A EXAMPLES OF COMPUTATIONS

A.1 STEP BY STEP EXAMPLE : AUTONOMOUS CONTROL

To measure whether the system

$$\frac{dx_1(t)}{dt} = \sin(x_1^2) + \log(1+x_2) + \frac{\operatorname{atan}(ux_1)}{1+x_2}$$
$$\frac{dx_2(t)}{dt} = x_2 - e^{x_1 x_2},$$

is controllable at a point x_e, with asymptotic control u_e, using Kalman condition we need to

1. differentiate the system with respect to its internal variables, obtain the Jacobian $A(x,u)$

$$A(x,u) = \begin{pmatrix} 2x_1\cos(x_1^2) + \frac{u(1+x_2)^{-1}}{1+u^2x_1^2} & (1+x_2)^{-1} - \frac{\operatorname{atan}(ux_1)}{(1+x_2)^2} \\ -x_2 e^{x_1x_2} & 1 - x_1 e^{x_1x_2} \end{pmatrix}$$

2. differentiate the system with respect to its control variables, obtain a matrix $B(x,u)$

$$B(x,u) = \begin{pmatrix} x_1((1+u^2x_1^2)(1+x_2))^{-1} \\ 0 \end{pmatrix}$$

3. evaluate A and B in $x_e = [0.5]$, $u_e = 1$

$$A(x_e,u_e) = \begin{pmatrix} 1.50 & 0.46 \\ -0.64 & 0.36 \end{pmatrix}, \quad B(x_e,u_e) = \begin{pmatrix} 0.27 \\ 0 \end{pmatrix}$$

4. calculate the controllability matrix given by (2).

$$C = [B, AB]((x_e,u_e)) = \left[\begin{pmatrix} 0.27 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.50 & 0.46 \\ -0.64 & 0.36 \end{pmatrix}\begin{pmatrix} 0.27 \\ 0 \end{pmatrix}\right] = \begin{pmatrix} 0.27 & 0.40 \\ 0 & -0.17 \end{pmatrix}$$

5. output $n-d$, with d the rank of the controllability matrix, the system is controllable if $n-d=0$

$$n - \operatorname{rank}(C) = 2 - 2 = 0 : \text{ System is controllable in } (x_e = [0.5], u_e = 1)$$

6. (optionally) if $n-d=0$, compute the control feedback matrix K as in (3)

$$K = \begin{pmatrix} -22.8 & 44.0 \end{pmatrix}.$$

A.2 STEP BY STEP EXAMPLE: STABILITY OF LINEAR PDE

To find the existence and behavior at infinite time of a solution, given a differential operator D_x and an initial condition u_0 we proceed as follows

1. find the Fourier polynomial $f(\xi)$ associated to D_x

$$D_x = 2\partial^2_{x_0} + 0.5\partial^2_{x_1} + \partial^4_{x_2} - 7\partial^2_{x_0,x_1} - 1.5\partial_{x_1}\partial^2_{x_2},$$
$$f(\xi) = -4\pi\xi_0^2 - \pi\xi_1^2 + 2\pi\xi_2^4 + 14\pi\xi_0\xi_1 + 3i\pi\xi_1\xi_2^2$$

2. find the Fourier transform $\tilde{u}_0(\xi)$ of u_0

$$u_0(x) = e^{-3ix_2}x_0^{-1}\sin(x_0)e^{2.5ix_1}e^{-x_2^2},$$
$$\tilde{u}_0(\xi) = \pi^{3/2}\mathbf{1}_{[-(2\pi)^{-1},(2\pi)^{-1}]}(\xi_0)\delta_0(\xi_1 - 2.5(2\pi)^{-1})e^{-\pi^2(\xi_2+3(2\pi)^{-1})^2}$$

3. find the set $\mathcal{F}$ of frequency ξ for which $\tilde{u}_0(\xi) \neq 0$

$$\mathcal{F} = [-(2\pi)^{-1},(2\pi)^{-1}] \times \{2.5(2\pi)^{-1}\} \times (-\infty,+\infty)$$

4. minimize $f(\xi)$ on $\mathcal{F}$

   $\min_{\mathcal{F}}(f(\xi)) = -22.6$

5. output (0,0) if this minimum is infinite, (1,0) is finite and negative, (1,1) if finite and positive. (optionally) output $\mathcal{F}$

   $\text{Out} = (1,0)$: there exists a solution u ; it does not vanish at $t \to +\infty$

