An Introduction to Fixed-Point A Posteriori Validation

Florent Bréhard florent.brehard@univ-lille.fr

Thursday, October 24, 2024

Compute a Rigorous Polynomial Approximation of $\dfrac{1}{P(x)}$

- Rational function $\rightsquigarrow$ not a polynomial (of finite degree)

Some Limitations of the Self-Validating Approach

Compute a Rigorous Polynomial Approximation of $\dfrac{1}{P(x)}$

- Rational function $\quad\rightsquigarrow\quad$ not a polynomial (of finite degree)

- **Taylor model:** $P = (P_0 + P_1 x + \cdots + P_n x^n, \varepsilon) \quad\rightsquigarrow$
  $Q = (Q_0 + Q_1 x + \cdots + Q_n x^n, \eta)$
  $P(x)Q(x) = 1 \quad\Rightarrow\quad P_n$ depends explicitly on $P_0, \ldots, P_n$ and $Q_0, \ldots, Q_{n-1}$
  How to determine η rigorously?

Some Limitations of the Self-Validating Approach

Compute a Rigorous Polynomial Approximation of $\dfrac{1}{P(x)}$

- Rational function $\rightsquigarrow$ not a polynomial (of finite degree)

- **Taylor model:** $P = (P_0 + P_1 x + \cdots + P_n x^n, \varepsilon) \quad \rightsquigarrow$
  $Q = (Q_0 + Q_1 x + \cdots + Q_n x^n, \eta)$
  $P(x)Q(x) = 1 \quad \Rightarrow \quad P_n$ depends explicitly on $P_0, \ldots, P_n$ and $Q_0, \ldots, Q_{n-1}$
  How to determine η rigorously?

- **Chebyshev model:** $P = (P_0 + P_1 x + \cdots + P_n T_n(x), \varepsilon) \quad \rightsquigarrow$
  $Q = (Q_0 + Q_1 x + \cdots + Q_n T_n(x), \eta)$
  $P(x)Q(x) = 1 \quad \Rightarrow \quad$ not a finite formula, since
  $T_n(x) T_m(x) = \frac{1}{2}(T_{n+m}(x) + T_{|n-m|}(x))$

Some Limitations of the Self-Validating Approach

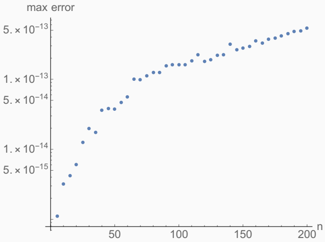
Matrix inverse via Gaussian elimination using interval arithmetics

Some Limitations of the Self-Validating Approach

Matrix inverse via Gaussian elimination using interval arithmetics

- **Example:** the Lehmer matrix $L_n = \left(\dfrac{\min(i,j)}{\max(i,j)}\right)_{1 \le i,j \le n}$ is well-conditionned

- Gaussian elimination (using binary64 FP arithmetic) computes L_n^{-1} accurately

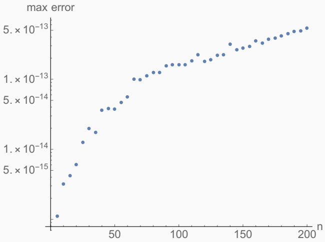


maximum error of $L_n^{-1} L_n - I_n$ computed using binary64 Gaussian elimination

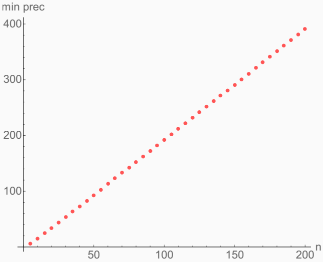
Some Limitations of the Self-Validating Approach

Matrix inverse via Gaussian elimination using interval arithmetics

- **Example:** the Lehmer matrix $L_n = \left(\dfrac{\min(i,j)}{\max(i,j)}\right)_{1 \leq i,j \leq n}$ is well-conditionned
- Gaussian elimination (using binary64 FP arithmetic) computes L_n^{-1} accurately
- Intervals of pivots using interval arithmetic grow much faster
  $\Rightarrow$ interval Gaussian elimination fails



maximum error of $L_n^{-1} L_n - I_n$ computed using binary64 Gaussian elimination



minimum FP precision needed s.t. interval Gaussian elimination on L_n does not fail

A Posteriori Validation Approach

- $\mathcal{F} : X \to Y$, X and Y Banach spaces, Solve $\mathcal{F}(x) = 0 \quad \rightsquigarrow \quad x^* \in X$
  $X = Y = (\mathcal{C}(I), \|\cdot\|_\infty)$, $\mathcal{F}(f) = gf - 1 = 0$, $f^* = \dfrac{1}{g}$

A Posteriori Validation Approach

- $\mathcal{F} : X \to Y$, X and Y Banach spaces, Solve $\mathcal{F}(x) = 0 \quad \rightsquigarrow \quad x^* \in X$
  $X = Y = (\mathcal{C}(I), \|\cdot\|_\infty)$, $\mathcal{F}(f) = gf - 1 = 0$, $f^* = \dfrac{1}{g}$

- $\widetilde{x} \in X$ a numerical solution: $\mathcal{F}(\widetilde{x}) \approx 0 \quad \rightsquigarrow \quad \widetilde{x} \approx x^*$
  Compute $\widetilde{f} = p = \sum_{i=0}^{n} a_i T_i \in \mathbb{R}_n[x] \subseteq X$ by interpolating $\dfrac{1}{g}$ at the Chebyshev nodes

A Posteriori Validation Approach

- $\mathcal{F} : X \to Y$, X and Y Banach spaces, Solve $\mathcal{F}(x) = 0 \quad \rightsquigarrow \quad x^* \in X$
  $X = Y = (\mathcal{C}(I), \| \cdot \|_\infty)$, $\mathcal{F}(f) = gf - 1 = 0$, $f^* = \dfrac{1}{g}$

- $\widetilde{x} \in X$ a numerical solution: $\mathcal{F}(\widetilde{x}) \approx 0 \quad \rightsquigarrow \quad \widetilde{x} \approx x^*$
  Compute $\widetilde{f} = p = \sum_{i=0}^{n} a_i T_i \in \mathbb{R}_n[x] \subseteq X$ by interpolating $\dfrac{1}{g}$ at the Chebyshev nodes

