Turnpike Control and Deep Learning

Borjan Geshkovski & Enrique Zuazua

FAU - AvH / CCM - Deusto, Bilbao / UAM-Madrid
enrique.zuazua@fau.de
borjan.geshkovski@uam.es
paginaspersonales.deusto.es/enrique.zuazua

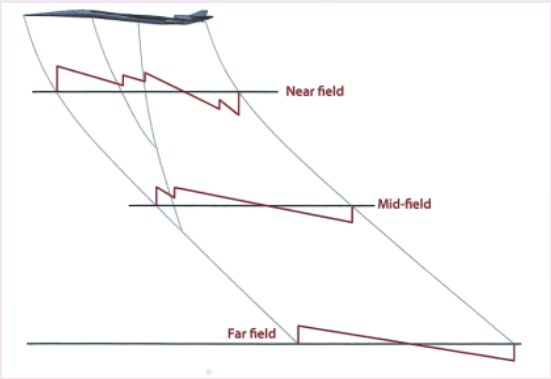
Second Symposium on Machine Learning and Dynamical Systems
August 2020
Fields Institute, Toronto

Outline

Sonic boom

Francisco Palacios, Boeing, Long Beach, California, Project Manager and Aerodynamics Engineer

- Goal: the development of supersonic aircrafts, sufficiently quiet to be allowed to fly supersonically over land.
- The pressure signature created by the aircraft must be such that, when reaching ground, (a) it can barely be perceived by humans, and (b) it results in admissible disturbances to man-made structures.
- This leads to an inverse design or control problem in long time horizons.



Juan J. Alonso and Michael R. Colonno, Multidisciplinary Optimization with Applications to Sonic-Boom Minimization, Annu. Rev. Fluid Mech. 2012, 44:505 – 526.

Many other challenging problems of high societal impact raise similar issues: climate change, sustainable growth, chronically deseases, design of long lasting devices and infrastructures...

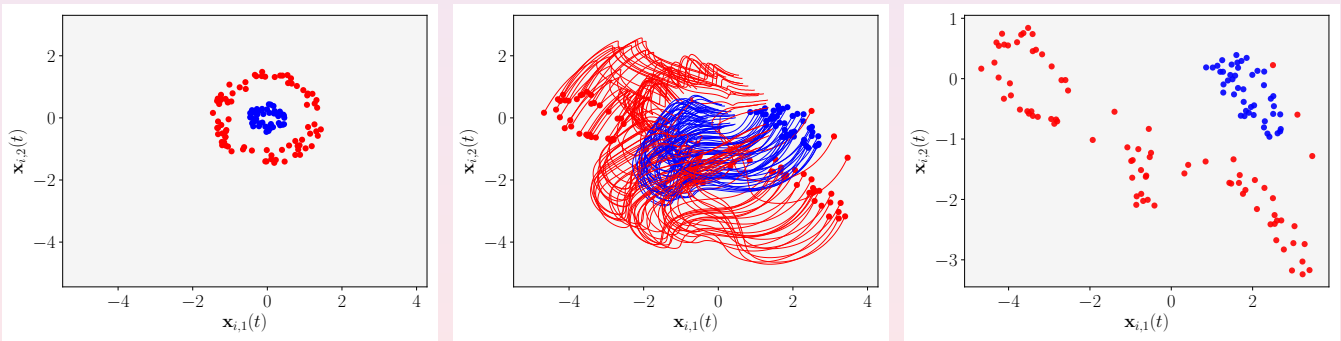
Deep learning

- Residual neural networks (ResNets) (He et al. '15) have become the building blocks of modern deep learning;
- Recent work (E '17, Haber & Ruthotto '17, Chen et al. '18) has reinterpreted ResNets as continuous-time controlled nonlinear dynamical systems:

$$\boxed{\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), u(t)) \qquad t \in (0, T)}$$

where $T > 0$ plays the role of the number of layers in the discrete-time setting, f has very specific form (sigmoid);

- Controls $u = u(t)$, corresponding to the free parameters of the ResNet, found by minimizing an appropriate nonnegative cost function $\mathcal{J}_T$ (*training*);



- What happens when $T \to \infty$, i.e. in the deep, high number of layers regime?[1]

[1] Suggested by our FAU colleague Daniel Tenbrinck.

Origins

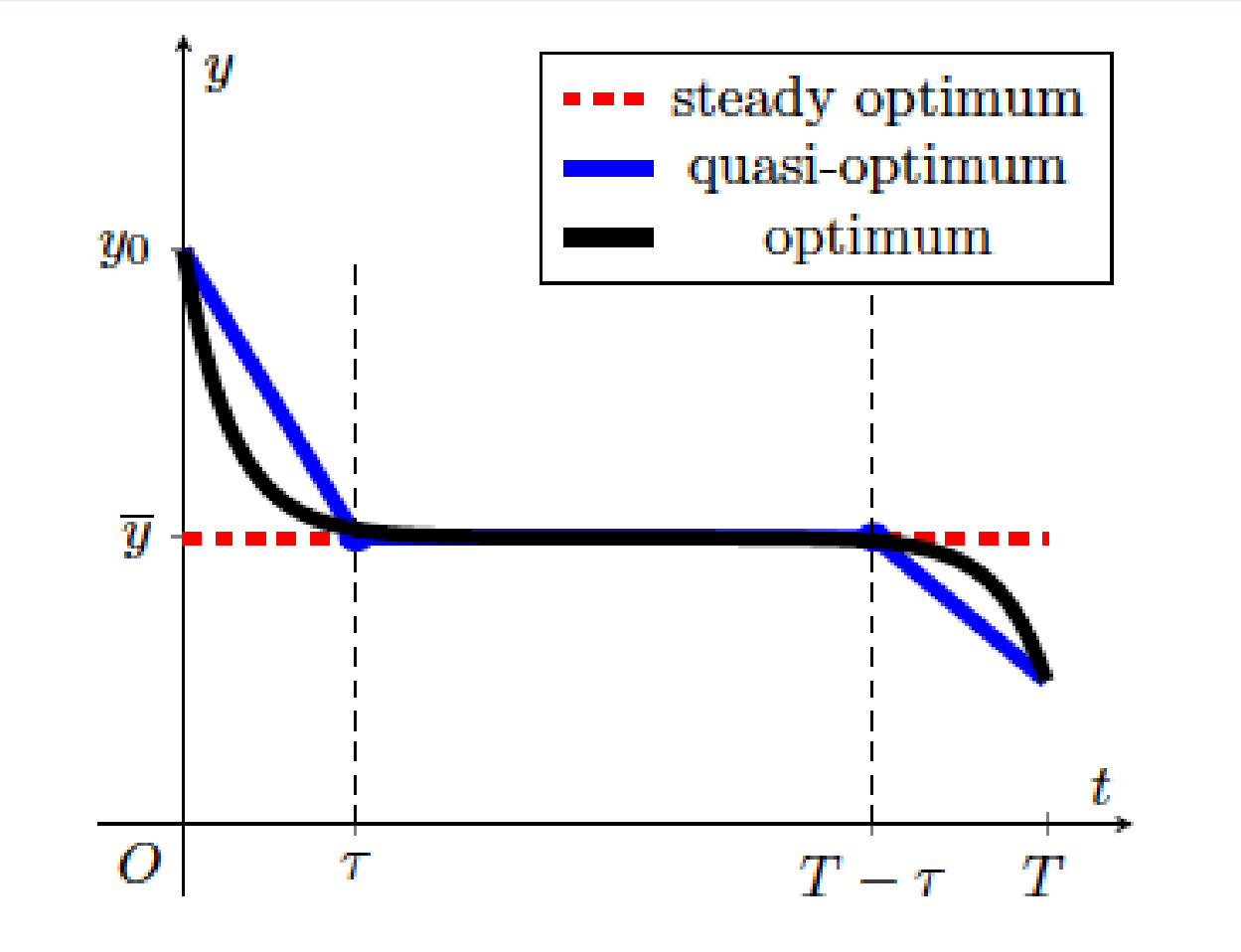
Although the idea goes back to John von Neumann in 1945, Lionel W. McKenzie traces the term to Robert Dorfman, Paul Samuelson, and Robert Solow's "Linear Programming and Economics Analysis" in 1958, referring to an American English word for a Highway:

> *... There is a fastest route between any two points; and if the origin and destination are close together and far from the turnpike, the best route may not touch the turnpike. But if the origin and destination are far enough apart, it will always pay to get on to the turnpike and cover distance at the best rate of travel, even if this means adding a little mileage at either end.*

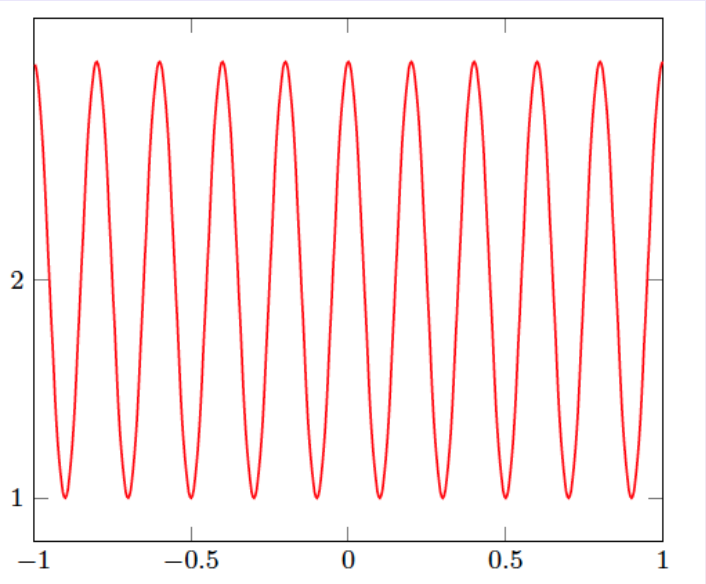
Substantiation and preliminary conclusion