A.3 Examples of inputs and outputs

A.3.1 Local stability

System	Speed of convergence at $x_e = [0.01]$
$\begin{cases} \frac{d}{dt}x_0 = -\frac{x_1}{\operatorname{atan}(8x_0x_2)} + \frac{0.01}{\operatorname{atan}(0.0008)} \\ \frac{d}{dt}x_1 = -\cos(9x_0) + \cos(0.09) \\ \frac{d}{dt}x_2 = x_0 - \sqrt{x_1 + x_2} - 0.01 + 0.1\sqrt{2} \end{cases}$	-1250
$\begin{cases} \frac{d}{dt}x_0 = -\frac{2x_2}{x_0 - 2x_2(x_1 - 5)} + 0.182 \\ \frac{d}{dt}x_1 = (x_1 + (x_2 - e^{x_1})(\tan(x_0) + 3))(\log(3) + i\pi) \\ \qquad +3.0\log(3) + 3.0i\pi \\ \frac{d}{dt}x_2 = \operatorname{asin}\left(x_0 \log\left(-\frac{4}{x_1}\right)\right) - \operatorname{asin}(0.06 + 0.01i\pi) \end{cases}$	-0.445
$\begin{cases} \frac{d}{dt}x_0 = e^{x_1 + e^{-\sin\left(x_0 - e^2\right)}} - 1.01e^{e^{-\sin\left(0.01 - e^2\right)}} \\ \frac{d}{dt}x_1 = 0.06 - 6x_1 \\ \frac{d}{dt}x_2 = -201 + \frac{x_0 + 2}{x_0^2 x_2} \end{cases}$	6.0 (locally stable)
$\begin{cases} \frac{d}{dt}x_0 = x_2 e^{-x_1}\sin(x_1) - 9.9 \cdot 10^{-5} \\ \frac{d}{dt}x_1 = 7.75.10^{-4} - \frac{e^{x_2}\operatorname{atan}(\operatorname{atan}(x_1))}{4e^{x_2} + 9} \\ \frac{d}{dt}x_2 = (x_1 - \operatorname{asin}(9))\, e^{-\frac{x_0}{\log(3) + i\pi}} \\ \qquad -(0.01 - \operatorname{asin}(9))\, e^{-\frac{0.01}{\log(3) + i\pi}} \end{cases}$	-0.0384
$\begin{cases} \frac{d}{dt}x_0 = -\frac{x_0\left(7 - \sqrt[4]{7}\sqrt{i}\right)}{9} - x_1 + 0.0178 - 0.00111\sqrt[4]{7}\sqrt{i} \\ \frac{d}{dt}x_1 = -0.000379 + e^{-\frac{63}{\cos((x_2 - 9)\operatorname{atan}(x_1)) + 7}} \\ \frac{d}{dt}x_2 = -x_0 - x_1 + \operatorname{asin}\left(\cos(x_0) + \frac{x_2}{x_0}\right) \\ \qquad -1.55 + 1.32i \end{cases}$	$3.52.10^{-11}$ (locally stable)

A.3.2 CONTROLLABILITY: AUTONOMOUS SYSTEMS

Autonomous system	Dimension of uncontrollable space at $x_e = [0.5]$, $u_e = [0.5]$
$\begin{cases} \frac{dx_0}{dt} = -\operatorname{asin}\left(\frac{x_1}{9} - \frac{4\tan(\cos(10))}{9}\right) \\ \qquad\qquad -\operatorname{asin}\left(\frac{4\tan(\cos(10))}{9} - 0.0556\right) \\ \frac{dx_1}{dt} = u - x_2 + \log\left(10 + \frac{\tan(x_1)}{u+x_0}\right) - 2.36 \\ \frac{dx_2}{dt} = 2x_1 + x_2 - 1.5 \end{cases}$	0 (controllable)
$\begin{cases} \frac{dx_0}{dt} = u - \operatorname{asin}(x_0) - 0.5 + \frac{\pi}{6} \\ \frac{dx_1}{dt} = x_0 - x_1 + 2x_2 + \operatorname{atan}(x_0) - 1.46 \\ \frac{dx_2}{dt} = \frac{5x_2}{\cos(x_2)} - 2.85 \end{cases}$	1
$\begin{cases} \frac{dx_0}{dt} = 6u + 6x_0 - \frac{6x_1}{x_0} \\ \frac{dx_1}{dt} = 0.75 + x_1^2 - \cos(u - x_2) \\ \frac{dx_2}{dt} = -x_0^2 + x_0 + \log(e^{x_2}) - 0.75 \end{cases}$	2
$\begin{cases} \frac{dx_0}{dt} = +x_0\left(\cos\left(\frac{u}{x_0+2x_2}\right) + \frac{\operatorname{asin}(u)}{x_1}\right) \\ \qquad\qquad -0.5\cos\left(\frac{1}{3}\right) - \frac{\pi}{6} \\ \frac{dx_1}{dt} = \frac{\pi x_1}{4(x_2+4)} - \frac{\pi}{36} \\ \frac{dx_2}{dt} = 2.5 - 108e^{0.5} - 12x_0x_2 + x_1 + 108e^u \end{cases}$	0 (controllable)
$\begin{cases} \frac{dx_0}{dt} = -10\sin\left(\frac{3x_0}{\log(8)} - 22\right) - 6.54 \\ \frac{dx_1}{dt} = \sin\left(9 + \frac{-x_1-4}{8x_2}\right) - 1 \\ \frac{dx_2}{dt} = 4\tan\left(\frac{4x_0}{u}\right) - 4\tan(4) \end{cases}$	1