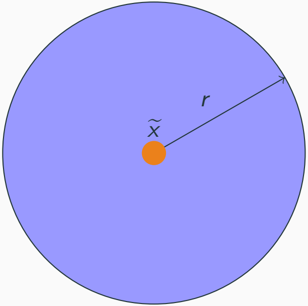
- A posteriori validation $\rightsquigarrow$ recover a rigorous bound $\varepsilon \geq \|\widetilde{x} - x^*\|$
  The pair (p, ε) is a Rigorous Polynomial Approximation for $f^* = \dfrac{1}{g}$

A Posteriori Validation Approach

- $\mathcal{F} : X \to Y$, X and Y Banach spaces, Solve $\mathcal{F}(x) = 0 \quad \rightsquigarrow \quad x^* \in X$
  $X = Y = (\mathcal{C}(I), \|\cdot\|_\infty)$, $\mathcal{F}(f) = gf - 1 = 0$, $f^* = \dfrac{1}{g}$
- $\widetilde{x} \in X$ a numerical solution: $\mathcal{F}(\widetilde{x}) \approx 0 \quad \rightsquigarrow \quad \widetilde{x} \approx x^*$
  Compute $\widetilde{f} = p = \sum_{i=0}^{n} a_i T_i \in \mathbb{R}_n[x] \subseteq X$ by interpolating $\dfrac{1}{g}$ at the Chebyshev nodes
- A posteriori validation $\rightsquigarrow$ recover a rigorous bound $\varepsilon \geq \|\widetilde{x} - x^*\|$
  The pair (p, ε) is a Rigorous Polynomial Approximation for $f^* = \dfrac{1}{g}$

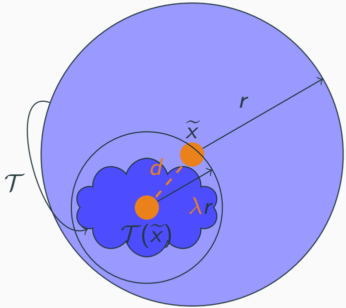
$\Rightarrow$ Use a fixed-point theorem!

- Convert $\mathcal{F}(x) = 0$ into an equivalent fixed-point equation $\mathcal{T}(x) = x$
  $\Rightarrow \quad \mathcal{T} : X \to X$ must be contracting



Banach Fixed-Point Theorem

- Convert $\mathcal{F}(x) = 0$ into an equivalent fixed-point equation $\mathcal{T}(x) = x$
  $\Rightarrow \quad \mathcal{T} : X \to X$ must be contracting



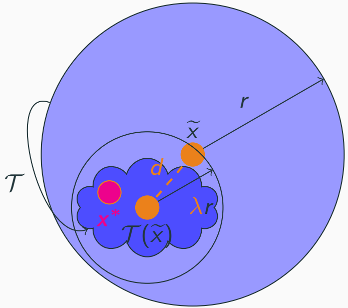
Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{x}, r)) \subseteq B(\widetilde{x}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

* it means that $\|\mathcal{T}(x) - \mathcal{T}(x')\| \leq \lambda \|x - x'\|$ for all $x, x' \in B(\widetilde{x}, r)$

- Convert $\mathcal{F}(x) = 0$ into an equivalent fixed-point equation $\mathcal{T}(x) = x$
  $\Rightarrow \quad \mathcal{T} : X \to X$ must be contracting



Banach Fixed-Point Theorem

If one can rigorously check:

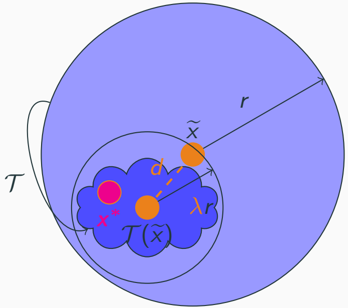
- $\mathcal{T}(B(\widetilde{x}, r)) \subseteq B(\widetilde{x}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

* it means that $\|\mathcal{T}(x) - \mathcal{T}(x')\| \leq \lambda \|x - x'\|$ for all $x, x' \in B(\widetilde{x}, r)$

Banach Fixed-Point Theorem

- Convert $\mathcal{F}(x) = 0$ into an equivalent fixed-point equation $\mathcal{T}(x) = x$
  $\Rightarrow \quad \mathcal{T} : X \to X$ must be contracting



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{x}, r)) \subseteq B(\widetilde{x}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

- Find "optimal" radius r that satisfies the theorem

* it means that $\|\mathcal{T}(x) - \mathcal{T}(x')\| \leq \lambda \|x - x'\|$ for all $x, x' \in B(\widetilde{x}, r)$

A Linear Example: Division of RPAs (1/3)

Compute $f = \dfrac{g}{h}, \quad h(x) \neq 0$ over I:

- Numerical interpolation $\quad \rightsquigarrow \quad \widetilde{f} \approx \dfrac{g}{h}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \|\cdot\|_\infty)$

A Linear Example: Division of RPAs (1/3)

Compute $f = \dfrac{g}{h}, \quad h(x) \neq 0$ over I:

- Numerical interpolation $\quad \rightsquigarrow \quad \widetilde{f} \approx \dfrac{g}{h}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \|\cdot\|_\infty)$

- We want to solve:

$$\mathcal{F}(f) = g \qquad \text{where} \qquad \mathcal{F} : X \to X, f \mapsto hf$$

$\Rightarrow$ $\mathcal{F}$ is linear

A Linear Example: Division of RPAs (1/3)

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Numerical interpolation $\quad\rightsquigarrow\quad \widetilde{f} \approx \dfrac{g}{h}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \|\cdot\|_\infty)$
- We want to solve:

$$\mathcal{F}(f) = g \qquad \text{where} \qquad \mathcal{F} : X \to X, f \mapsto hf$$

$\Rightarrow$ $\mathcal{F}$ is linear

- Since $\mathcal{F}$ is linear, it coincides with its Fréchet derivative $\mathrm{D}\mathcal{F}_f : X \to X$

$$\mathcal{F}(f+\delta_f) = h(f+\delta_f) = \underbrace{hf}_{\mathcal{F}(f)} + \underbrace{h\delta_f}_{\mathrm{D}\mathcal{F}_f(\delta_f)} \qquad \rightsquigarrow \qquad \mathrm{D}\mathcal{F}_f(\delta_f) = \mathcal{F}(\delta_f) = h\delta_f$$

A Linear Example: Division of RPAs (2/3)

Compute $f = \dfrac{g}{h}, \quad h(x) \neq 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{h}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{h}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(hf - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \frac{g}{h}$$

A Linear Example: Division of RPAs (2/3)

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{h}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{h}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(hf - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \frac{g}{h}$$

