge } e \text{ of } \mathcal{G}, \\ u \text{ is continuous} & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \\ \displaystyle\sum_{e \succ \mathrm{v}} \frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v}) = 0 & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \end{cases}$$

where the symbol $e \succ \mathrm{v}$ means that the sum ranges over all edges of vertex v and where $\frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v})$ is the outgoing derivative of u at v (*Kirchhoff's condition*).

The spectral problem on metric graphs

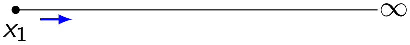
We are interested in solutions (γ, u), with $u \neq 0$, of the differential system

$$\begin{cases} -u'' = \gamma u & \text{on each edge } e \text{ of } \mathcal{G}, \\ u \text{ is continuous} & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \\ \displaystyle\sum_{e \succ \mathrm{v}} \frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v}) = 0 & \text{for every vertex } \mathrm{v} \text{ of } \mathcal{G}, \end{cases}$$

where the symbol $e \succ \mathrm{v}$ means that the sum ranges over all edges of vertex v and where $\frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v})$ is the outgoing derivative of u at v (*Kirchhoff's condition*).

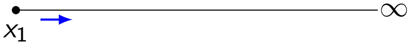
Remark: we always have $\dim E_1 = 1$ with $\gamma_1 = 0$, considering constant functions.

Kirchoff's condition: degree one nodes



$$\lim_{\substack{t \to 0 \\ t>0}} \frac{u(x_1 + t) - u(x_1)}{t} = 0$$

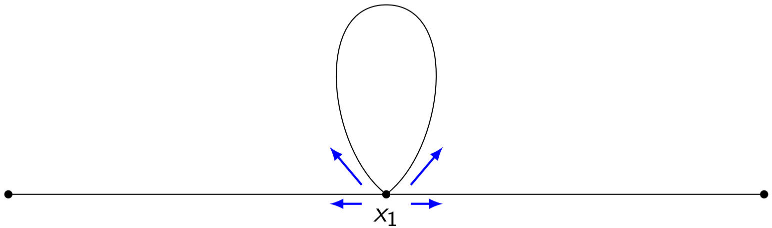
Kirchoff's condition: degree one nodes



$$\lim_{\substack{t \to 0 \\ t>0}} \frac{u(x_1+t)-u(x_1)}{t} = 0$$

In other words, the derivative of u at x_1 vanishes: this is the usual Neumann condition.

Kirchoff's condition in general: outgoing derivatives



$$\sum_{e \succ \mathrm{v}} \frac{\mathrm{d}u}{\mathrm{d}x_e}(\mathrm{v}) = 0$$

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

$$\begin{cases} -u'' + \lambda u = \gamma_2 |u|^{p-2} u \quad \text{on every edge of } \mathcal{G}, \\ \\ \end{cases} \qquad (\mathcal{P}_p)$$

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

$$\begin{cases} -u'' + \lambda u = \gamma_2 |u|^{p-2} u & \text{on every edge of } \mathcal{G}, \\ u \text{ is continuous} & \text{on } \mathcal{G}, \end{cases} \qquad (\mathcal{P}_p)$$

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

$$\begin{cases} -u'' + \lambda u = \gamma_2 |u|^{p-2} u & \text{on every edge of } \mathcal{G}, \\ u \text{ is continuous} & \text{on } \mathcal{G}, \\ \displaystyle\sum_{e \succ v} \frac{\mathrm{d}u}{\mathrm{d}x_e}(v) = 0 & \text{for every vertex } v. \end{cases} \tag{$\mathcal{P}_p$}$$

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

$$\begin{cases} -u'' + \lambda u = \gamma_2 |u|^{p-2} u & \text{on every edge of } \mathcal{G}, \\ u \text{ is continuous} & \text{on } \mathcal{G}, \\ \displaystyle\sum_{e \succ v} \frac{\mathrm{d}u}{\mathrm{d}x_e}(v) = 0 & \text{for every vertex } v. \end{cases} \qquad (\mathcal{P}_p)$$

When $p = 2$, the solutions of ($\mathcal{P}_p$) are the eigenfunctions in E_2.