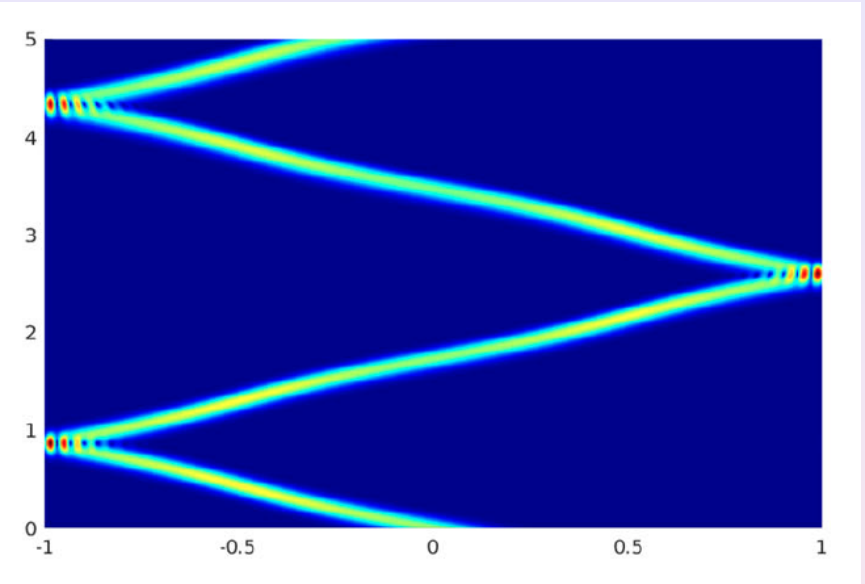
We implement turnpike (or nearby) strategies most often. And it is indeed a good idea to do it!

But this requires that the system under consideration to be controllable/stabilisable.



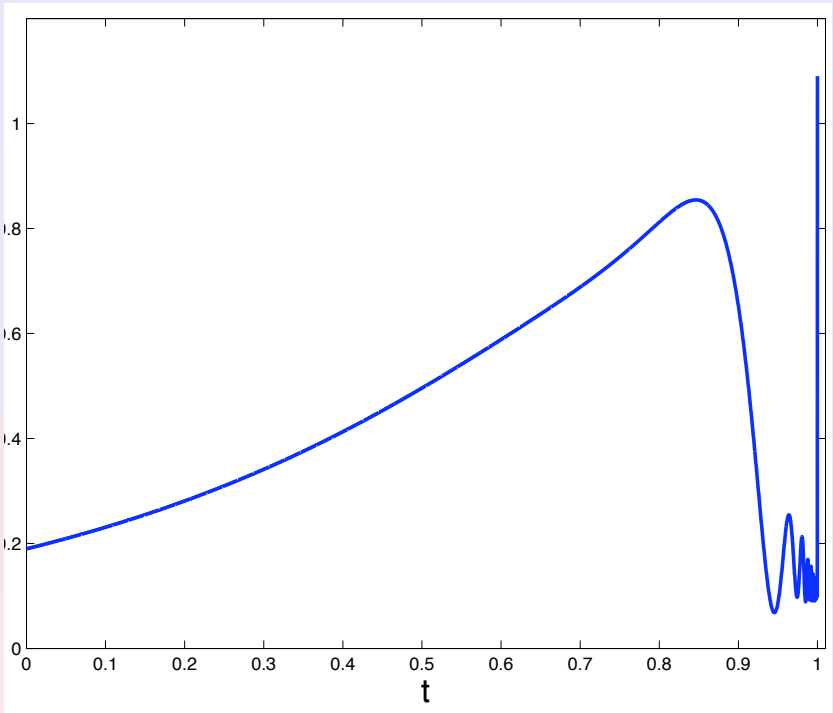
Wave propagation: Why do not we see the turnpike?

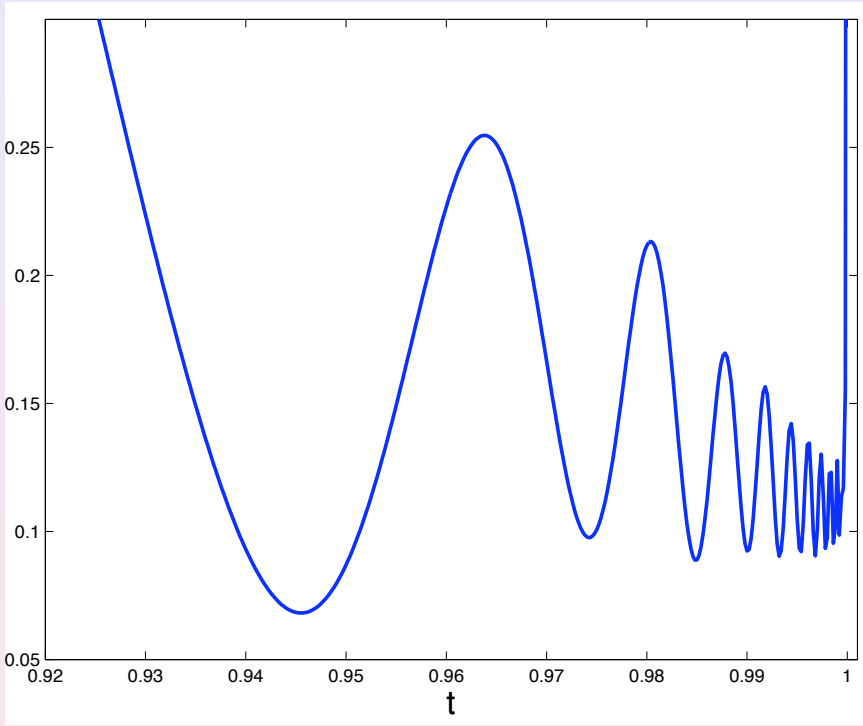




- Typical controls for the wave equation exhibit an oscillatory behaviour, and this independently of the length of the control time-horizon.
- Nobody would be surprised about this fact that seems to be intrinsically linked to the oscillatory (even periodic in some particular cases) nature of the wave equation solutions.
- Waves propagate with finite speed and it is natural to control them through anti-waves when they reach the actuator location.

Heat and diffusion processes: Why do not we see the turnpike either?





Typical controls for the heat equation exhibit **unexpected** oscillatory and concentration effects. This was observed by R. Glowinski and J. L. Lions in the 80's in their works in the numerical analysis of controllability problems for heat and wave equations.

Why? Lazy controls?

Optimal controls are normally characterised as boundary traces of solutions of the **adjoint problem** through the optimality system or the Pontryagin Maximum Principle, and solutions of the adjoint system of the heat equation

$$-p_t - \Delta p = 0$$

look precisely this way.

Large and oscillatory near $t = T$ they decay and get smoother when t gets down to $t = 0$. And this is independent of the time control horizon $[0, T]$. The same occurs to wave-like equations

where controls are given by the solutions of the adjoint system

$$p_{tt} - \Delta p = 0$$

that exhibit endless oscillations.

First conclusion:
Typical control problems for wave and heat equations do not seem to exhibit the turnpike property.
Note however that these are the controls of L^2-minimal norm. There are many other possibilities for successful control strategies.

The control problem for diffusion : A closer look

Let $n \geq 1$ and $T > 0$, Ω be a simply connected, bounded domain of $\mathbb{R}^n$ with smooth boundary Γ, $Q = (0, T) \times \Omega$ and $\Sigma = (0, T) \times \Gamma$:

$$\begin{cases} y_t - \Delta y = f 1_\omega & \text{in} \quad Q \\ y = 0 & \text{on} \quad \Sigma \\ y(x, 0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{1}$$

1_ω = the characteristic function of ω of Ω where the control is active.