- Check the contraction property of $\mathcal{T}$:

$$\begin{aligned} \|\mathcal{T}(f) - \mathcal{T}(f')\| &= \|[f - \widetilde{\varphi}(hf - g)] - [f' - \widetilde{\varphi}(hf' - g)]\| \\ &= \|(1 - \widetilde{\varphi}h)(f - f')\| \quad \leq \quad \underbrace{\|1 - \widetilde{\varphi}h\|}_{:=\lambda} \|f - f'\| \end{aligned}$$

A Linear Example: Division of RPAs (2/3)

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{h}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{h}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(hf - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \frac{g}{h}$$

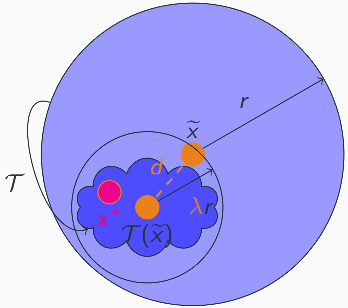
- Check the contraction property of $\mathcal{T}$:

$$\begin{aligned} \|\mathcal{T}(f) - \mathcal{T}(f')\| &= \|[f - \widetilde{\varphi}(hf - g)] - [f' - \widetilde{\varphi}(hf' - g)]\| \\ &= \|(1 - \widetilde{\varphi}h)(f - f')\| \quad \leq \quad \underbrace{\|1 - \widetilde{\varphi}h\|}_{:=\lambda} \|f - f'\| \end{aligned}$$

$\Rightarrow$ We need $\lambda = \|1 - \widetilde{\varphi}h\| < 1$

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Check the stability condition to apply the Banach fixed-point theorem



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{f}, r)) \subseteq B(\widetilde{f}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

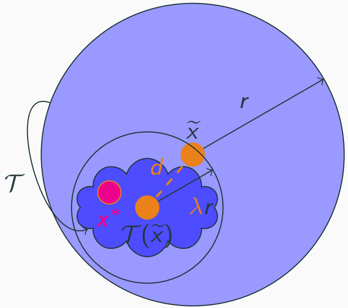
Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

- The stability condition is encoded as:

$$d + \lambda r \leq r, \qquad d = \|\mathcal{T}(\widetilde{f}) - \widetilde{f}\| = \|\widetilde{\varphi}(h\widetilde{f} - g)\|$$

Compute $f = \dfrac{g}{h}$, $\quad h(x) \neq 0$ over I:

- Check the stability condition to apply the Banach fixed-point theorem



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{f}, r)) \subseteq B(\widetilde{f}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

- The stability condition is encoded as:

$$d + \lambda r \leq r, \qquad d = \|\mathcal{T}(\widetilde{f}) - \widetilde{f}\| = \|\widetilde{\varphi}(h\widetilde{f} - g)\|$$

$\Rightarrow$ the "best" bound r is $r = \dfrac{d}{1 - \lambda}$

Division of RPAs — The Algorithm

Algorithm RPADiv

- **Input:** RPAs $\boldsymbol{g}$ and $\boldsymbol{h}$, approximation degree $n \in \mathbb{N}$
- **Output:** degree-n RPA $\boldsymbol{f}$ representing $\frac{\boldsymbol{g}}{\boldsymbol{h}}$ rigorously

1. Compute $\widetilde{f} \approx \frac{g}{h}$ using degree-n Chebyshev interpolation
2. Compute $\widetilde{\varphi} \approx \frac{1}{h}$ using degree-n Chebyshev interpolation
3. Compute $\lambda = \|1 - \widetilde{\varphi}h\|$ and FAIL if $\lambda \geq 1$
4. Compute $d = \|\widetilde{\varphi}(h\widetilde{f} - g)\|$
5. Compute $r = \frac{d}{1-\lambda}$ and RETURN RPA $\boldsymbol{f} = (\widetilde{f}, r)$

Algorithm RPADiv is correct

If RPADiv$(\boldsymbol{g}, \boldsymbol{h})$ does not fail, then it returns an RPA $\boldsymbol{f}$ such that for all $g \in \boldsymbol{g}, h \in \boldsymbol{h}$, we have $f = \frac{g}{h} \in \boldsymbol{f}$.

A Nonlinear Example: Square Root of an RPA (1/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Numerical interpolation $\quad \rightsquigarrow \quad \widetilde{f} \approx \sqrt{g}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \| \cdot \|_\infty)$

A Nonlinear Example: Square Root of an RPA (1/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Numerical interpolation $\quad\rightsquigarrow\quad \widetilde{f} \approx \sqrt{g}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \|\cdot\|_\infty)$
- We want to solve:

$$\mathcal{F}(f) = g \qquad \text{where} \qquad \mathcal{F} : X \to X, f \mapsto f^2$$

$\Rightarrow$ $\mathcal{F}$ is nonlinear (quadratic)

A Nonlinear Example: Square Root of an RPA (1/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Numerical interpolation $\quad \rightsquigarrow \quad \widetilde{f} \approx \sqrt{g}$ in $\mathbb{R}_n[x] \subseteq X = (\mathcal{C}(I), \|\cdot\|_\infty)$
- We want to solve:

$$\mathcal{F}(f) = g \qquad \text{where} \qquad \mathcal{F} : X \to X, f \mapsto f^2$$

$\Rightarrow$ $\mathcal{F}$ is nonlinear (quadratic)

- Since $\mathcal{F}$ is nonlinear, its Fréchet derivative $\mathrm{D}\mathcal{F}_f : X \to X$ depends on f:

$$\mathcal{F}(f + \delta_f) = (f + \delta_f)^2 = \underbrace{f^2}_{\mathcal{F}(f)} + \underbrace{2f\delta_f}_{\mathrm{D}\mathcal{F}_f(\delta_f)} + \delta_f^2 \qquad \rightsquigarrow \qquad \mathrm{D}\mathcal{F}_f(\delta_f) = 2f\delta_f$$

A Nonlinear Example: Square Root of an RPA (2/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

A Nonlinear Example: Square Root of an RPA (2/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{2f}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{2f} \approx \dfrac{1}{2\sqrt{g}}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(f^2 - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \pm\sqrt{g}$$

A Nonlinear Example: Square Root of an RPA (2/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{2f}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{2f} \approx \dfrac{1}{2\sqrt{g}}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(f^2 - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \pm\sqrt{g}$$