The nonlinear Schrödinger equation on metric graphs

We generalize the spectral problem by introducing a *nonlinear term* (which appears in models of optic fibers, Bose-Einstein condensates...).

Given $p \geq 2$, we are interested in solutions of

$$\begin{cases} -u'' + \lambda u = \gamma_2 |u|^{p-2} u & \text{on every edge of } \mathcal{G}, \\ u \text{ is continuous} & \text{on } \mathcal{G}, \\ \displaystyle\sum_{e \succ v} \frac{\mathrm{d}u}{\mathrm{d}x_e}(v) = 0 & \text{for every vertex } v. \end{cases} \qquad (\mathcal{P}_p)$$

When $p = 2$, the solutions of $(\mathcal{P}_p)$ are the eigenfunctions in E_2.

Question

What about $p > 2$?

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n \geq 1} \subseteq \,]2, +\infty[$ *be a sequence of exponents which converges to* 2

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n\geq 1} \subseteq \,]2, +\infty[$ *be a sequence of exponents which converges to* 2 *and* $(u_{p_n})_{n\geq 1} \subseteq H^1(\mathcal{G})$ *be a sequence of nonzero solutions to the problems* $(\mathcal{P}_{p_n})$.

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n\geq 1} \subseteq]2, +\infty[$ *be a sequence of exponents which converges to* 2 *and* $(u_{p_n})_{n\geq 1} \subseteq H^1(\mathcal{G})$ *be a sequence of nonzero solutions to the problems* $(\mathcal{P}_{p_n})$. *Assume that* $(u_{p_n})_n$ *converges weakly in* $H^1(\mathcal{G})$ *to a function* $u_* \in H^1(\mathcal{G})$.

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n\geq 1} \subseteq]2, +\infty[$ *be a sequence of exponents which converges to* 2 *and* $(u_{p_n})_{n\geq 1} \subseteq H^1(\mathcal{G})$ *be a sequence of nonzero solutions to the problems* $(\mathcal{P}_{p_n})$. *Assume that* $(u_{p_n})_n$ *converges weakly in* $H^1(\mathcal{G})$ *to a function* $u_* \in H^1(\mathcal{G})$. *Then,* u_* *belongs to* E_2

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n\geq 1} \subseteq \,]2, +\infty[$ *be a sequence of exponents which converges to* 2 *and* $(u_{p_n})_{n\geq 1} \subseteq H^1(\mathcal{G})$ *be a sequence of nonzero solutions to the problems* $(\mathcal{P}_{p_n})$*. Assume that* $(u_{p_n})_n$ *converges weakly in* $H^1(\mathcal{G})$ *to a function* $u_* \in H^1(\mathcal{G})$*. Then,* u_* *belongs to* E_2 *and one has*

$$\int_{\mathcal{G}} u_* \ln |u_*| \varphi \,\mathrm{d}x = 0 \qquad \forall \varphi \in E_2.$$

The quasilinear regime $p \approx 2$ ($p > 2$)

Proposition

Let $(p_n)_{n\geq 1} \subseteq]2, +\infty[$ *be a sequence of exponents which converges to* 2 *and* $(u_{p_n})_{n\geq 1} \subseteq H^1(\mathcal{G})$ *be a sequence of nonzero solutions to the problems* $(\mathcal{P}_{p_n})$. *Assume that* $(u_{p_n})_n$ *converges weakly in* $H^1(\mathcal{G})$ *to a function* $u_* \in H^1(\mathcal{G})$. *Then,* u_* *belongs to* E_2 *and one has*

$$\int_{\mathcal{G}} u_* \ln |u_*| \varphi \, \mathrm{d}x = 0 \qquad \forall \varphi \in E_2.$$

We say that $u_* \in E_2$ is a *solution of the reduced problem* if the above condition holds.