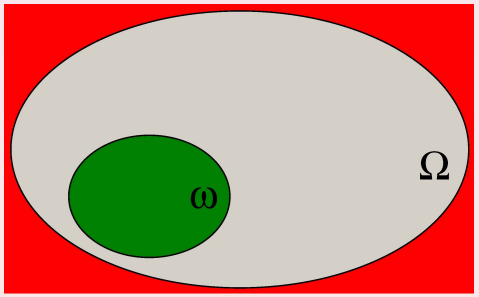
We know that $y^0 \in L^2(\Omega)$ and $f \in L^2(Q)$ so that (9) admits a unique solution

$$y \in C\left([0, T]; L^2(\Omega)\right) \cap L^2\left(0, T; H_0^1(\Omega)\right).$$

$$y = y(x, t) = \textit{solution} = \textit{state}, \; f = f(x, t) = \textit{control}$$

Goal: Drive the dynamics to equilibrium by means of a suitable choice of the control

$$y(\cdot, T) \equiv y^*(x).$$



We address this problem fro a classical optimal control / least square approach:

$$min\ \frac{1}{2}\left[\int_0^T \int_\omega |f|^2 dxdt + \int_\Omega |y(x,T) - y^*(x)|^2 dx\right].$$

According to Pontryagin's Maximum Principle the Optimality System (OS) reads

$$y_t - \Delta y = \varphi 1_\omega \text{ in } Q$$

$$-\varphi_t - \Delta\varphi = 0 \text{ in } Q$$

$$y = 0 \text{ on } \Sigma$$

$$y(x,0) = y^0(x) \text{ in } \Omega$$

$$\varphi(x,T) = y(x,T) - y^*(x) \text{ in } \Omega$$

$$\varphi = 0 \text{ on } \Sigma.$$

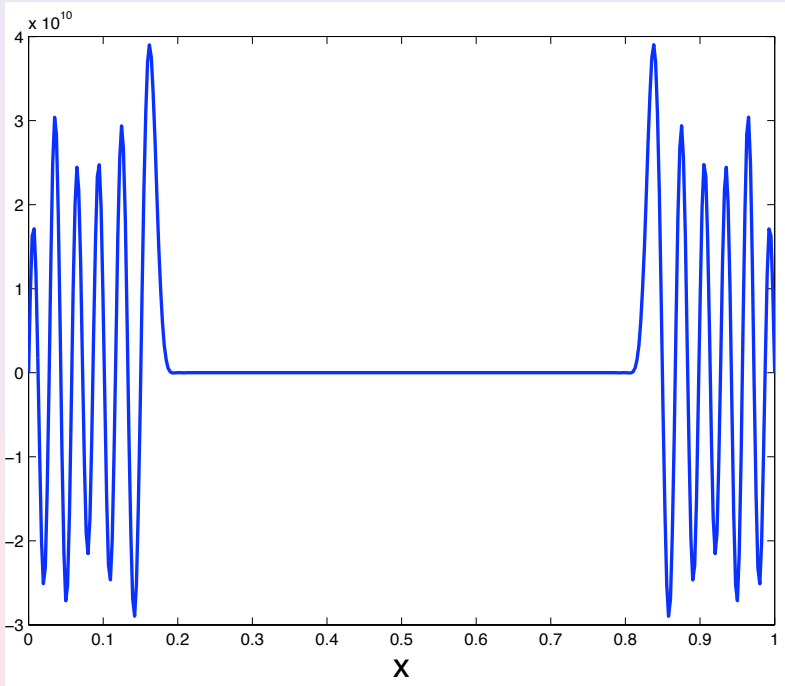
And the optimal control is:

$$f(x,t) = \varphi(x,t) \quad \text{in}\, \omega \times (0,T).$$

The minimizer φ^T saturates the regularity properties required to assure the well-posedness of the functional:

$$\mathcal{H} = \{\varphi^T : \varphi(x,0) \in L^2(\Omega)\}$$

This is a huge space, allowing an exponential increase of Fourier coefficients at high frequencies. And, because of this, we observe the tendency of the control to concentrate all the action in the final time instant $t = T$, incompatible with turnpike effects[2]



Tychonoff's monster (1935)

Théorèmes d'unicité pour l'équation de la chaleur

A. Tychonoff (Moscou)

[2] A. Münch & E. Z., Inverse Problems, 2010

Remedy: Better balanced controls

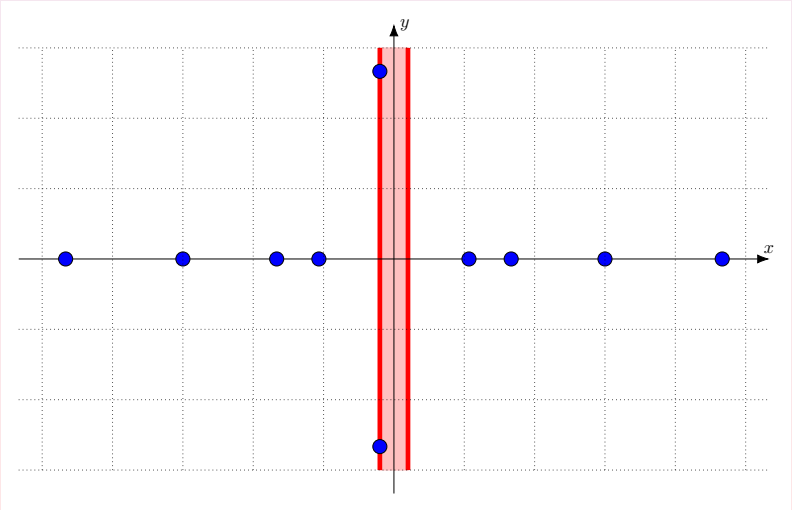
Let us now consider the control f minimising a compromise between the norm of the state and the control among the class of admissible controls:

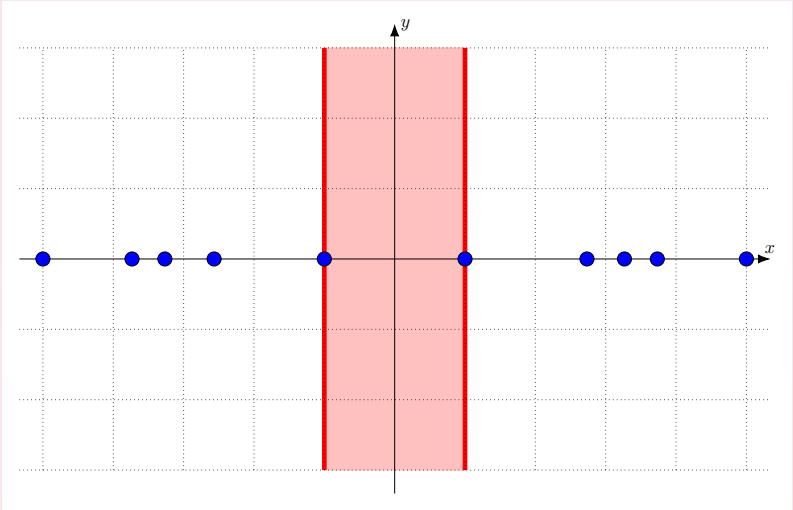
$$min\ \frac{1}{2}\left[\int_0^T\int_\Omega |y|^2 dxdt + \int_0^T\int_\omega |f|^2 dxdt + \int_\Omega |y(x,T)-y^*(x)|^2\right].$$

Then the Optimality System reads

$$y_t - \Delta y = -\varphi 1_\omega \text{ in } Q$$
$$-\varphi_t - \Delta\varphi = y \text{ in } Q$$
$$y = \varphi = 0 \text{ on } \Sigma$$
$$y(x,0) = y^0(x) \text{ in } \Omega$$
$$\varphi(x,T) = y(x,T) - y^*(x) \text{ in } \Omega$$

We now observe a coupling between φ and y on the adjoint state equation!





New Optimality System Dynamics

What is the dynamic behaviour of solutions of the new fully coupled OS?
For the sake of simplicity, assume $\omega = \Omega$.

The dynamical system now reads

$$y_t - \Delta y = -\varphi$$
$$\varphi_t + \Delta \varphi = -y$$

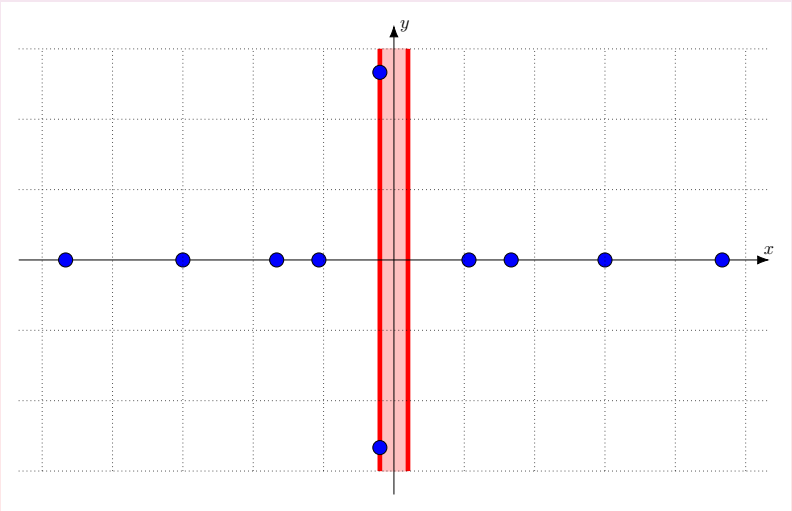
This is a forward-backward parabolic system.

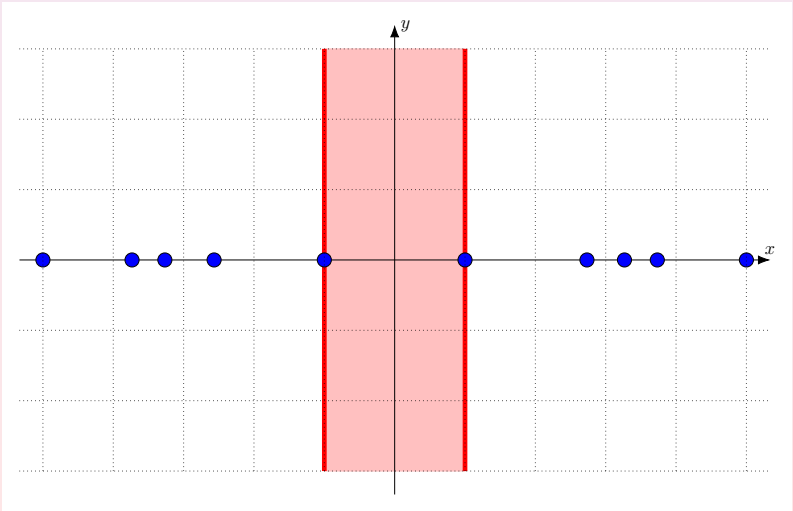
A spectral decomposition exhibits the characteristic values

$$\mu_j^{\pm} = \pm\sqrt{1+\lambda_j^2}$$

where $(\lambda_j)_{j\geq 1}$ are the (positive) eigenvalues of $-\Delta$.