- Check the contraction property of $\mathcal{T}$, for $f, f' \in B(\widetilde{\ }, r)$:

$$\begin{aligned} \|\mathcal{T}(f) - \mathcal{T}(f')\| &= \|[f - \widetilde{\varphi}(f^2 - g)] - [f' - \widetilde{\varphi}({f'}^2 - g)]\| \\ &= \|[1 - \widetilde{\varphi}(f + f')](f - f')\| \quad \leq \quad \underbrace{\|1 - \widetilde{\varphi}(f + f')\|}_{\leq \lambda(r)} \|f - f'\| \end{aligned}$$

A Nonlinear Example: Square Root of an RPA (2/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{2f}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{2f} \approx \dfrac{1}{2\sqrt{g}}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(f^2 - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \pm\sqrt{g}$$

- Check the contraction property of $\mathcal{T}$, for $f, f' \in B(\widetilde{\ }, r)$:

$$\begin{aligned} \|\mathcal{T}(f) - \mathcal{T}(f')\| &= \|[f - \widetilde{\varphi}(f^2 - g)] - [f' - \widetilde{\varphi}(f'^2 - g)]\| \\ &= \|[1 - \widetilde{\varphi}(f + f')](f - f')\| \quad \leq \quad \underbrace{\|1 - \widetilde{\varphi}(f + f')\|}_{\leq \lambda(r)} \|f - f'\| \end{aligned}$$

$\Rightarrow$ $\mathcal{T}$ is $\lambda(r)$ contracting over $B(\widetilde{f}, r)$, where:

$$\lambda(r) := \underbrace{\|1 - 2\widetilde{\varphi}\widetilde{f}\|}_{\lambda_0} + \underbrace{2\|\widetilde{\varphi}\|}_{\lambda_1} r$$

A Nonlinear Example: Square Root of an RPA (2/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Construct a Newton-like fixed-point operator $\mathcal{T} : X \to X$:

$$\mathcal{T}(f) = f - \mathcal{A}(\mathcal{F}(f) - g) \qquad \text{where} \quad \mathcal{A} \approx \mathrm{D}\mathcal{F}_f^{-1} : X \to X$$

- $\mathrm{D}\mathcal{F}_f^{-1} : \delta_f \mapsto \dfrac{\delta_f}{2f}$, so we define $\mathcal{A}(\delta_f) = \widetilde{\varphi}\delta_f$ using $\widetilde{\varphi} \approx \dfrac{1}{2f} \approx \dfrac{1}{2\sqrt{g}}$

$$\mathcal{T}(f) = f - \widetilde{\varphi}(f^2 - g) \qquad \rightsquigarrow \qquad \mathcal{T}(f) = f \quad \Leftrightarrow \quad f = \pm\sqrt{g}$$

- Check the contraction property of $\mathcal{T}$, for $f, f' \in B(\widetilde{\;}, r)$:

$$\begin{aligned} \|\mathcal{T}(f) - \mathcal{T}(f')\| &= \|[f - \widetilde{\varphi}(f^2 - g)] - [f' - \widetilde{\varphi}(f'^2 - g)]\| \\ &= \|[1 - \widetilde{\varphi}(f + f')](f - f')\| \quad \leq \quad \underbrace{\|1 - \widetilde{\varphi}(f + f')\|}_{\leq \lambda(r)} \|f - f'\| \end{aligned}$$

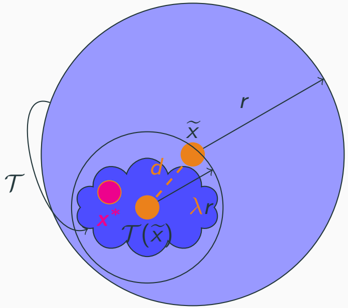
$\Rightarrow$ $\mathcal{T}$ is $\lambda(r)$ contracting over $B(\widetilde{f}, r)$, where:

$$\lambda(r) := \underbrace{\|1 - 2\widetilde{\varphi}\widetilde{f}\|}_{\lambda_0} + \underbrace{2\|\widetilde{\varphi}\|}_{\lambda_1} r$$

$\Rightarrow$ r must be small enough to ensure $\lambda(r) \leq 1$

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Check the stability condition to apply the Banach fixed-point theorem



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{f}, r)) \subseteq B(\widetilde{f}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

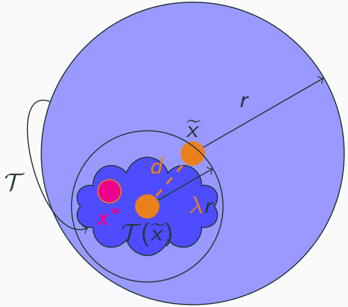
Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

The stability condition is encoded as:

$$d + \lambda(r)r \le r, \quad d = \|\mathcal{T}(\widetilde{f}) - \widetilde{f}\| = \|\widetilde{\varphi}(\widetilde{f}^2 - g)\|$$

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Check the stability condition to apply the Banach fixed-point theorem



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{f}, r)) \subseteq B(\widetilde{f}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

The stability condition is encoded as:

$$d + \lambda(r) r \leq r, \quad d = \|\mathcal{T}(\widetilde{f}) - \widetilde{f}\| = \|\widetilde{\varphi}(\widetilde{f}^2 - g)\|$$

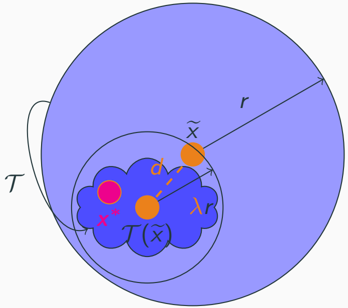
- $\lambda_1 r^2 + (\lambda_0 - 1) r + d \leq 0$ has positive real solutions iff:

$$\Delta := (1 - \lambda_0)^2 - 4\lambda_1 d \geq 0$$

A Nonlinear Example: Square Root of an RPA (3/3)

Compute $f = \sqrt{g}, \quad g(x) > 0$ over I:

- Check the stability condition to apply the Banach fixed-point theorem



Banach Fixed-Point Theorem

If one can rigorously check:

- $\mathcal{T}(B(\widetilde{f}, r)) \subseteq B(\widetilde{f}, r)$
- $\mathcal{T}$ is λ-contracting* over $B(\widetilde{x}, r)$ with $\lambda < 1$

Then $\mathcal{T}$ has a unique fixed-point x^* in $B(\widetilde{x}, r)$

The stability condition is encoded as:

$$d + \lambda(r) r \leq r, \quad d = \|\mathcal{T}(\widetilde{f}) - \widetilde{f}\| = \|\widetilde{\varphi}(\widetilde{f}^2 - g)\|$$