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4} \int_{\mathcal{G}} \varphi^2(x) \big(1 - 2 \ln |\varphi(x)|\big) \, \mathrm{d}x$$

is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points.

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4} \int_{\mathcal{G}} \varphi^2(x) \big(1 - 2 \ln |\varphi(x)|\big) \, \mathrm{d}x$$

is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points. We thus have two goals:

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4}\int_{\mathcal{G}} \varphi^2(x)\big(1 - 2\ln|\varphi(x)|\big)\,\mathrm{d}x$$

is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points. We thus have two goals:

1. find all nonzero critical points $\varphi_* \in E_2$ of $\mathcal{J}_*$;

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4} \int_{\mathcal{G}} \varphi^2(x) \big(1 - 2 \ln |\varphi(x)|\big) \, \mathrm{d}x$$

is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points. We thus have two goals:

1. find all nonzero critical points $\varphi_* \in E_2$ of $\mathcal{J}_*$;
2. determine the *nondegenerate* critical points $\varphi_* \in E_2$, namely those for which the Hessian $\mathcal{J}_*''(\varphi_*)$ is invertible

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4} \int_{\mathcal{G}} \varphi^2(x)\big(1 - 2\ln|\varphi(x)|\big)\,\mathrm{d}x$$

is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points. We thus have two goals:

1. find all nonzero critical points $\varphi_* \in E_2$ of $\mathcal{J}_*$;
2. determine the *nondegenerate* critical points $\varphi_* \in E_2$, namely those for which the Hessian $\mathcal{J}_*''(\varphi_*)$ is invertible (when it is defined, which is not always the case, but this is another story);

Variational formulation

The functional $\mathcal{J}_* : E_2 \to \mathbb{R}$

$$\mathcal{J}_*(\varphi) := \frac{1}{4} \int_{\mathcal{G}} \varphi^2(x) \big(1 - 2 \ln |\varphi(x)|\big) \, \mathrm{d}x$$

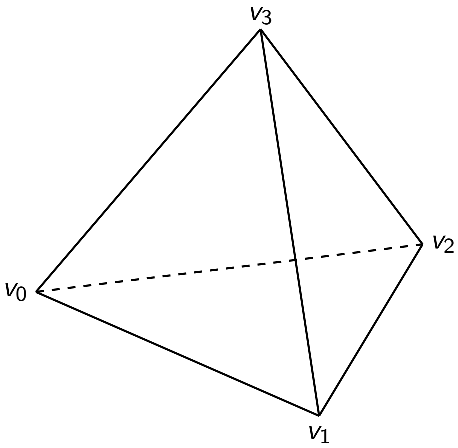
is of class $\mathcal{C}^1$, and the solutions of the reduced problem coincide with its critical points. We thus have two goals:

1. find all nonzero critical points $\varphi_* \in E_2$ of $\mathcal{J}_*$;
2. determine the *nondegenerate* critical points $\varphi_* \in E_2$, namely those for which the Hessian $\mathcal{J}_*''(\varphi_*)$ is invertible (when it is defined, which is not always the case, but this is another story);

Using a “Lyapunov-Schmidt” argument, we can show **existence and uniqueness results around a nondegenerate critical point** for $(\mathcal{P}_p)$, when $p \approx 2$.

The tetrahedron

In the remainder of the talk, we will only consider the following graph $\mathcal{G}_{\mathrm{t}}$.



Second eigenspace and symmetries of $\mathcal{G}_t$

One may explicitly determine the second eigenspace for $\mathcal{G}_t$. It turns out that $\dim E_2 = 3$.

Second eigenspace and symmetries of $\mathcal{G}_t$

One may explicitly determine the second eigenspace for $\mathcal{G}_t$. It turns out that $\dim E_2 = 3$.