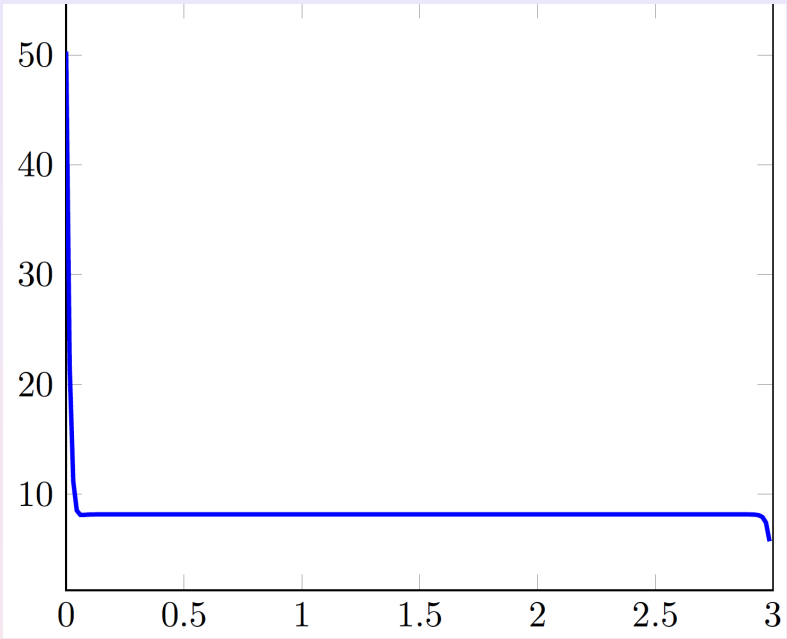
Thus, the system is the superposition of growing + diminishing real exponentials.





The turnpike property for the heat equation

This new dynamic behaviour, combining exponentially stable and unstable branches, is compatible with the turnpike behavior.
Controls and trajectories exhibit the expected dynamics:



The turnpike behaviour is ensured by modifying the optimality criterion for the choice of the control, to weight both state and control and provided $T \gg 1$.

The same occurs for wave propagation:
M. Gugat, E. Trélat, E. Zuazua, Systems and Control Letters, 90 (2016), 61-70.

[Controllability] + [Coercive in state + control cost] → Turnpike

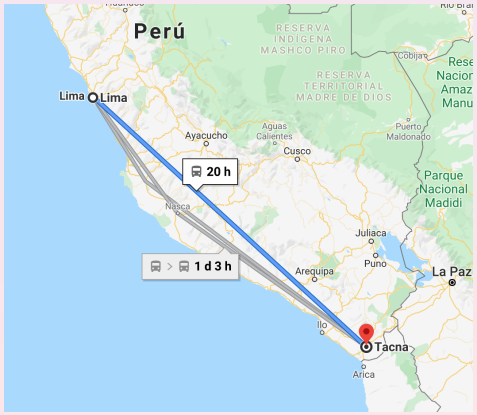
Mainly motivated by applications to economic models and game theory there was a literature concerned with this kind of stationary behavior in the transient time for long horizon control problems. In that context, such type of result goes under the name of *turnpike theory* which was mostly investigated in the finite dimensional case.

A. J. Zaslavski, *Turnpike properties in the calculus of variations and optimal control*. Nonconvex Optimization and its Applications, 80. Springer, New York, 2006.

L. Grüne, Economic receding horizon control without terminal constraints Automatica, 49, 725-734, 2013

But our main motivation originated from the optimal shape design in aeronautics and other PDE problems.

In recent years a number of model cases have been well understood in the infinite dimensional PDE context. But there is still a long way to go...



Dakar 2019

Linear theory. Joint work with A. Porretta, SIAM J. Cont. Optim., 2013.

The same methods apply in the inifinite-dimensional context, covering in particular linear heat and wave equations

Consider the finite dimensional dynamics

$$\begin{cases} x_t + Ax = Bu \\ x(0) = x_0 \in \mathbb{R}^N \end{cases} \tag{2}$$

where $A \in M(N,N)$, $B \in M(N,M)$, with control $u \in L^2(0,T;\mathbb{R}^M)$.
Given a matrix $C \in M(N,N)$, and some $x^* \in \mathbb{R}^N$, consider the optimal control problem

$$\min_u \; J^T(u) = \frac{1}{2}\int_0^T (|u(t)|^2 + |C(x(t) - x^*)|^2)dt\,.$$

There exists a unique optimal control $u(t)$ in $L^2(0,T;\mathbb{R}^M)$, characterized by the optimality condition

$$u = -B^*p\,, \qquad \begin{cases} x_t + Ax = -BB^*p \\ x(0) = x_0 \end{cases}\,, \qquad \begin{cases} -p_t + A^*p = C^*C(x - x^*) \\ p(T) = 0 \end{cases} \tag{3}$$

The steady state control problem

The same problem can be formulated for the steady-state model

$$Ax = Bu.$$

Then there exists a unique minimum $\bar{u}$, and a unique optimal state $\bar{x}$, of the stationary control problem

$$\min_u \; J_s(u) = \frac{1}{2}(|u|^2 + |C(x - x^*)|^2) \tag{4}$$

which is nothing but a constrained minimization in $\mathbb{R}^N$.
The optimal control $\bar{u}$ and state $\bar{x}$ satisfy

$$\bar{u} = -B^* \bar{p}\,, \qquad A\bar{x} = B\bar{u}\,, \quad \text{and} \quad A^* \bar{p} = C^* C(\bar{x} - x^*)\,.$$

We assume that

$$(A, B) \text{ is controllable,} \tag{5}$$

or, equivalently, that the matrices A, B satisfy the Kalman rank condition

$$Rank \left[B \; AB \; A^2B \ldots A^{N-1}B \right] = N \; . \tag{6}$$

Concerning the cost functional, we assume that the matrix C is such that (void assumption when $C = Id$)

$$(A, C) \text{ is observable} \tag{7}$$

which means that the following algebraic condition holds:

$$Rank \left[C \; CA \; CA^2 \ldots CA^{N-1} \right] = N \; . \tag{8}$$

$$x_t + Ax = Bu$$

$$J^T(u) = \frac{1}{2} \int_0^T (|u(t)|^2 + |C(x(t) - x^*)|^2) dt$$

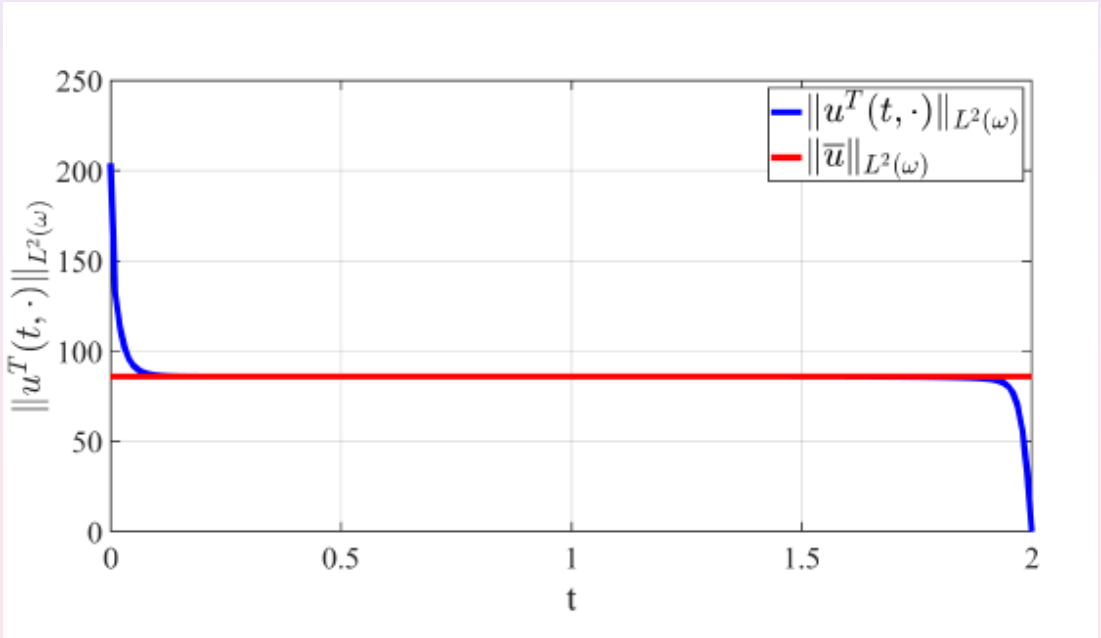
$$\begin{cases} x_t + Ax = Bu \\ -p_t + A^* p = C^* Cx \end{cases}$$

Under the above controllability and observability assumptions, we have the following result.