- $\lambda_1 r^2 + (\lambda_0 - 1) r + d \leq 0$ has positive real solutions iff:

$$\Delta := (1 - \lambda_0)^2 - 4\lambda_1 d \geq 0$$

$\Rightarrow$ Return the smallest root: $r := \dfrac{1 - \lambda_0 - \sqrt{\Delta}}{2\lambda_1}$

Square Root of an RPA — The Algorithm

Algorithm RPASqrt

- **Input:** RPA $\boldsymbol{g}$, approximation degree $n \in \mathbb{N}$
- **Output:** degree-n RPA $\boldsymbol{f}$ representing $\sqrt{\boldsymbol{g}}$ rigorously

---

1. Compute $\widetilde{f} \approx \sqrt{g}$ using degree-n Chebyshev interpolation
2. Compute $\widetilde{\varphi} \approx \dfrac{1}{2\widetilde{f}}$ using degree-n Chebyshev interpolation
3. Compute $\lambda_0 = \|1 - 2\widetilde{\varphi}\widetilde{f}\|$ and FAIL if $\lambda_0 \geq 1$
4. Compute $\lambda_1 = 2\|\widetilde{\varphi}\|$
5. Compute $d = \|\widetilde{\varphi}(\widetilde{f}^2 - g)\|$
6. Compute $\Delta = (1 - \lambda_0)^2 - 4\lambda_1 d$ and FAIL if $\Delta < 0$
5. Compute $r := \dfrac{1 - \lambda_0 - \sqrt{\Delta}}{2\lambda_1}$ and RETURN RPA $\boldsymbol{f} = (\widetilde{f}, r)$

Algorithm RPADiv is correct

If RPASqrt($\boldsymbol{g}$) does not fail, then it returns an RPA $\boldsymbol{f}$ such that for all $g \in \boldsymbol{g}$, we have $f = \sqrt{g} \in \boldsymbol{f}$.

Further Examples of Fixed-Point A Posteriori Validation (1/2)

- Roots of univariate polynomials:

$$p(x) = 0 \quad \rightsquigarrow \quad z_1, \ldots, z_n \quad \text{s.t.} \quad p(x) = (x - z_1) \ldots (x - z_n)$$

- Roots of univariate analytic functions:

$$f(x) = 0 \quad \rightsquigarrow \quad \text{isolate one/all root(s) of } f$$

- Roots of systems of multivariate polynomial/analytic functions:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \qquad\qquad \vdots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases} \quad \rightsquigarrow \quad \text{isolate one/all solutions}$$

See for instance: Siegfried M. Rump, Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*. 2010;19:287-449.

Further Examples of Fixed-Point A Posteriori Validation (2/2)

- Solving linear systems:

$$Ax = b \quad \rightsquigarrow \quad \text{recover } x = (x_1, \ldots, x_n)$$

- Eigenvalue problems:

$$Av = \lambda v \quad \rightsquigarrow \quad \text{recover eigenvalue/eigenvector pair } (\lambda, v)$$

$$A \quad \rightsquigarrow \quad \text{diagonalize } A = PDP^{-1} : \quad \begin{cases} D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \text{eigenvalues} \\ \qquad P \in \mathbb{C}^{n \times n} = \text{eigenvectors} \end{cases}$$

- Ordinary differential equations (ODEs):

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(0) = v \in \mathbb{R}^N \end{cases} \quad \rightsquigarrow \quad \text{compute } \boldsymbol{p} = (p, \varepsilon) \text{ for } y$$

- Partial Differential Equations (PDEs), Delay Differential Equations (DDEs), etc.

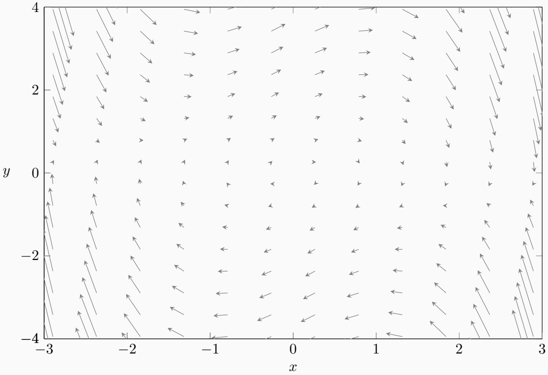
See for instance: Siegfried M. Rump, Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*. 2010;19:287-449.

Rigorous Numerics for Hilbert's 16th Problem

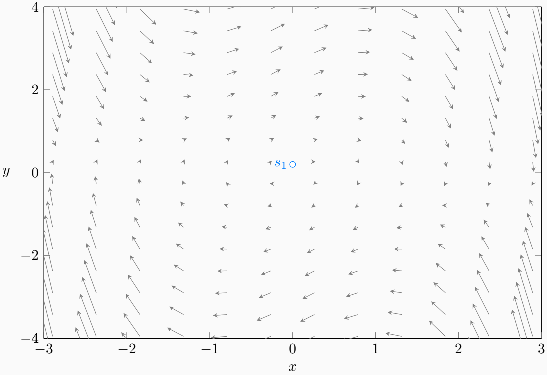
Florent Bréhard, Nicolas Brisebarre, Mioara Joldes, Damien Pous, Warwick Tucker
florent.brehard@univ-lille.fr

Thursday, October 24, 2024

$$\begin{cases} \dot{x} = P(x,y) \\ \dot{y} = Q(x,y) \end{cases} \qquad P, Q \in \mathbb{R}[x,y]$$

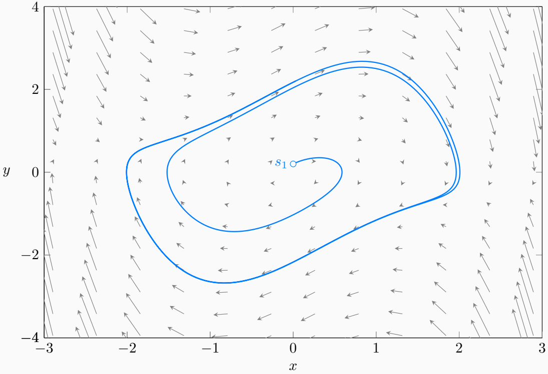


$$\begin{cases} \dot{x} = P(x,y) \\ \dot{y} = Q(x,y) \end{cases} \qquad P, Q \in \mathbb{R}[x,y]$$