A.3.3 CONTROLLABILITY: NON-AUTONOMOUS SYSTEMS

Non-autonomous system	Local controllability at $x_e = [0.5]$, $u_e = [0.5]$
$\begin{cases} \frac{dx_0}{dt} = (x_2 - 0.5)\, e^{-\operatorname{asin}(8)} \\ \frac{dx_1}{dt} = e^{t+0.5} - e^{t+x_1} + \frac{-x_1 + e^{\frac{x_0}{u}}}{x_2} + 1 - 2e \\ \frac{dx_2}{dt} = t(x_2 - 0.5)\left(\operatorname{asin}(6) + \sqrt{\tan(8)}\right) \end{cases}$	False
$\begin{cases} \frac{dx_0}{dt} = \frac{\operatorname{atan}(\sqrt{x_2})}{x_0 - 1} - 2\operatorname{atan}\left(\frac{\sqrt{2}}{2}\right) \\ \frac{dx_1}{dt} = -\frac{u}{-\sqrt{x_0}x_1 + 3} + x_2 + \log(x_0) \\ \qquad + \log(2) - 0.5 + (1/(6 - \sqrt{2})) \\ \frac{dx_2}{dt} = -70t(x_0 - 0.5) \end{cases}$	False
$\begin{cases} \frac{dx_0}{dt} = \frac{x_0 + 7}{\sin(x_0 e^u) + 3} \\ \frac{dx_1}{dt} = -\frac{9x_2 e^{-\sin\left(\sqrt{\log(x_1)}\right)}}{x_0} \\ \frac{dx_2}{dt} = t + \operatorname{asin}(tx_2 + 4) \end{cases}$	False
$\begin{cases} \frac{dx_0}{dt} = 0.5 - x_2 + \tan(x_0) - \tan(0.5) \\ \frac{dx_1}{dt} = \frac{t}{x_1(t + \cos(x_1(t+u)))} - \frac{t}{0.5(t + \cos(0.5t + 0.25))} \\ \frac{dx_2}{dt} = 2.75 - x_0\,(u + 4) - x_0 \end{cases}$	True
$\begin{cases} \frac{dx_0}{dt} = u\,(u - x_0 - \tan(8)) + 0.5(\tan(8)) \\ \frac{dx_1}{dt} = -\frac{6t\left(-2 + \frac{\pi}{2}\right)}{x_0 x_1} - 12t\,(4 - \pi) \\ \frac{dx_2}{dt} = -7(u - 0.5) - 7\tan(\log(x_2)) \\ \qquad + 7\tan(\log(0.5)) \end{cases}$	True

A.3.4 STABILITY OF PARTIAL DIFFERENTIAL EQUATIONS USING FOURIER TRANSFORM

PDE $\partial_t u + D_x u = 0$ and initial condition	Existence of a solution, $u \to 0$ at $t \to +\infty$
$\begin{cases} D_x = 2\partial_{x_0}\left(2\partial_{x_0}^4\partial_{x_2}^4 + 3\partial_{x_1}^3 + 3\partial_{x_1}^2\right) \\ u_0 = \delta_0(-18x_0)\delta_0(-62x_2)e^{89ix_0 - 8649x_1^2 + 89ix_1 - 59ix_2} \end{cases}$	False , False
$\begin{cases} D_x = -4\partial_{x_0}^4 - 5\partial_{x_0}^3 - 6\partial_{x_0}^2\partial_{x_1}^2\partial_{x_2}^2 + 3\partial_{x_0}^2\partial_{x_1} - 4\partial_{x_1}^6 \\ u_0 = (162x_0x_2)^{-1}\left(e^{i(-25x_0 + 96x_2)}\sin(54x_0)\sin(3x_2)\right) \end{cases}$	True , False
$\begin{cases} D_x = \partial_{x_1}\left(4\partial_{x_0}^5\partial_{x_1} + 4\partial_{x_0}^2 - 9\partial_{x_0}\partial_{x_2}^6\right. \\ \left. +2\partial_{x_1}^3\partial_{x_2}^5 - 4\partial_{x_1}^3\partial_{x_2}^4 - 2\partial_{x_2}\right) \\ u_0 = (33x_0)^{-1}\left(e^{86ix_0 - 56ix_1 - 16x_2^2 + 87ix_2}\sin(33x_0)\right) \end{cases}$	True , False
$\begin{cases} D_x = -6\partial_{x_0}^7\partial_{x_2}^2 + \partial_{x_0}^5\partial_{x_2}^6 - 9\partial_{x_0}^4\partial_{x_1}^2 - 9\partial_{x_0}^4\partial_{x_2}^4 \\ +7\partial_{x_0}^2\partial_{x_2}^6 + 4\partial_{x_0}^2\partial_{x_2}^5 - 6\partial_{x_1}^6 \\ u_0 = \delta_0(88x_1)e^{-2x_0(2312x_0 + 15i)} \end{cases}$	True , True

B Mathematical definitions

B.1 Notions of stability

Let us consider a system

$$\frac{dx(t)}{dt} = f(x(t)). \tag{7}$$

x_e is an attractor, if there exists $\rho > 0$ such that

$$|x(0) - x_e| < \rho \Longrightarrow \lim_{t \to +\infty} x(t) = x_e. \tag{8}$$

But, counter intuitive as it may seem, this is not enough for asymptotic stability to take place.