Theorem

For some $\gamma > 0$ for $T > 0$ large enough we have

$$||x^T(t) - \bar{x}|| + ||u^T(t) - \bar{u}|| \leq C[\exp(-\mu t) + \exp(-\mu(T-t))].$$



Note the presence of the two boundary layers at $t = 0$ and $t = T$ and that the state and control x^T and u^T are defined in $[0, T]$, that varies as $T \to \infty$.

Proofs

Proof # 1: Dissipativity

$$\frac{d}{dt}[(x-\bar{x})(p-\bar{p})] = -\left[B^*(p-\bar{p})|^2 + |C(x-\bar{x})|^2\right]$$

That is the starting point of a turnpike proof. Note however that it is much trickier than the classical Lyapunov stability: Two boundary layers at $t = 0$ and $t = T$, moving time-horizon $[0, T]$...

Proof #2 : Riccati

- First consider the infinite horizon Linear Quadratic Regulator (LQR) problem for $0 \leq t < +\infty$ with null target $x^* \equiv 0$.
- Employ Riccati theory to describe the optimal trajectory in a feedback manner.
- Take the cut-off of this optimal Riccati trajectory from $[0, \infty)$ into $[0, T]$.
- Correct the boundary layer at $t = T$ to match the terminal conditions of the Optimality System in $[0, T]$.

Proof # 3: Singular perturbations Implement the change of variables $t \to sT$ so that the time variable $t \in [0, T]$ becomes $s \in [0, 1]$. Then the control problem

$$x_t + Ax = Bu, \quad t \in [0, T]$$

becomes

$$\frac{1}{T}x_s + Ax = Bu, \quad t \in [0, 1]$$

As $T \to \infty$ this indicates the trend towards steady state control.

Hyperbolicity

It is a direct consequence of the hyperbolicity of the underlying dynamics, whose steady state solutions are characterised by the system

$$A\bar{x} + BB^*\bar{p} = 0$$

$$-A^*\bar{p} + C^*C\bar{x} = C^*Cx^*$$

generated by the operator matrix

$$\tilde{A} = \begin{pmatrix} A & BB^* \\ C^*C & -A^* \end{pmatrix}$$

Note however that the hyperbolicity of this matrix operator needs of controllability/observability conditions.

A major technical difficulty for nonlinear problems

Consider now the semilinear heat equation:

$$\begin{cases} y_t - \Delta y + y^3 = f1_\omega & \text{in} \quad Q \\ y = 0 & \text{on} \quad \Sigma \\ y(x,0) = y^0(x) & \text{in} \quad \Omega \end{cases} \tag{9}$$

$$\min_f \left[\frac{1}{2} \int_0^T \int_\Omega |y - y_d|^2 dxdt + \int_0^T \int_\omega f^2 dxdt \right].$$

The optimality system reads:

$$y_t - \Delta y + y^3 = -\varphi 1_\omega \text{ in } Q$$

$$y = 0 \text{ on } \Sigma$$

$$y(x,0) = y^0(x) \text{ in } \Omega$$

$$-\varphi_t - \Delta\varphi + 3y^2\varphi = y - y_d \text{ in } Q$$

$$\varphi = 0 \text{ on } \Sigma$$

$$\varphi(x,T) = 0 \text{ in } \Omega.$$

Linearisation of the OS

And the linearised optimality system, around the optimal steady solution $(\bar{y}, \bar{\varphi})$ is as follows:

$$z_t - \Delta z + 3(\bar{y})^2 z = -\psi 1_\omega \text{ in } Q$$

$$z = 0 \text{ on } \Sigma$$

$$z(x,0) = 0 \text{ in } \Omega$$

$$-\psi_t - \Delta\psi + 3(\bar{y})^2\psi = (1 - 6\bar{y}\bar{\varphi})z \text{ in } Q$$

$$\psi = 0 \text{ on } \Sigma$$

$$\psi(x,T) = 0 \text{ in } \Omega.$$

This is the optimality system for a LQ control problem of the model

$$z_t - \Delta z + 3(\bar{y})^2 z = f 1_\omega$$

and the cost

$$\min_f \left[\frac{1}{2}\int_0^T \int_\Omega |z|^2 dxdt + \int_0^T \int_\omega \rho(x) f^2 dxdt\right]$$

$$\rho(x) = 1 - 6\bar{y}(x)\bar{\varphi}(x).$$

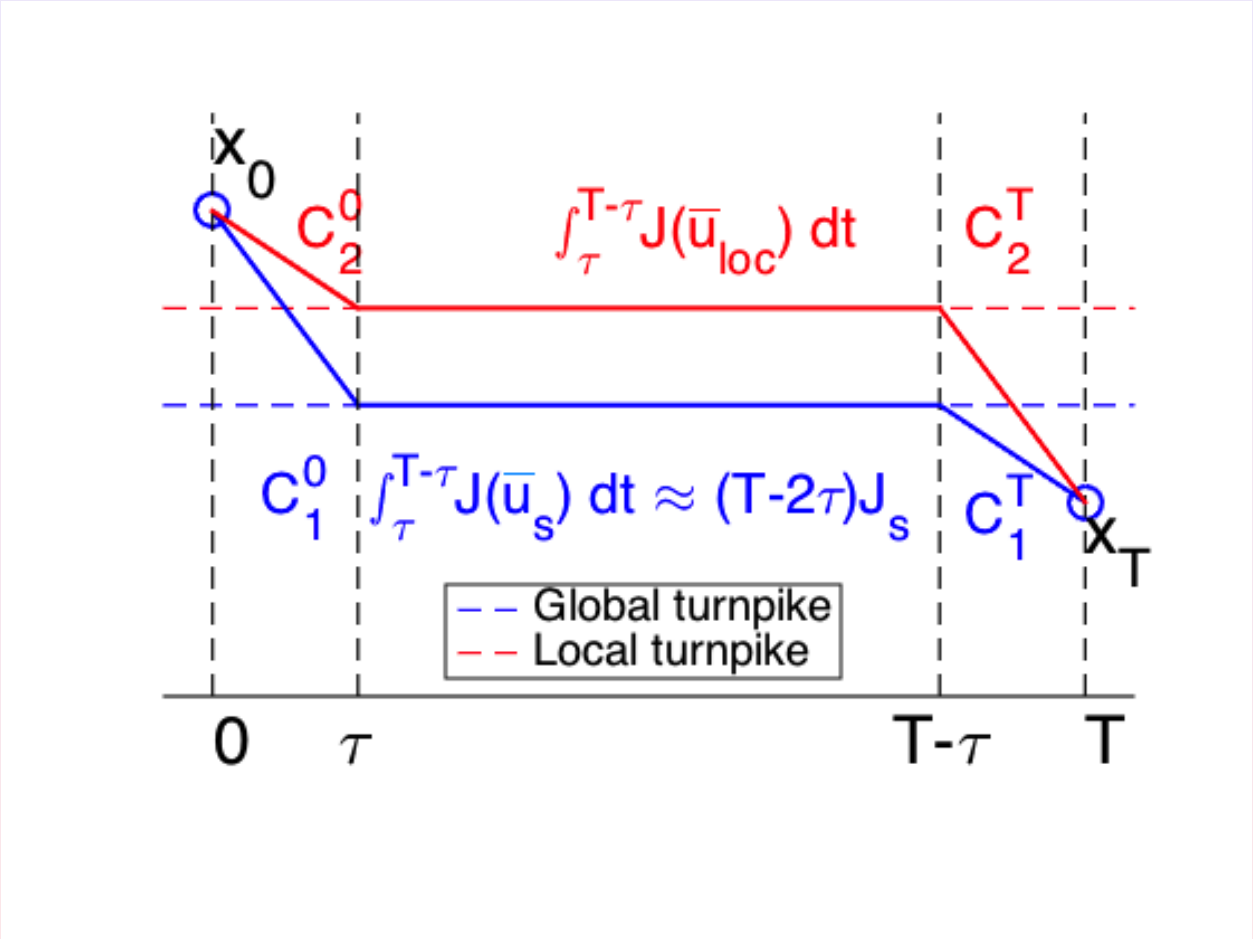
And the turnpike property holds as soon as

$$\rho(x) \geq \delta > 0.$$

This holds if $\bar{y}$ and φ are small enough, and this requires the **smallness of the target**.