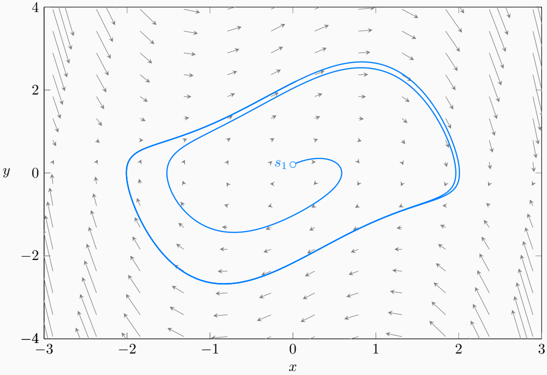
Polynomial Vector Fields in the Plane

$$\begin{cases} \dot{x} = P(x, y) \\ \dot{y} = Q(x, y) \end{cases} \qquad P, Q \in \mathbb{R}[x, y]$$



Polynomial Vector Fields in the Plane

$$\begin{cases} \dot{x} = P(x,y) \\ \dot{y} = Q(x,y) \end{cases} \qquad P, Q \in \mathbb{R}[x,y]$$



Examples:

- $(v, a) = (\dot{x}, \dot{v}) = (v, Q(x, v))$ in mechanics
- $(\dot{u}, \dot{i}) = (P(u, i), Q(u, i))$ in electricity

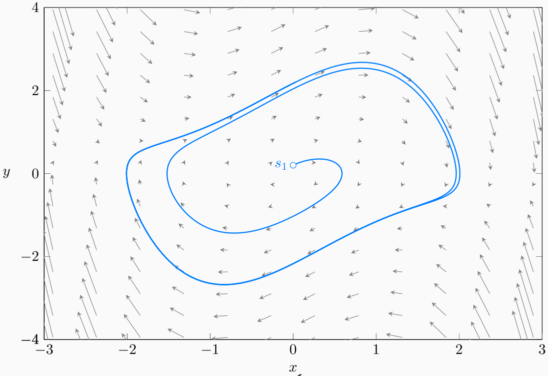
Limit Cycles

limit cycle = isolated periodic orbit

(For example, $(\dot{x}, \dot{y}) = (-y, x)$ produces a continuum of periodic orbits but no limit cycles!)

limit cycle = isolated periodic orbit

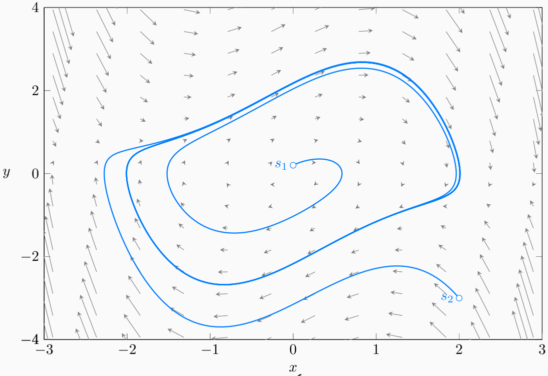
(For example, $(\dot{x}, \dot{y}) = (-y, x)$ produces a continuum of periodic orbits but no limit cycles!)



Van der Pol oscillator $\begin{cases} \dot{x} = y \\ \dot{y} = \mu(1 - x^2)y - x \end{cases}$

limit cycle = isolated periodic orbit

(For example, $(\dot{x}, \dot{y}) = (-y, x)$ produces a continuum of periodic orbits but no limit cycles!)

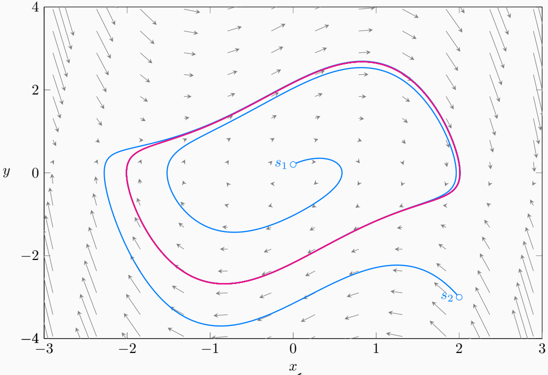


Van der Pol oscillator $\begin{cases} \dot{x} = y \\ \dot{y} = \mu(1 - x^2)y - x \end{cases}$

Limit Cycles

limit cycle = isolated periodic orbit

(For example, $(\dot{x}, \dot{y}) = (-y, x)$ produces a continuum of periodic orbits but no limit cycles!)

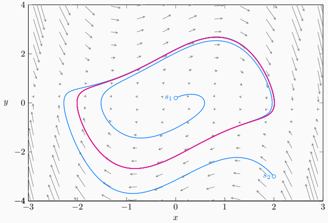


Van der Pol oscillator $\begin{cases} \dot{x} = y \\ \dot{y} = \mu(1 - x^2)y - x \end{cases}$

Hilbert's 16th problem (second part)

For a given integer n, what is the maximum number $\mathcal{H}(n)$ of limit cycles a polynomial vector field of degree at most n in the plane can have?

D. Hilbert, International Congress of Mathematicians, Paris, 1900

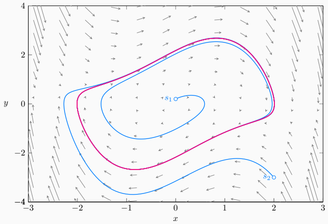


Hilbert's 16th problem (second part)

For a given integer n, what is the maximum number $\mathcal{H}(n)$ of limit cycles a polynomial vector field of degree at most n in the plane can have?

D. Hilbert, International Congress of Mathematicians, Paris, 1900

- 1923: H. Dulac (incorrectly) proved that a *single* polynomial vector field has a finite number of limit cycles
- 1981: Y. S. Il'Yashenko found a major gap in Dulac's proof
- 1991: New proofs of Dulac's result by Y. S. Il'Yashenko and J. Écalle
- But even $\mathcal{H}(2) < \infty$ is open!



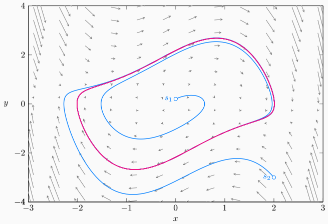
Hilbert's 16th Problem

Hilbert's 16th problem (second part)

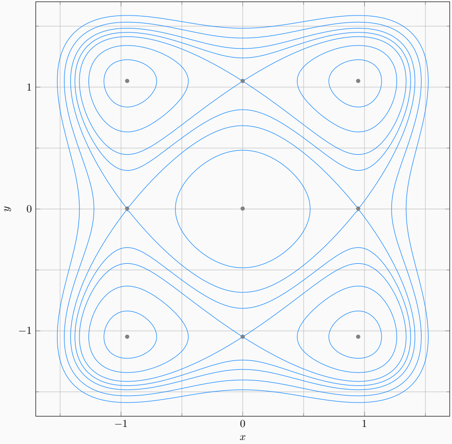
For a given integer n, what is the maximum number $\mathcal{H}(n)$ of limit cycles a polynomial vector field of degree at most n in the plane can have?