Definition B.1. *We say that x_e is a locally (asymptotically) stable equilibrium if the two following conditions are satisfied:*

(i) x_e is a stable point, i.e. for every $\varepsilon > 0$, there exists $\eta > 0$ such that

$$|x(0) - x_e| < \eta \Longrightarrow |x(t) - x_e| < \varepsilon, \ \forall\, t \geq 0. \tag{9}$$

(ii) x_e is an attractor, i.e. there exists $\rho > 0$ such that

$$|x(0) - x_e| < \rho \Longrightarrow \lim_{t \to +\infty} x(t) = x_e. \tag{10}$$

In fact, the SMT of Subsection 3.1 deals with an even stronger notion of stability, namely the exponential stability defined as follows:

Definition B.2. *We say that x_e is an exponentially stable equilibrium if x_e is locally stable equilibrium and, in addition, there exist $\rho > 0$, $\lambda > 0$, and $M > 0$ such that*

$$|x(0) - x_e| < \rho \Longrightarrow |x(t)| \leq M e^{-\lambda t} |x(0)|.$$

In this definition, λ is called the exponential convergence rate, which is the quantity predicted in our first task. Of course, if x_e is locally exponentially stable it is in addition locally asymptotically stable.

B.2 Controllability

We give here a proper mathematical definition of controllability. Let us consider a non-autonomous system

$$\frac{dx(t)}{dt} = f(x(t), u(t), t), \tag{11}$$

such that $f(x_e, u_e) = 0$.

Definition B.3. *Let $\tau > 0$, we say that the nonlinear system (11) is locally controllable at the equilibrium x_e in time τ with asymptotic control u_e if, for every $\varepsilon > 0$, there exists $\eta > 0$ such that, for every $(x_0, x_1) \in \mathbb{R}^n \times \mathbb{R}^n$ with $|x_0 - x_e| \leq \eta$ and $|x_1 - x_e| \leq \eta$ there exists a trajectory (x, u) such that*

$$\begin{aligned} x(0) = x_0, \quad x(\tau) = x_1 \\ |u(t) - u_e| \leq \varepsilon, \ \ \forall\, t \in [0, \tau]. \end{aligned} \tag{12}$$

An interesting remark is that if the system is autonomous, the local controllability does not depend on the time τ considered, which explains that it is not precised in Theorem 3.2.

B.3 Tempered distribution

We start by recalling the multi-index notation: let $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}^n$, $x \in \mathbb{R}^n$, and $f \in C^\infty(\mathbb{R}^n)$, we denote

$$\begin{aligned} x^\alpha &= x_1^{\alpha_1} \times \cdots \times x_n^{\alpha_n} \\ \partial_x^\alpha f &= \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n} f. \end{aligned} \tag{13}$$

α is said to be a multi-index and $|\alpha| = \sum_{i=1}^n |\alpha_i|$. Then we give the definition of the Schwartz functions:

Definition B.4. *A function $\phi \in C^\infty$ belongs to the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ if, for any multi-index α and β,*

$$\sup_{x \in \mathbb{R}^n} |x^\alpha \partial_x^\beta \phi| < +\infty. \tag{14}$$

Finally, we define the space of tempered distributions:

Definition B.5. *A tempered distribution $\phi \in \mathcal{S}'(\mathbb{R}^n)$ is a linear form u on $\mathcal{S}(\mathbb{R}^n)$ such that there exists $p > 0$ and $C > 0$ such that*

$$|\langle u, \phi \rangle| \leq C \sum_{|\alpha|, |\beta| < p} \sup_{x \in \mathbb{R}^n} |x^\alpha \partial_x^\beta \phi|, \quad \forall\, \phi \in \mathcal{S}(\mathbb{R}^n). \tag{15}$$

C Additional experiments

C.1 Prediction of speed of convergence with higher precision

In Section 5.1, λ is predicted with a 10% margin error. Prediction of λ to better accuracy can be achieved by training models on data rounded to 2, 3 or 4 significant digits, and measuring the number of exact predictions on the test sample. Overall, we predict λ with two significant digits in 59.2% of test cases. Table 8 summarizes the results for different precisions (for transformers with 6 layers and a dimensionality of 512).

Table 8: Exact prediction of local convergence speed to given precision.

	Degree 2	Degree 3	Degree 4	Degree 5	Degree 6	Overall
2 digits	83.5	68.6	55.6	48.3	40.0	59.2
3 digits	75.3	53.2	39.4	33.4	26.8	45.7
4 digits	62.0	35.9	25.0	19.0	14.0	31.3

D Proofs of theorems

D.1 Analysis of Problem 2

The proofs of Theorem 3.2, of validity of the feedback matrix given by the expression (3), and of the extension of Theorem 3.2 to the non-autonomous system given by condition (4) can be found in Coron (2007). We give here the key steps of the proof for showing that the matrix K given by (3) is a valid feedback matrix to illustrate the underlying mechanisms:

- Setting $V(x(t)) = x(t)^{tr} C_T^{-1} x(t)$, where x is solution to $x'(t) = f(x, u_e + K.(x - x_e))$, and
$$C_T = \left(e^{-AT} \left[\int_0^T e^{-At} B B^{tr} e^{-A^{tr} t} dt \right] e^{-A^{tr} T} \right). \tag{16}$$
- Showing, using the form of C_T, that
$$\frac{d}{dt}(V(x(t))) = -|B^{tr} C_T^{-1} x(t)|^2 - |B^{tr} e^{-TA^{tr}} C_T^{-1} x(t)|^2$$

- Showing that, if for any $t \in [0,T]$, $|B^{tr}C_T^{-1}x(t)|^2 = 0$, then for any $i \in \{0,...,n-1\}$,
$$x^{tr}C_T^{-1}A^iB = 0, \quad \forall\, t \in [0,T].$$
- Deducing from the controllability condition (2), that
$$x(t)^{tr}C_T^{-1} = 0, \quad \forall\, t \in [0,T].$$
and therefore from the invertibility of C_T^{-1},
$$x(t) = 0, \quad \forall\, t \in [0,T].$$
- Concluding from the previous and LaSalle invariance principle that the system is locally exponentially stable.