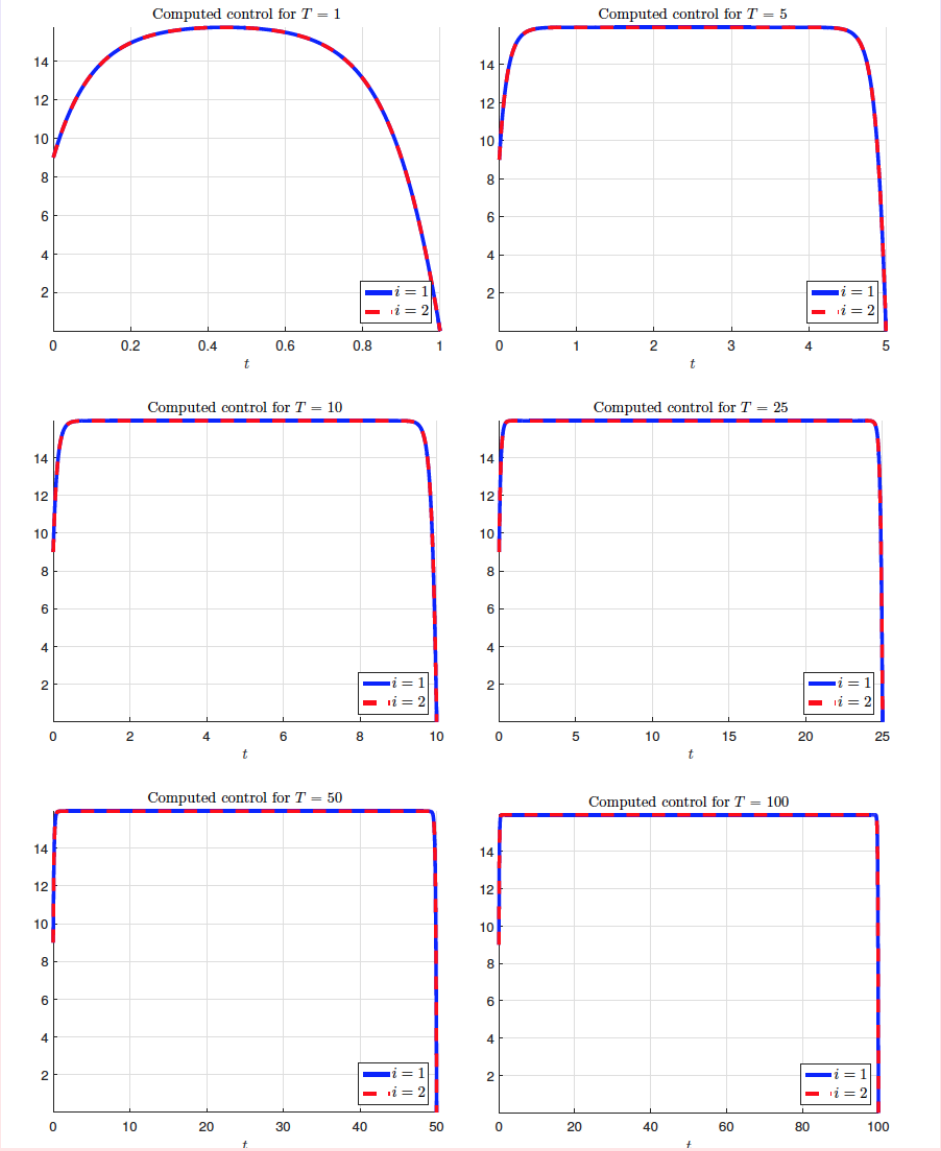
Heuristic explanation and Tip

In applications and daily life we use a quasi-turnpike principle that is very robust and universal too, even in the context of multiple steady optima (local or global).



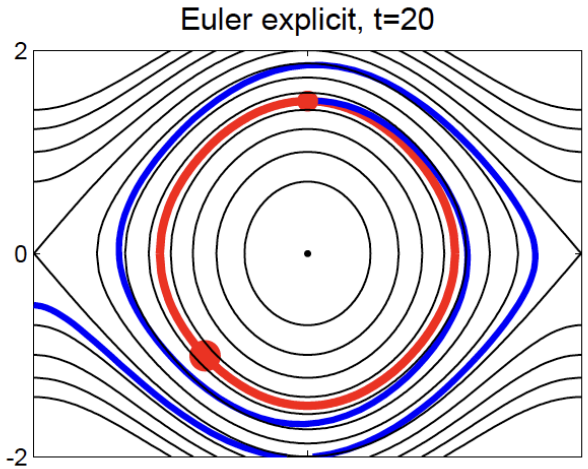
Simulations for nonlinear heat equations with arbitrary targets (S. Volkwein)

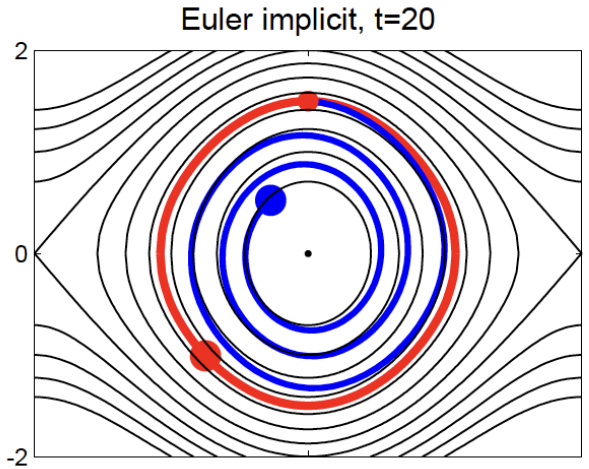
Numerical simulations show that the turnpike property is quite robust and the smallness of the target does not seem to be needed.

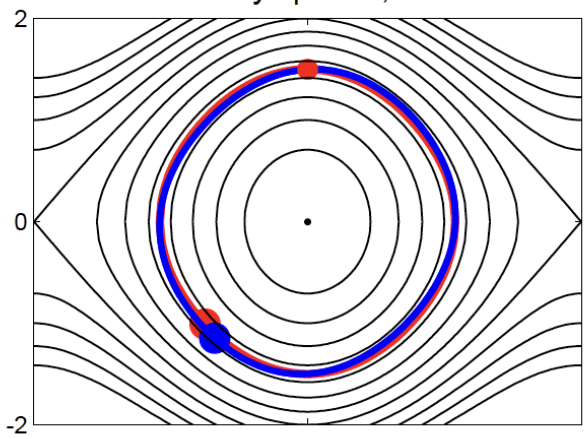


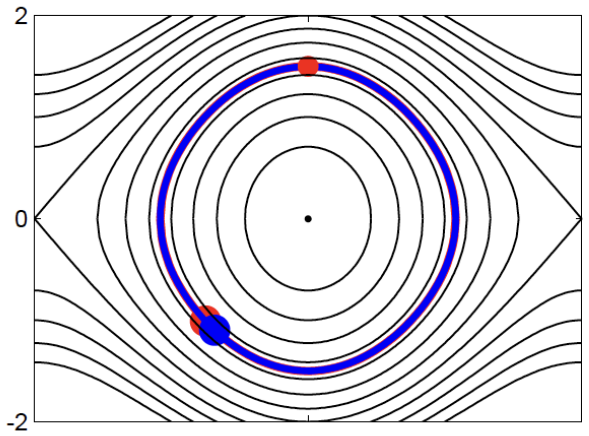
Warning! Long time numerics plays a key role: Geometric/Symplectic integration; Well balanced numerical schemes...

Numerical integration of the pendulum (A. Marica)









An open problem and biblio

Further extend the turnpike theory for nonlinear PDE, getting rid of the smallness condition on the target, which in numerical simulations seems to be unnecessary.

- A. Porretta, E. Z., SIAM J. Control. Optim., 51 (6) (2013), 4242-4273.
- A. Porretta, E. Z., Springer INdAM Series "Mathematical Paradigms of Climate Science", F. Ancona et al. eds, 15, 2016, 67-89.
- E. Trélat, E. Z., JDE, 218 (2015) , 81-114.
- M. Gugat, E. Trélat, E. Z., Systems and Control Letters, 90 (2016), 61-70.
- E. Z., Annual Reviews in Control, 44 (2017) 199-210.
- E.Trélat, C. Zhang, E. Z., SIAM J. Control Optim. **56** (2018), no. 2, 1222–1252.
- V. Hernández-Santamaria, M. Lazar, E.Z. Numerische Mathematik (2019) 141:455-493.
- D. Pighin, N. Sakamoto, E. Z., IEEE CDC Proceedings, Nice, 2019.
- G. Lance, E. Trélat, E. Z., Systems & Control Letters 142 (2020) 104733.
- J. Heiland, E. Z., arXiv:2007.13621, 2020.
- C. Esteve, H. Kouhkouh, D. Pighin, E. Z., arxiv.org/pdf/2006.10430, 2020.
- M. Gugat, M. Schuster and E. Z., SEMA/SIMAI Springer Series, 2020.

And further interesting work by collaborators: S. Zamorano (NS), M. Warma & S. Zamorano, Fractional heat,...

Our thanks to our FAU colleague Daniel Tenbrinck. He suggested to us to explore turnpike for Neural Networks.

Outline

Supervised learning..

Goal: Find an approximation of a function $f_\rho : \mathbb{R}^d \to \mathbb{R}^m$ from a dataset

$$\{\vec{x}_i, \vec{y}_i\}_{i=1}^N \subset \mathbb{R}^{d\times N} \times \mathbb{R}^{m\times N}$$

drawn from an unknown probability measure ρ on $\mathbb{R}^d \times \mathbb{R}^m$.

- **Classification**: match points (images) to respective labels (cat, dog).

$\longrightarrow$ Popular method: **training a neural network**.

..via neural networks

1. A **neural network** is a scheme: for any $i \leq N$

$$\begin{cases} \mathbf{x}_i^{k+1} = \sigma(w^k \mathbf{x}_i^k + b^k) \quad \text{for} \quad k \in \{0, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d, \end{cases} \tag{NN}$$

where

- $w^k \in \mathbb{R}^{d_{k+1} \times d_k}$ and $b^k \in \mathbb{R}^{d_k}$ are **controls**
- $\sigma(x) = \tanh(x)$ or $\sigma(x) = \max\{x, 0\}$
- $N_{layers} \geq 1$ **depth**

2. **Training**: minimize cost:

$$\inf_{\{w^k, b^k\}_{k=0}^{N_{layers}}} \underbrace{\frac{1}{N} \sum_{i=1}^{N} \text{loss}\left(\varphi\left(\mathbf{x}_i^{N_{layers}}\right), \vec{y}_i\right)}_{:=\phi\left(\mathbf{x}^{N_{layers}}\right)} + \frac{\alpha}{2} \left\| \left\{w^k, b^k\right\}_k \right\|_{\ell^2}^2$$

where

- e.g. $\text{loss}(x, y) = \|x - y\|_{\ell^p}^p$ for $p = 1, 2$;
- $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ (possibly nonlinear)

$$\varphi(x) = w^{N_{layers}} x + b^{N_{layers}}.$$

Residual neural networks

ResNets[3]: for any $i \leq N$

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + h\sigma(w^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d, \end{cases} \tag{ResNet}$$

where $h = 1$, **width** $d_k \equiv d$ is constant, .