Moreover, the group

$$G_t := S_4 \times \{\pm 1\}$$

acts on E_2 due to the fact that "all vertices of the tetrahedron are the same" and that one may replace φ by $-\varphi$.

Second eigenspace and symmetries of $\mathcal{G}_t$

One may explicitly determine the second eigenspace for $\mathcal{G}_t$. It turns out that $\dim E_2 = 3$.

Moreover, the group

$$G_t := S_4 \times \{\pm 1\}$$

acts on E_2 due to the fact that "all vertices of the tetrahedron are the same" and that one may replace φ by $-\varphi$.

In this way, we obtain an *isometric group action*

$$G_t \times E_2 \to E_2 : (g, \varphi) \mapsto g \cdot \varphi,$$

such that $J_*(g \cdot \varphi) = J_*(\varphi)$ for all $(g, \varphi) \in G_t \times E_2$.

Critical points created by the symmetries

The presence of such a rich symmetry group entails the existence of four distinct families of critical points, due to the *principle of symmetric criticality*.

Critical points created by the symmetries

The presence of such a rich symmetry group entails the existence of four distinct families of critical points, due to the *principle of symmetric criticality*.

Theorem (Principle of symmetric criticality, Palais, 1979)

Assume that the action of the topological group G on the Hilbert space E is isometric.

Critical points created by the symmetries

The presence of such a rich symmetry group entails the existence of four distinct families of critical points, due to the *principle of symmetric criticality*.

Theorem (Principle of symmetric criticality, Palais, 1979)

Assume that the action of the topological group G on the Hilbert space E is isometric. If $J \in \mathcal{C}^1(E, \mathbb{R})$ is invariant under this action

Critical points created by the symmetries

The presence of such a rich symmetry group entails the existence of four distinct families of critical points, due to the *principle of symmetric criticality*.

Theorem (Principle of symmetric criticality, Palais, 1979)

Assume that the action of the topological group G on the Hilbert space E is isometric. If $J \in \mathcal{C}^1(E, \mathbb{R})$ is invariant under this action and if u is a critical point of J restricted to

$$\mathrm{Fix}(G) := \left\{ u \in E \mid \forall g \in G, g \cdot u = u \right\},$$

Critical points created by the symmetries

The presence of such a rich symmetry group entails the existence of four distinct families of critical points, due to the *principle of symmetric criticality*.

Theorem (Principle of symmetric criticality, Palais, 1979)

Assume that the action of the topological group G on the Hilbert space E is isometric. If $J \in \mathcal{C}^1(E, \mathbb{R})$ is invariant under this action and if u is a critical point of J restricted to

$$\mathrm{Fix}(G) := \Big\{ u \in E \mid \forall g \in G, g \cdot u = u \Big\},$$

then u is a critical point of J.

A natural question

Critical point theory (using the principle of symmetric criticality, Morse theory, etc), will give relations on the number of critical points and the existence of some specific symmetric critical points.

A natural question

Critical point theory (using the principle of symmetric criticality, Morse theory, etc), will give relations on the number of critical points and the existence of some specific symmetric critical points.

However, it cannot classify all critical points of J_*.

A natural question

Critical point theory (using the principle of symmetric criticality, Morse theory, etc), will give relations on the number of critical points and the existence of some specific symmetric critical points.

However, it cannot classify all critical points of J_*.

Question

Does $\mathcal{J}_*$ possess critical points other than the ones of the four aforementioned families?

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

Strategy of the proof:

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

Strategy of the proof:

1. locating small "boxes" containing all critical points of $\mathcal{J}_*$, by root finding methods.

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

Strategy of the proof:

1. locating small "boxes" containing all critical points of $\mathcal{J}_*$, by root finding methods.
2. proving uniqueness of critical points inside each box using second order information.

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

Strategy of the proof:

1. locating small "boxes" containing all critical points of $\mathcal{J}_*$, by root finding methods.
2. proving uniqueness of critical points inside each box using second order information.