D.2 Analysis of Problem 3

In this section we prove Proposition 3.1. We study the problem

$$\partial_t u + \sum_{|\alpha|\leq k} a_\alpha \partial_x^\alpha u = 0 \text{ on } \mathbb{R}_+ \times \mathbb{R}^n, \tag{17}$$

with initial condition

$$u(0,\cdot) = u_0 \in \mathcal{S}'(\mathbb{R}^n), \tag{18}$$

and we want to find a solution $u \in C^0([0,T],\mathcal{S}'(\mathbb{R}^n))$.

Denoting $\widetilde{u}$ the Fourier transform of u with respect to x, the problem is equivalent to

$$\partial_t \widetilde{u}(t,\xi) + \sum_{|\alpha|\leq k} a_\alpha (i\xi)^\alpha \widetilde{u}(t,\xi) = 0, \tag{19}$$

with initial condition $\widetilde{u}_0 \in \mathcal{S}(\mathbb{R}^n)$. As the only derivative now is with respect to time, we can check that

$$\widetilde{u}(t,\xi) = \widetilde{u}_0(\xi) e^{-f(\xi)t}, \tag{20}$$

where $f(\xi) = \sum_{|\alpha|\leq k} a_\alpha (i\xi)^\alpha$, is a weak solution to (19) belonging to the space $C^0([0,+\infty),\mathcal{D}'(\mathbb{R}^n))$. Indeed, first of all we can check that for any $t \in [0,+\infty)$, $\xi \to \exp(-f(\xi)t)$ is a continuous function and $\widetilde{u}_0$ belongs to $\mathcal{S}'(\mathbb{R}^n) \subset \mathcal{D}'(\mathbb{R}^n)$, thus $\widetilde{u}(t,\cdot)$ belongs to $\mathcal{D}'(\mathbb{R}^n)$. Besides, $t \to e^{-f(\xi)t}$ is a C^∞ function whose derivative in time are of the form $P(\xi)e^{-f(\xi)t}$ where $P(\xi)$ is a polynomial function. $\widetilde{u}$ is continuous in time and $\widetilde{u} \in C^0([0,+\infty),\mathcal{D}'(\mathbb{R}^n))$. Now we check that it is a weak solution to (19) with initial condition $\widetilde{u}_0$. Let $\phi \in C_c^\infty([0,+\infty)\times\mathbb{R}^n)$ the space of smooth functions with compact support, we have

$$\begin{aligned} &-\langle \widetilde{u}, \partial_t \phi\rangle + \sum_{|\alpha|\leq k} a_\alpha (i\xi)^\alpha \langle \widetilde{u},\phi\rangle + \langle \widetilde{u}_0,\phi\rangle \\ =&-\langle \widetilde{u}_0, \partial_t (e^{-\overline{f(\xi)}t}\phi)\rangle - \langle \widetilde{u}_0, \overline{f(\xi)} e^{-\overline{f(\xi)}t}\phi\rangle + \langle \widetilde{u}_0, e^{-\overline{f(\xi)}t}\overline{f(\xi)}\phi\rangle + \langle \widetilde{u}_0,\phi\rangle \\ =&0. \end{aligned} \tag{21}$$

Hence, u defined by (20) is indeed a weak solution of (19) in $C^0([0,+\infty),\mathcal{D}'(\mathbb{R}^n))$. Now, this does not answer our question as this only tells us that at time $t > 0$, $u(t,\cdot) \in \mathcal{D}'(\mathbb{R}^n)$ which is a less regular space than the space of tempered distribution $\mathcal{S}'(\mathbb{R}^n)$. In other words, at $t = 0$, $\widetilde{u} = \widetilde{u}_0$ has a higher regularity by being in $\mathcal{S}'(\mathbb{R}^n)$ and we would like to know if equation (19) preserves this regularity. This is more than a regularity issue as, if not, one cannot define a solution u as the inverse Fourier Transform of $\widetilde{u}$ because such function might not exist. Assume now that there exists a constant C such that

$$\forall \xi \in \mathbb{R}^n\ ,\ \widetilde{u}_0(\xi) = 0 \ \text{ or } \ \mathrm{Re}(f(\xi)) > C. \tag{22}$$

$$\forall\, \xi \in \mathbb{R}^n, \ \ \mathbf{1}_{\mathrm{supp}(\widetilde{u}_0)} e^{-f(\xi)t} \leq e^{-Ct}. \tag{23}$$

This implies that, for any $t > 0$, $\widetilde{u} \in \mathcal{S}'(\mathbb{R}^n)$. Besides, defining for any $p \in \mathbb{N}$,

$$\mathcal{N}_p(\phi) = \sum_{|\alpha|,|\beta|<p} \sup_{\xi\in\mathbb{R}^n} |\xi^\alpha \partial_\xi^\beta \phi(\xi)|, \tag{24}$$

then for $t_1, t_2 \in [0, T]$,

$$\mathcal{N}_p((e^{-f(\xi)t_1} - e^{-f(\xi)t_2})\phi) = \sum_{|\alpha|,|\beta|<p} \sup_{\xi\in\mathbb{R}^n} |\xi^\alpha P_\beta(\xi, \phi)|, \tag{25}$$

where $P_\beta(\xi, \phi)$ is polynomial with $f(\xi)$, $\phi(\xi)$, and their derivatives of order strictly smaller than p. Besides, each term of this polynomial tend to 0 when t_1 tends to t_2 on $\text{supp}(\widetilde{u_0})$, the set of frequency of u_0. Indeed, let β_1 be a multi-index, $k \in \mathbb{N}$, and $Q_i(\xi)$ be polynomials in ξ, where $i \in \{0, ..., k\}$.