- "**layer = timestep**"[4]; $h = \frac{T}{N_{layers}}$ for given $T > 0$:

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i. \end{cases} \tag{nODE}$$

- Supervised Learning is an **optimal control problem**:

$$\inf_{[w,b]^\top \in L^2(0,T;\mathbb{R}^{d_u})} \phi(\mathbf{x}(T)) + \frac{\alpha}{2} \left\| [w, b]^\top \right\|^2_{L^2(0,T;\mathbb{R}^{d_u})} \tag{SL}$$

where

- $\mathbf{x}(t) = [\mathbf{x}_1(t), \dots, \mathbf{x}_N(t)]^\top$ solutions to (nODE)

[3] He et al. '15

[4] E, Haber & Ruthotto '17

Objective

- $\mathbf{x}^0 := [\vec{x}_1, \ldots, \vec{x}_N]^\top$, $u := [w, b]^\top$
- ϕ continuous & nonnegative
- Asssume σ glob. Lipschitz & $\sigma(0) = 0$ and put (nODE) in the form

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), u(t)) & \text{in } (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}. \end{cases} \tag{nODE}$$

Question: What happens to a global minimizer u^T solving (SL), and corresponding state $\mathbf{x}^T$ to (nODE) when $T \to \infty$?

Interest:

$$T \to \infty \quad \sim \quad N_{\text{layers}} \to \infty.$$

Regularization

Caution before proceeding..

- For (nODE) $\longrightarrow$ L^2–regularization **may not be enough** for existence of minimizers. Due to the nonlinearity σ and lack of compactness.

$\longrightarrow$ enhance to Sobolev regularization:

$$\inf_{[w,b]^\top \in H^1(0,T;\mathbb{R}^{d_u})} \phi(\mathbf{x}(T)) + \frac{\alpha}{2} \left\| [w,b]^\top \right\|^2_{H^1(0,T;\mathbb{R}^{d_u})} \tag{SL}$$

- **Not a problem for** the "simpler" version

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w(t)\sigma(\mathbf{x}_i(t)) + b(t) & \text{for } \ t \in (0,T) \\ \mathbf{x}_i(0) = \vec{x}_i, \end{cases} \tag{nODE$_2$}$$

motivated by equivalent definition of NN:

$$\begin{cases} \mathbf{x}_i^{k+1} = w^k \sigma(\mathbf{x}_i^k) + b^k & \text{for } \ k \in \{1, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = w^0 \vec{x}_i. \end{cases}$$

All results to follow also hold for (nODE$_2$) with H^1 replaced by L^2.

Time-scaling

Key idea: Time-Scaling.

1. Given some $u^1(t)$ and solution $\mathbf{x}^1(t)$ to

$$\begin{cases} \dot{\mathbf{x}}^1(t) = \mathbf{f}(\mathbf{x}^1(t), u^1(t)) & \text{in } (0,1) \\ \mathbf{x}^1(0) = \mathbf{x}^0, \end{cases}$$

   then $u^T(t) := \frac{1}{T} u^1(\frac{t}{T})$ is such that $\mathbf{x}^T(t) := \mathbf{x}^1(\frac{t}{T})$ solves (nODE) for $t \in [0, T]$.

2. Then:

$$\begin{aligned} &\inf_{u^T \in L^2(0,T;\mathbb{R}^{d_u})} \phi(\mathbf{x}^T(T)) + \frac{\alpha}{2} \int_0^T \left\| u^T(t) \right\|^2 dt \\ &= \frac{1}{T} \inf_{u^T \in L^2(0,T;\mathbb{R}^{d_u})} T\phi(\mathbf{x}^T(T)) + \frac{\alpha}{2} \int_0^1 \left\| T u^T(sT) \right\|^2 ds \\ &= \frac{1}{T} \inf_{u^1 \in L^2(0,1;\mathbb{R}^{d_u})} T\phi(\mathbf{x}^1(1)) + \frac{\alpha}{2} \int_0^1 \left\| u^1(s) \right\|^2 ds. \end{aligned}$$

Zero training error asymptotics

Recall:

$$\mathbf{x}^\dagger \in \arg\min(\phi) \quad \Longleftrightarrow \quad \phi(\mathbf{x}^\dagger) = \min_{\mathbb{R}^{d_x}} \phi.$$

Theorem (Esteve et al. '20): For any $T > 0$, let u^T be minimizer in (SL), $\mathbf{x}^T$ associated solution to (nODE).
Under controllability/reachability assumptions, there exist a sequence $\{T_n\}_{n=1}^{+\infty}$ of positive times and $\mathbf{x}^\dagger \in \arg\min(\phi)$, such that

$$\left\| \mathbf{x}^{T_n}(T_n) - \mathbf{x}^\dagger \right\| \longrightarrow 0 \qquad \text{as} \quad n \to \infty.$$

Setting $u_n(t) = \frac{1}{T_n}\, u^{T_n}(\frac{t}{T_n})$ for $t \in [0, T_n]$, we also have

$$\left\| u_n - u^1 \right\|_{H^1(0,1;\mathbb{R}^{d_u})} \longrightarrow 0 \qquad \text{as} \quad n \to \infty$$

where u^1 solves

$$\inf_{\substack{u \in H^1(0,1;\mathbb{R}^{d_u}) \\ \text{subject to} \\ \mathbf{x}(1) \in \arg\min(\phi)}} \frac{\alpha}{2} \left\| u \right\|^2_{H^1(0,1;\mathbb{R}^{d_u})}.$$

$\longrightarrow$ Not a turnpike result!

Figure: Here $N_{\text{layers}} = \left\lfloor T^{\frac{3}{2}} \right\rfloor$ and thus $h = \frac{1}{\sqrt{T}}$, and we consider $\alpha = 1$.

Turnpike

Recall training error, assuming $\text{loss}(x, y) = \|x - y\|^2$:

$$\phi(\mathbf{x}(T)) := \frac{1}{N} \sum_{i=1}^{N} \|\varphi(\mathbf{x}_i(T)) - \vec{y}_i\|^2 ; \tag{10}$$

$\varphi : \mathbb{R}^d \to \mathbb{R}^m$ not surjective a priori!

Question: Can we have quantitative estimates for the time T required to reach the zero training error regime?

$\longrightarrow$ Consider enhanced cost

$$J_T(u) := \frac{1}{2} \int_0^T \phi(\mathbf{x}(t))dt + \frac{\alpha}{2} \|u\|^2_{H^1(0,T;\mathbb{R}^{d_u})}$$

The optimal steady states

- The steady optimal control/learning problem associated to J_T consists in minimizing

$$J_s(u^s) := \frac{1}{2}\phi(\mathbf{x}^s) + \frac{\alpha}{2}\|u^s\|^2$$

over $u^s \in \mathbb{R}^{d_u}$, where $\mathbf{x}^s \in \mathbb{R}^{d_x}$ is a steady state of (nODE):

$$\mathbf{f}(\mathbf{x}^s, u^s) = 0.$$

- Due to
  1. form of controls $u = [w, b]^\top$ and $\mathbf{f}(x, u) = \sigma(wx + b)$;
  2. $\sigma(0) = 0$

$\longrightarrow$ optimal steady-state pair is

$$(u^s, \mathbf{x}^s) = (0_{\mathbb{R}^{d_u}}, \mathbf{x}^\dagger)$$

for some $\mathbf{x}^\dagger \in \mathbb{R}^{d_x}$ such that

$$\phi(\mathbf{x}^\dagger) = \min_{\mathbb{R}^{d_x}} \phi,$$

i.e. $\mathbf{x}^\dagger \in \arg\min(\phi)$.

Turnpike property

Theorem (Esteve et al. '20): Under controllability/reachability assumptions, for any sufficiently large $T > 0$, consider a solution u^T to

$$\inf_{u \in H^1(0,T;\mathbb{R}^{d_u})} \frac{1}{2} \int_0^T \phi(\mathbf{x}(t))dt + \frac{\alpha}{2} \|u\|^2_{H^1(0,T;\mathbb{R}^{d_u})}$$

and let $\mathbf{x}^T$ be the associated state, solution to (nODE).
Then

$$\left\| u^T \right\|_{H^1(0,T;\mathbb{R}^{d_u})} \leq C$$

and there exists $\mathbf{x}^\dagger \in \arg\min(\phi)$ such that

$$\left\| \mathbf{x}^T(t) - \mathbf{x}^\dagger \right\| \leq \gamma \left(e^{-\mu t} + e^{-\mu(T-t)} \right)$$

$\forall t \in [0, T]$ and for some $C > 0$, $\gamma > 0$ and $\mu > 0$, all independent of T.

