The interplay of Deep Learning and Control Theory

Borjan Geshkovski

AG "Mathematics of Deep Learning", FAU Erlangen-Nurnberg
December 9th, 2020

Supervised learning

Supervised learning

Goal: Find an approximation of a function $f(\cdot)$ from a dataset

$$\{\vec{x}_i, \vec{y}_i = f(\vec{x}_i)\}_{i=1}^{N}$$

drawn from an unknown probability distribution $p = p(x, y)$ on $\mathbb{R}^d \times \mathbb{R}^m$.

- **Classification:** $f : \mathbb{R}^d \to \{1, \ldots, m\}$, thus labels $\vec{y}_i \in \{1, \ldots, m\}$.
- **Regression:** $f : \mathbb{R}^d \to \mathbb{R}^m$, thus labels $\vec{y}_i \in \mathbb{R}^m$.

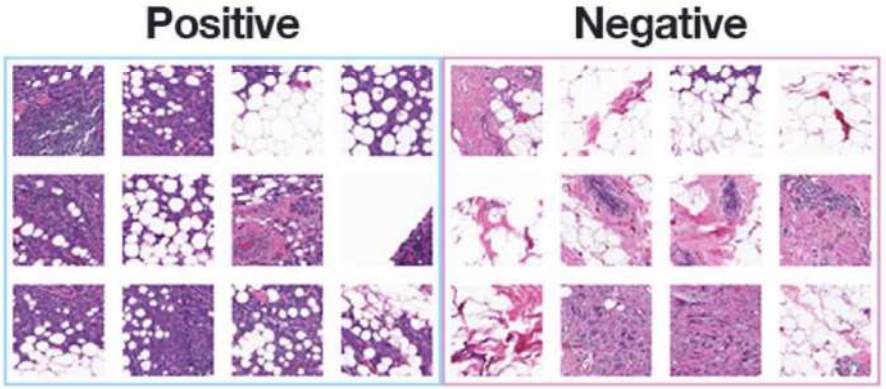


Figure: Classification – $f : \mathbb{R}^{9216} \to \{-1, +1\}$

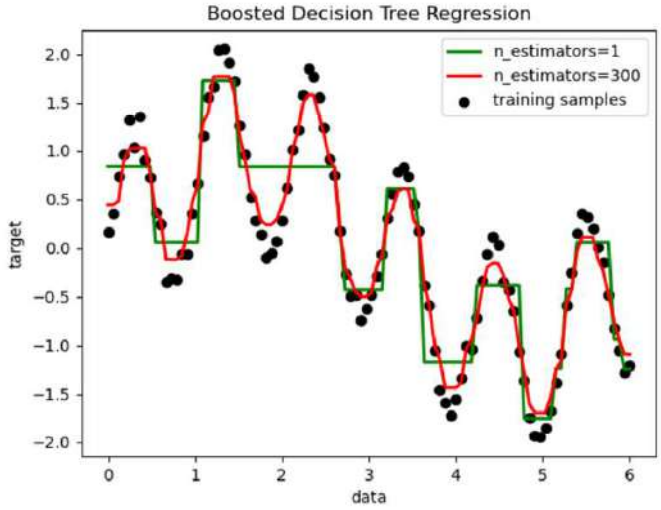


Figure: Regression – $f : [0, 6] \to \mathbb{R}$

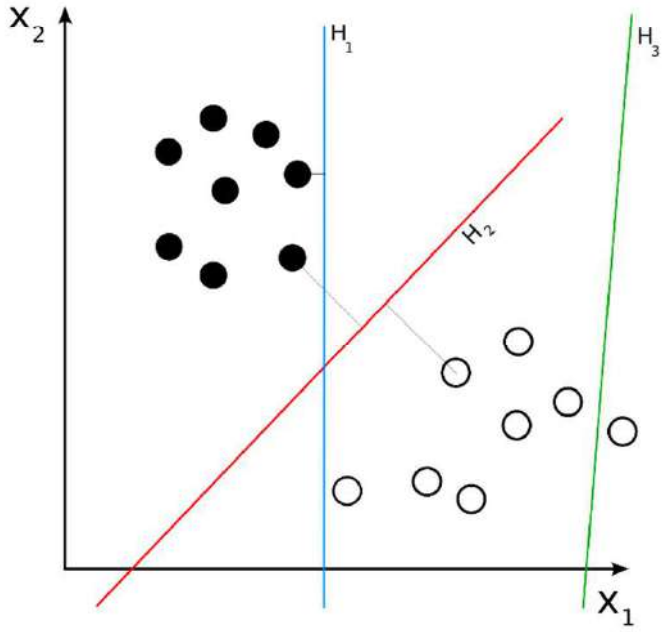
How to solve such tasks?

Suppose we are looking for $f : \mathbb{R}^d \to \{-1, 1\}$, thus given data $\vec{x}_i \in \mathbb{R}^2$ and $\vec{y}_i \in \{-1, 1\}$ for $i \leq N$.

- A simple idea:

$$\min_{w \in \mathbb{R}^2} \sum_{i=1}^{N} \left\| \text{sign}(w^\top \vec{x}_i) - \vec{y}_i \right\|^2$$

  and $x \mapsto \text{sign}(w^\top x)$ will be approximation candidate;
- But data is not linearly separable in general!