$$\begin{aligned} &\left| \mathbf{1}_{\text{supp}(u_0)} \partial_\xi^{\beta_1} \phi(\xi) \left(\sum_{i=0}^{k} Q_i(\xi) t_1^i e^{-f(\xi)t_1} - Q_i(\xi) t_2^i e^{-f(\xi)t_2} \right) \right| \\ &\leq \sum_{i=0}^{k} \max_{\text{supp}(\widetilde{u}_0)} \left| t_1^i e^{-f(\xi)t_1} - t_2^i e^{-f(\xi)t_2} \right| \max_{\xi\in\mathbb{R}^n} \left| \partial_\xi^{\beta_1} \phi(\xi) Q_i(\xi, t) \right|. \end{aligned} \tag{26}$$

From (22), the time-dependant terms in the right-hand sides converge to 0 when t_1 tends to t_2. This implies that $u \in C^0([0, T], \mathcal{S}'(\mathbb{R}^n))$. Finally let us show the property of the behavior at infinity. Assume that $C > 0$, one has, for any $\phi \in S(\mathbb{R}^n)$

$$\langle \widetilde{u}(t, \cdot), \phi \rangle = \langle \widetilde{u}_0, \mathbf{1}_{\text{supp}(\widetilde{u}_0)} e^{-\overline{f(\xi)}t} \phi \rangle. \tag{27}$$

Let us set $g(\xi) = e^{-\overline{f(\xi)}t}\phi(\xi)$, one has for two multi-index α and β

$$|\xi^\alpha \partial_\xi^\beta g(\xi)| \leq |\xi^\alpha Q(\xi) e^{-f(\xi)t}|, \tag{28}$$

where Q is a sum of polynomials, each multiplied by $\phi(\xi)$ or one of its derivatives. Thus $\xi^\alpha Q(\xi)$ belongs to $\mathcal{S}(\mathbb{R}^n)$ and therefore, from assumption (22),

$$|\xi^\alpha \partial_\xi^\beta g(\xi)| \mathbf{1}_{\text{supp}(u_0)} \leq \max_{\xi\in\mathbb{R}^n} |\xi^\alpha Q(\xi)| e^{-Ct}, \tag{29}$$

which goes to 0 when $t \to +\infty$. This imply that $\widetilde{u}(t, \cdot) \to 0$ in $\mathcal{S}'(\mathbb{R}^n)$ when $t \to +\infty$, and hence $u(t, \cdot) \to 0$. This ends the proof of Proposition 3.1.

Let us note that one could try to find solutions with lower regularity, where u is a distribution of $\mathcal{D}'(\mathbb{R}_+ \times \mathbb{R}^n)$, and satisfies the equation

$$\partial_t u + \sum_{|\alpha|\leq k} a_\alpha \partial_x^\alpha u = \delta_{t=0} u_0 \text{ on } \mathbb{R}_+ \times \mathbb{R}^n. \tag{30}$$

This could be done using for instance Malgrange-Erhenpreis theorem, however, studying the behavior at $t \to +\infty$ may be harder mathematically, hence this approach was not considered in this paper.

E Size of the problem space

Lample and Charton (2020) provide the following formula to calculate the number of functions with m operators:

$$\begin{aligned} E_0 &= L \\ E_1 &= (q_1 + q_2 L)L \\ (m+1)E_m &= (q_1 + 2q_2 L)(2m-1)E_{m-1} - q_1(m-2)E_{m-2} \end{aligned}$$

Where L is the number of possible leaves (integers or variables), and q_1 and q_2 the number of unary and binary operators. In the stability and controllability problems, we have $q_1 = 9$, $q_2 = 4$ and $L = 20 + q$, with q the number of variables.

Replacing, we have, for a function with q variables and m operators

$$E_0(q) = 20 + q$$
$$E_1(q) = (89 + 4q)(20 + q)$$
$$(m + 1)E_m(q) = (169 + 8q)(2m - 1)E_{m-1} - 4(m - 2)E_{m-2}$$

In the stability problem, we sampled systems of n functions, with n variables, n from 2 to 6. Functions have between 3 and $2n + 2$ operators. The number of possible systems is

$$PS_{st} = \sum_{n=2}^{6} \left(\sum_{m=3}^{2n+2} E_m(n) \right)^n > E_{14}(6)^6 \approx 3.10^{212}$$

(since $E_m(n)$ increases exponentially with m and n, the dominant factor in the sum is the term with largest m and n)

In the autonomous controllability problem, we generated systems with n functions (n between 3 and 6), and $n + p$ variables (p between 1 and $n/2$). Functions had between $n + p$ and $2n + 2p + 2$ operators. The number of systems is

$$PS_{aut} = \sum_{n=3}^{6} \left(\sum_{p=1}^{n/2} \sum_{m=n+p}^{2(n+p+1)} E_m(n + p) \right)^n > E_{20}(9)^6 \approx 4.10^{310}$$

For the non-autonomous case, the number of variables in $n + p + 1$, n is between 2 and 3 and $p = 1$, therefore

$$PS_{naut} = \sum_{n=2}^{3} \left(\sum_{m=n+1}^{2(n+2)} E_m(n + 2) \right)^n > E_{10}(5)^3 \approx 5.10^{74}$$

Because expressions with undefinite or degenerate jacobians are skipped, the actual problem space size will be smaller by several orders of magnitude. Yet, problem space remains large enough for overfitting by memorizing problems and solutions to be impossible.