Due to the absence of final time cost:

Corollary (Esteve et al. '20): In fact,

$$\left\| \mathbf{x}^T(t) - \mathbf{x}_d \right\| \leq \gamma\, e^{-\mu t}$$

$\forall t \in [0, T]$ and for some $\gamma > 0$ and $\mu > 0$ independent of T.

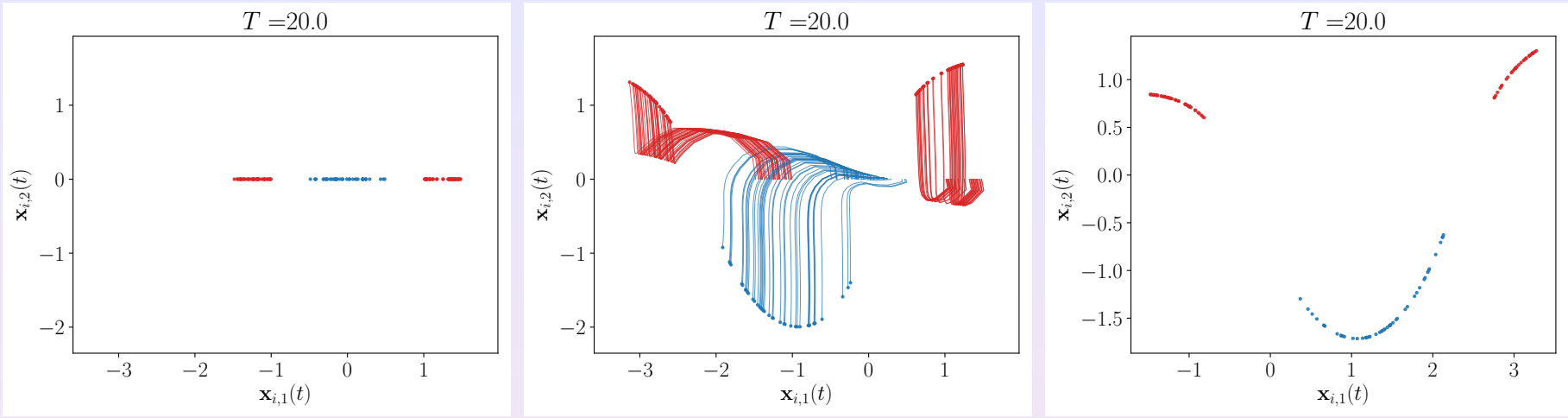
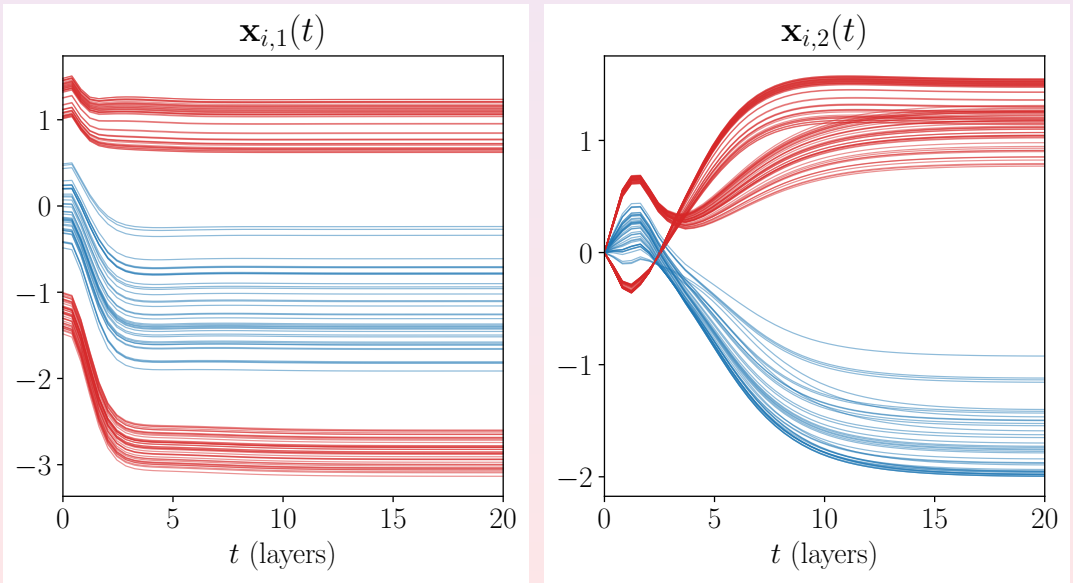


Figure: Optimal trajectories of solutions to the above learning problem: a simple flow separates the points and ensures the turnpike property. Here $T = 20$, $N_{\text{layers}} = 50$, $\alpha = 2$.



Variable width

Variable width ResNets: view width as auxiliary continuous variable

1. **Integro-differential equation**[5]

$$\partial_t \mathbf{x}_i(t,\zeta) = \sigma\left(\int_\Omega w(t,\zeta,\xi)\mathbf{x}_i(t,\xi)d\xi + b(t,\zeta)\right) \qquad \text{in } (0,T)\times\Omega.$$

   - e.g. Ω = image $\times (0,1) \subset \mathbb{R}^3$; asymptotics theorems apply here;

2. **Switched systems**: Changing widths over layers as switched systems over time:

$$\dot{\mathbf{x}}(t) = \mathbf{f}_{\rho(t)}(\mathbf{x}(t), u(t))$$

given M vector fields $\mathbf{f}_1, \dots, \mathbf{f}_M$ and switching signal $\rho : [0,T] \to \{1,\dots,M\}$;

$\longrightarrow$ **Quasi-turnpike strategy:**

#1 increase the dimension to the "optimal system" $\mathbf{f}_{j^*}$,
#2 use the turnpike for fixed width
#3 switch back.

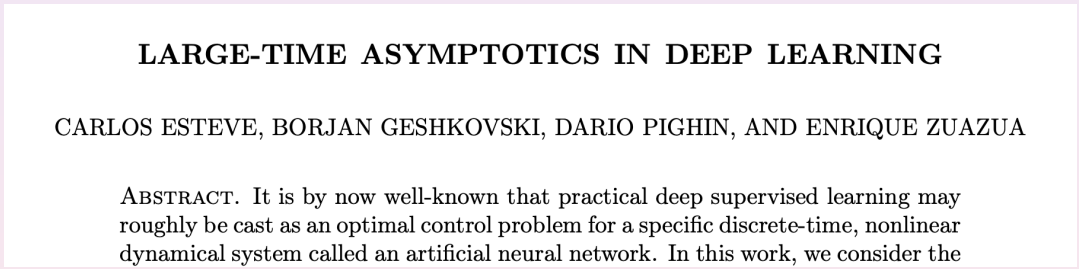
The optimal system $\mathbf{f}_{j^*}$? $\longrightarrow$ optimal with respect to cost.
What are the switching times? How many?

[5] Liu & Markowich '19

Outlook

1. Long-time behavior depends on the cost functional to be minimized.
2. Results should be complemented by ML subfields (e.g. CNN design, training algorithms..)

Many other open problems and extensive bibliography can be found in our paper:



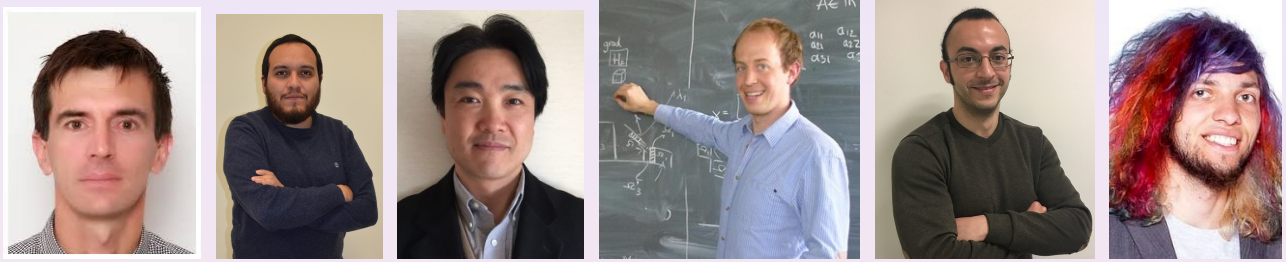
LARGE-TIME ASYMPTOTICS IN DEEP LEARNING

CARLOS ESTEVE, BORJAN GESHKOVSKI, DARIO PIGHIN, AND ENRIQUE ZUAZUA

ABSTRACT. It is by now well-known that practical deep supervised learning may roughly be cast as an optimal control problem for a specific discrete-time, nonlinear dynamical system called an artificial neural network. In this work, we consider the

https://arxiv.org/abs/2008.02491

Team, collaborators, funding



- E. Trélat (Paris Sorbonne), A. Porretta (Roma 2), M. Gugat (FAU), D. Pighin (Innovalia), C. Esteve (UAM & Deusto), M. Lazar (Dubrovnik), V. Hernández-Santamaria, N. Sakamoto (Nanzan), J. Heiland (Magdeburg), H. Kouhkouh (Padova), M. Schuster (FAU).
- Funded by the ERC Advanced Grant DyCon and an Alexander von Humboldt Professorship and Marie-Sklodowska Curie ITN "ConFlex"

Alexander von Humboldt Stiftung/Foundation

DyCon DYNAMIC CONTROL erc

ConFlex www.conflex.org

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 765579.

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

COM Fundación Deusto Fundazioa

UAM Universidad Autónoma de Madrid

Thank you for your attention.