D. Hilbert, International Congress of Mathematicians, Paris, 1900

- 1923: H. Dulac (incorrectly) proved that a *single* polynomial vector field has a finite number of limit cycles
- 1981: Y. S. Il'Yashenko found a major gap in Dulac's proof
- 1991: New proofs of Dulac's result by Y. S. Il'Yashenko and J. Écalle
- But even $\mathcal{H}(2) < \infty$ is open!
- Some lower bounds: $\mathcal{H}(2) \geq 4$, $\mathcal{H}(3) \geq 13$, $\mathcal{H}(4) \geq 28$.

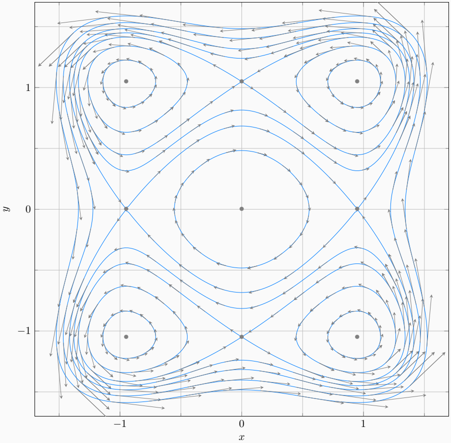


Infinitesimal Hilbert's 16th Problem



$$H(x,y) = (x^2 - 0.9)^2 + (y^2 - 1.1)^2$$

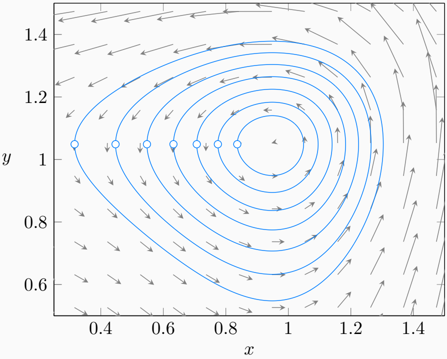
T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011



$$H(x,y) = (x^2 - 0.9)^2 + (y^2 - 1.1)^2$$

$$\begin{cases} \dot{x} = -\partial_y H(x,y) = 4y(y^2 - 1.1) \\ \dot{y} = \partial_x H(x,y) = 4x(x^2 - 0.9) \end{cases}$$

T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011

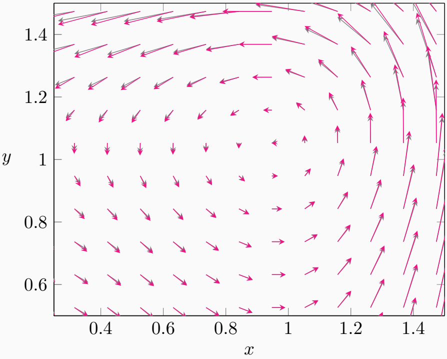


$$H(x,y) = (x^2 - 0.9)^2 + (y^2 - 1.1)^2$$

$$\begin{cases} \dot{x} = 4y(y^2 - 1.1) \\ \dot{y} = 4x(x^2 - 0.9) \end{cases}$$

T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011

Infinitesimal Hilbert's 16th Problem

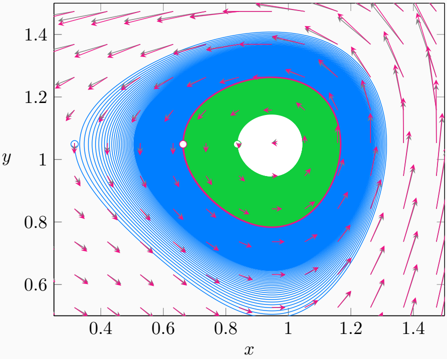


$$H(x, y) = (x^2 - 0.9)^2 + (y^2 - 1.1)^2$$

$$\begin{cases} \dot{x} = 4y(y^2 - 1.1) \\ \dot{y} = 4x(x^2 - 0.9) - 0.4y + 0.46x^2y \end{cases}$$

T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011

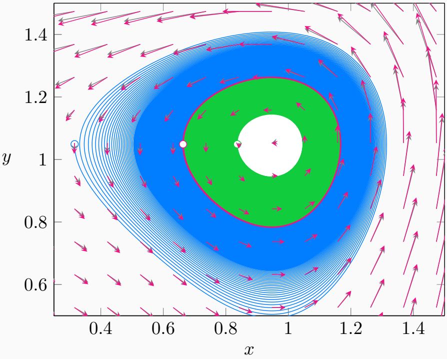
Infinitesimal Hilbert's 16th Problem



$$H(x, y) = (x^2 - 0.9)^2 + (y^2 - 1.1)^2$$

$$\begin{cases} \dot{x} = 4y(y^2 - 1.1) \\ \dot{y} = 4x(x^2 - 0.9) - 0.4y + 0.46x^2y \end{cases}$$

T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011



T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011

Infinitesimal Hilbert's 16th problem

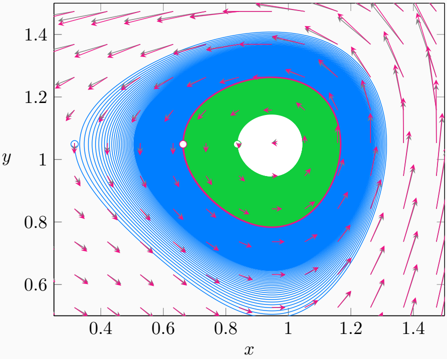
For a given integer n, what is the maximal number $\mathcal{Z}(n)$ of limit cycles a perturbed Hamiltonian vector field of the form:

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

can have when $\varepsilon \to 0$, with:

- $H(x,y)$ a polynomial potential function of degree $n+1$
- f,g polynomial perturbations of degree n

Infinitesimal Hilbert's 16th Problem



T. Johnson, A quartic system with twenty-six limit cycles, *Experimental Mathematics*, 2011

Infinitesimal Hilbert's 16th problem

For a given integer n, what is the maximal number $\mathcal{Z}(n)$ of limit cycles a perturbed Hamiltonian vector field of the form:

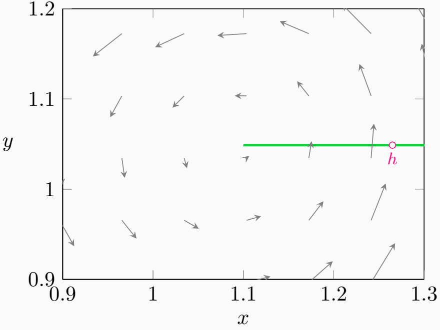
$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

can have when $\varepsilon \to 0$, with:

- $H(x,y)$ a polynomial potential function of degree $n+1$
- f, g polynomial perturbations of degree n

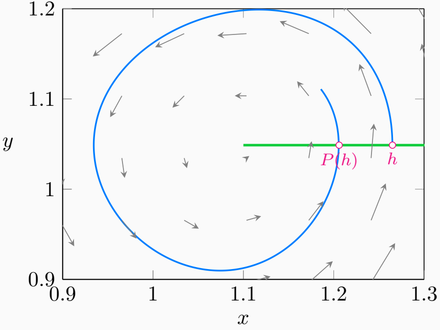
- $\mathcal{Z}(n) < \infty$ for all n
- Pessimistic upper bounds

A Fundamental Tool: the Poincaré-Pontryagin Theorem



$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

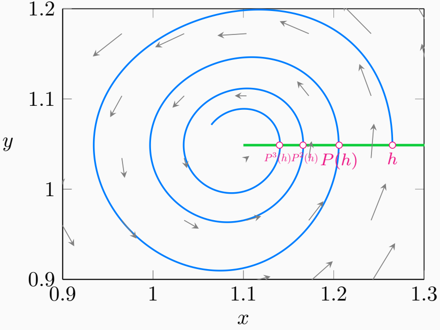
A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$

$$\begin{cases} \dot{x} = -\partial_y H(x, y) + \varepsilon f(x, y) \\ \dot{y} = \partial_x H(x, y) + \varepsilon g(x, y) \end{cases}$$

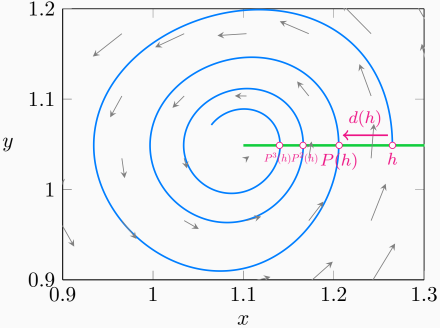
A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

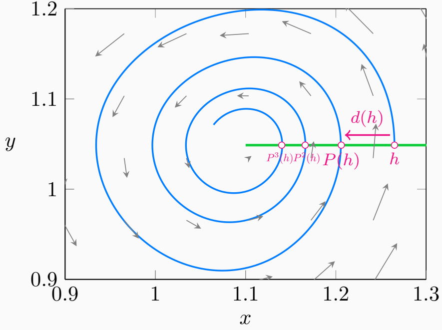
A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$
- Displacement $d(h) = P(h) - h$

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

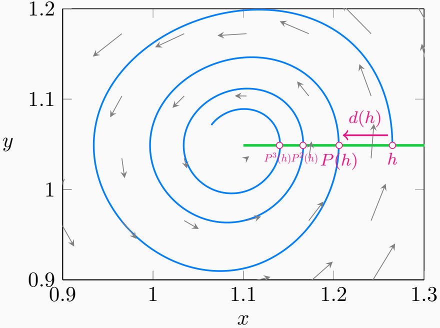
A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$
- Displacement $d(h) = P(h) - h$
- Limit cycle $\Leftrightarrow$ isolated zero of d

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

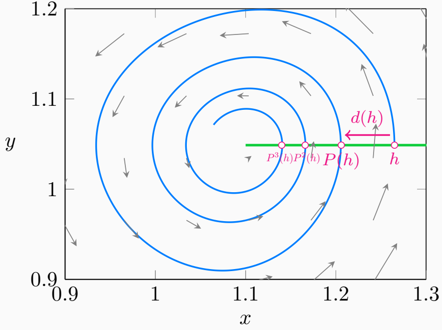
A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$
- Displacement $d(h) = P(h) - h$
- Limit cycle $\Leftrightarrow$ isolated zero of d
- Abelian integral $\mathcal{I}(h)$:

$$\oint_{H^{-1}(h)} f(x,y)\mathrm{d}y - g(x,y)\mathrm{d}x$$

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$



- Poincaré first return map $P(h)$
- Displacement $d(h) = P(h) - h$
- Limit cycle $\Leftrightarrow$ isolated zero of d
- Abelian integral $\mathcal{I}(h)$:

$$\oint_{H^{-1}(h)} f(x,y)\mathrm{d}y - g(x,y)\mathrm{d}x$$

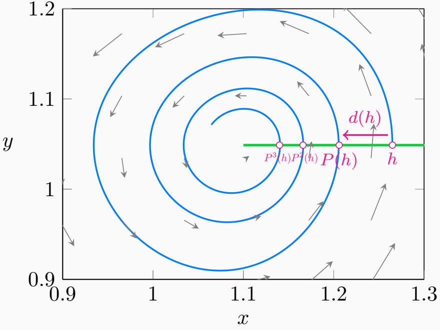
$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

Poincaré-Pontryagin theorem

The Abelian integral $\mathcal{I}(h)$ approximates the displacement function $d(h)$ for small ε:

$$d(h) = \varepsilon(\mathcal{I}(h) + O(\varepsilon)) \qquad \text{when } \varepsilon \to 0$$

A Fundamental Tool: the Poincaré-Pontryagin Theorem



- Poincaré first return map $P(h)$
- Displacement $d(h) = P(h) - h$
- Limit cycle $\Leftrightarrow$ isolated zero of d
- Abelian integral $\mathcal{I}(h)$:

$$\oint_{H^{-1}(h)} f(x,y)\mathrm{d}y - g(x,y)\mathrm{d}x$$

$$\begin{cases} \dot{x} = -\partial_y H(x,y) + \varepsilon f(x,y) \\ \dot{y} = \partial_x H(x,y) + \varepsilon g(x,y) \end{cases}$$

Poincaré-Pontryagin theorem

The Abelian integral $\mathcal{I}(h)$ approximates the displacement function $d(h)$ for small ε:

$$d(h) = \varepsilon(\mathcal{I}(h) + O(\varepsilon)) \qquad \text{when } \varepsilon \to 0$$

limit cycles $\equiv$ changes of sign of $\mathcal{I}(h)$ $\equiv$ simple zeros of $\mathcal{I}(h)$