After a careful implementation and some computing time...

A computer-assisted answer

Theorem (De Coster, G., Troestler (2024))

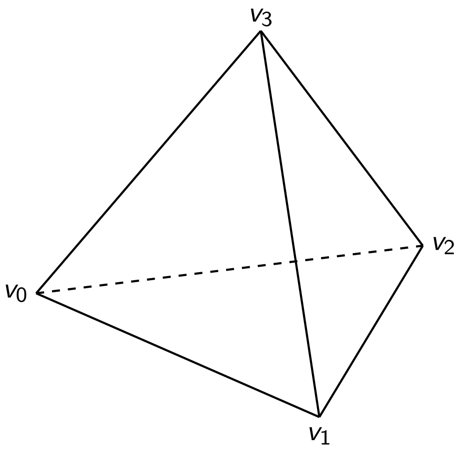
All critical points of $\mathcal{J}_* : E_2 \to \mathbb{R}$ (for the tetrahedron graph) belong to one of the four families obtained thanks to the symmetries.

Strategy of the proof:

1. locating small "boxes" containing all critical points of $\mathcal{J}_*$, by root finding methods.
2. proving uniqueness of critical points inside each box using second order information.

After a careful implementation and some computing time...
Things worked out!

Thanks for your attention!



References

Floating point arithmetic

- Computerphile,
  Floating Point Numbers,
  https://www.youtube.com/watch?v=PZRI1IfStY0

- Jean-Michel Muller, Nicolas Brunie, Florent de Dinechin, Claude-Pierre Jeannerod, Mioara Joldes, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Serge Torres,
  Handbook of Floating-Point Arithmetic, Second Edition, Birkhäuser (2018).

- https://en.wikipedia.org/wiki/Double-precision_floating-point_format

References

Numerical errors and computer bugs in general

Thomas Huckle, Tobias Neckel, *Bits and Bugs, A Scientific and Historical Review of Software Failures in Computational Science*, Society for Industrial and Applied Mathematics (2019).

https://sites.math.rutgers.edu/~sg1108/Math373/Matherrors.html

References

Historical examples

More about Ariane 501 (in French):

AstronoGeek,
L'explosion d'Ariane 501 était due à... une ligne de code.
https://www.youtube.com/watch?v=n4QheSThrC8

More about the Vancouver stock exchange:

Christ Stewart,
The code that sunk a stock market.
https://www.youtube.com/watch?v=RdDIANVl7dc

The Patriot Missile Failure:

Douglas N. Arnold,
The Patriot Missile Failure. https://www-users.cse.umn.edu/~arnold/disasters/patriot.html

References

Interval-arithmetic and computer-assisted proofs

Javier Gómez-Serrano
Computer-assisted proofs in PDE: a survey, SeMA Journal, 76, no. 3, 459-484 (2019).

Warwick Tucker
Validated Numerics, A Short Introduction to Rigorous Computations, Princeton University Press (2011).

Warwick Tucker,
The Lorenz attractor exists, Comptes Rendus de l'Académie des Sciences - Series I - Mathematics, Volume 328, Issue 12, 1197–1202 (1999).

References

The principle of symmetric criticality and variational methods

Palais, Richard S.,
The principle of symmetric criticality, Comm. Math. Phys. 69, no. 1, 19–30 (1979).

Willem, Michel
Minimax theorems, Progr. Nonlinear Differential Equations Appl., 24 Birkhäuser Boston, Inc., Boston, MA (1996).

References

NLS on metric graphs

De Coster C., Dovetta S., Galant D., Serra E., Troestler C., *Constant sign and sign changing NLS ground states on noncompact metric graphs.* ArXiV preprint: https://arxiv.org/abs/2306.12121.

Galant D., *The nonlinear Schrödinger equation on metric graphs.* PhD thesis (UMONS and UPHF), available on my webpage damien-gal.github.io.