F Model architecture

The networks used in this paper are very close to the one described in Vaswani et al. (2017). They use an encoder/decoder architecture. The encoder stack contains 6 transformer layers, each with a 8 head self-attention layer, a normalization layer, and a one layer feed forward network with 2048 hidden units. Inputs is fed through trainable embedding and positional embedding, and the encoder stack learns a representation of dimension 512. The decoder contains 6 transformer layers, each with a (8-head) self-attention layer, a cross attention (pointing to the encoder output) layer, normalization and feed forward linear layer. Representation dimension is the same as the encoder (512). The final output is sent to a linear layer that decodes the results.

The training loss is the cross entropy between the model predicted output and actual result from the dataset. During training, we use the Adam optimizer, with a learning rate of 0.0001 and scheduling (as in Vaswani et al. (2017)). Mini-batch size varies from one problem to the other, typically between 32 and 128 examples.

During training, we use 8 GPU. The model is distributed across GPUs, so that all of them have access to the same shared copy of the model. At each iteration, every GPU processes an independently generated batch, and the optimizer updated the model weights using the gradients accumulated by all GPU. Overall, this is equivalent to training on a single GPU, but with 8 times larger batches.

G Algorithmic complexity

Let n be the system degree, p the number of variables and q the average length (in tokens) of functions in the system. In all problems considered here, we have $p = O(n)$. Differentiating

or evaluating an expression with q tokens is O(q), and calculating the Jacobian of our system is $O(npq)$, i.e. $O(n^2q)$.

In the stability experiment, calculating the eigenvalues of the Jacobian will be $O(n^3)$ in most practical situations. In the autonomous controllability experiments, construction of the $n \times np$ Kalman matrix is $O(n^3p)$, and computing its rank, via singular value decomposition or any equivalent algorithm, will be $O(n^3p)$ as well. The same complexity arise for feedback matrix computations (multiplication, exponentiation and inversion are all $O(n^3)$ for a square n matrix). As a result, for controllability, complexity is $O(n^4)$. Overall, the classical algorithms have a complexity of $O(n^2q)$ for Jacobian calculation, and $O(n^3)$ (stability) and $O(n^4)$ (controllability) for the problem specific computations.

Current transformer architectures are quadratic in the length of the sequence, in our case nq, so a transformer will be $O(n^2q^2)$ (in speed and memory usage). Therefore, the final comparison will depend on how q, the average length of equations, varies with n, the number of parameters. If $q = O(1)$ or $O(log(n))$, transformers have a large advantage over classical methods. This means sparse Jacobians, a condition often met in practice. For controllability, the advantage remains if $q = O(n^{1/2})$, and the two methods are asymptotically equivalent if $q = O(n)$.

However, current research is working on improving transformer complexity to log-linear or linear. If this happened (and there seem to be no theoretical reason preventing it), transformers would have lower asymptotic complexity in all cases.

H Learning curves

Although all generated datasets included more than 50 million examples, most models were trained on less. The following curves shows how performance increases with the number of training examples, for the end to end stability problem (i.e. predicting whether systems of degree 2 to 5 are stable). There are twelve curves corresponding to as many experiments over shuffled versions of the dataset (i.e. different experiments used different parts of the dataset).

Overall, less than 10 million examples are needed to achieve close to optimal accuracy. Learning curves from different experiments are close, which proves the stability of the learning process.

I Out-of-distribution generalization

In all our experiments, trained models are tested on held-out samples generated with the same procedure as the training data, and our results prove that the model can generalize out of the training data. However, training and test data come from the same statistical distribution (iid). This would not happen in practical cases: problems would come from some unknown distribution over problem space. Therefore, it is interesting to investigate how the model performs when the test set follows a different statistical distribution. This provides insight about how learned properties generalize, and may indicate specific cases over which the model struggles.

To this purpose, we modified the data generator to produce new test datasets for end to end stability prediction (section 5.1). Four modifications were considered:

1. **Unary operators:** varying the distribution of operators in the system. In the training data, unary operators are selected at random from a set of nine, three trigonometric functions, three inverse trigonometric functions, logarithm and exponential, and square root (the four basic operations are always present). In this set of experiments, we generated four test sets, without trigonometric functions, without logs and exponentials, only with square roots, and with a different balance of operators (mostly square roots).
2. **Variables and integers:** varying the distribution of variables in the system. In the training data, 30% of the leaves are numbers, the rest variables. We changed

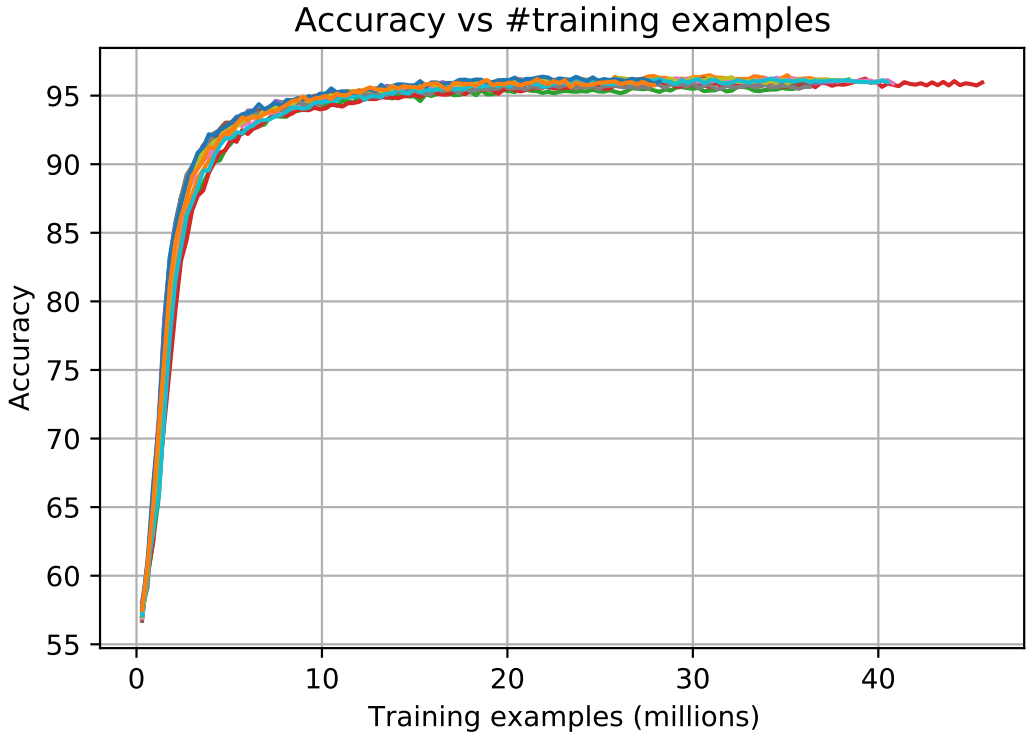

Figure 1: **End to end stability accuracy vs number of training examples.** 12 models, trained over shuffled versions of the same dataset.

this probability to 0.1, 0.5 and 0.7. This has no impact on expression length, but higher probabilities make the Jacobians more sparse.

3. **Expression lengths:** making expressions longer than in the train set. In the training data, for a system of n equations, we generate functions with 3 to $2n+3$ operators. In this experiments, we tried functions between $n+3$ and $3n+3$ and $2n+3$ and $4n+3$. This means that the test sequences are, on average, much longer that those seen at training, a known weakness of sequence to sequence models.

4. **Larger degree:** our models were trained on systems with 2 to 5 equations, we tried to test it on systems with 6 equations. Again, this usually proves difficult for transformers.

Note that the two first sets of experiments feature out-of-distribution tests, exploring different distributions over the same problem space as the training data. The two last sets, on the other hand, explore a different problem space, featuring longer sequences.

Table 9 presents the results of these experiments. Changing the distribution of operators, variables and integers has little impact on accuracy, up to two limiting cases. First, over systems of degree five (the largest in our set, and more difficult for the transformers) change in operator distribution has a small adverse impact on performance (but not change in variable distribution). Second, which the proportion of integers become very large, and therefore Jacobians become very sparse, the degree of the systems has less impact on performance. But overall results remain over 95%, and the model proves to be very resistant to changes in distribution over the same problem space.

Over systems with longer expressions, overall accuracy tends to decreases. Yet, systems of two or three equations are not affected by a doubling of the number of operators (and sequence length), compared to the training data. Most of the loss in performance concentrates on larger degrees, which suggests that it results from the fact that the transformer is presented

at test time with much longer sequences that what it saw at training. In any case, all results but one are well above the fastText baseline (60.5%).

When tested on systems with six equations, the trained model predicts stability in 78.7% of cases. This is a very interesting result, where the model is extrapolating out of the problem space (i.e. no system of six equations have been seen during training) with an accuracy well above chance level, and the fastText baseline.

Table 9: **End to end stability: generalization over different test sets.**

	Overall	Degree 2	Degree 3	Degree 4	Degree 5
Baseline: training distribution	96.4	98.4	97.3	95.9	94.1
Unary operators: no trigs	95.7	98.8	97.3	95.5	91.2
Unary operators: no logs	95.3	98.2	97.1	95.2	90.8
Unary operators: no logs and trigs	95.7	98.8	97.7	95.2	91.0
Unary operators: less logs and trigs	95.9	98.8	96.8	95.0	93.1
Variables and integers: 10% integers	96.1	98.6	97.3	94.7	93.8
Variables and integers: 50% integers	95.6	97.8	96.7	94.3	93.1
Variables and integers: 70% integers	95.7	95.7	95.9	95.7	95.5
Expression lengths: n+3 to 3n+3	89.5	96.5	92.6	90.0	77.9
Expression lengths: 2n+3 to 4n+3	79.3	93.3	88.3	73.4	58.2
System degree: degree 6	78.7				