Neural networks

Neural network: for any $i \leq N$

$$\begin{cases} \mathbf{x}_i^{k+1} = \sigma(w^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d, \end{cases} \tag{NN$_1$}$$

- $w^k \in \mathbb{R}^{d_{k+1} \times d_k}$ and $b^k \in \mathbb{R}^{d_k}$ are *controls*;
- $N_{layers} \geq 1$ given **depth**; $d_k \geq 1$ called **widths** with $d_0 = d$ and $d_{N_{layers}} = m$.
- $\sigma \in \mathrm{Lip}(\mathbb{R})$ & $\sigma(0) = 0$ defined componentwise:

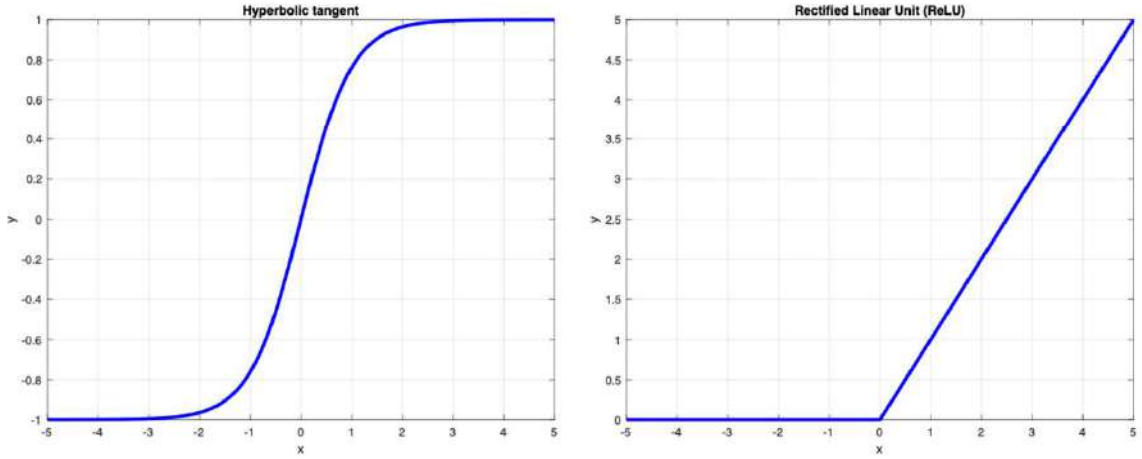


Figure: Sigmoid: $\tanh(x)$ and ReLU: $\max\{x, 0\}$

- ML jargon: multilayer perceptron / fully-connected.

Universal approximation

- Neural networks are universal approximators[1]: if σ not polynomial, the set

$$H := \left\{ f \,:\, f(x) = \sum_{j=1}^{n} \alpha_j \sigma(\beta_j^\top x + \gamma_j); \quad \alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times d}, \gamma \in \mathbb{R}^n, n \geq 1 \right\}$$

  is dense in $C^0([-1,1]^d)$. So here width n is large.
- Plethora of extensions[2]; Dual view of large depth has also been studied[3].
- Maiorov and Pinkus '99: $\exists \sigma$ such that $f \in C^0([-1,1]^d)$ may approximated by a two-hidden layer NN with $(2d+1)(4d+3)$ neurons in layer 1 and $4d+3$ neurons in layer 2. Uses Kolmogorov-Arnold rpz:

$$f(x) = \sum_{j=0}^{2d} \Phi \left(\sum_{i=1}^{d} \alpha_i \phi(x_i + \eta j) + j \right)$$

- Results do not say how to find the parameters.

---

[1] Cybenko '89

[2] Hornik '89, Barron '90s, Pinkus '99, Burger et al. '01, deVore, Daubechies et al. '19, Kutyniok et al. '19, etc..

[3] Kidger & Lyons '19: "Universal Approximation with Deep Narrow Networks" and references

"Training" a NN

Training $\Longleftrightarrow$ **Optimization**: $\lambda > 0$ fixed,

$$\min_{\left\{w^k, b^k\right\}_{k=0}^{N_{layers}}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(P\mathbf{x}_i^{N_{layers}}, \vec{y}_i\right)}_{\text{training error}} + \lambda \underbrace{\left\|\left\{w^k, b^k\right\}_k\right\|_{\ell^p}^p}_{\text{regularization}}$$

1. **Regression:** $\vec{y}_i \in \mathbb{R}^m$, and

$$Px = w^{N_{layers}} x + b^{N_{layers}} \in \mathbb{R}^m, \quad \text{loss}(x, y) = \|x - y\|^2.$$

2. **Classification:**
   - 2 classes: $\vec{y}_i \in \{-1, 1\}$,

$$\text{loss}(Px, y) = \left\|\tanh(w^{N_{layers}} x + b^{N_{layers}}) - y\right\|^2$$

   or $Px = w^{N_{layers}} x + b^{N_{layers}} \in \mathbb{R}$ and

$$\text{loss}(Px, y) = \log\left(1 + \exp\left(-yPx\right)\right)$$

   - $m \geq 2$: cross-entropy loss.

We shall assume that P is given, possibly picked at random.

Residual neural networks

ResNets: fix $d_k \equiv d$; for any $i \leq N$

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

where $h = 1$.

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

$$\boxed{\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i. \end{cases}} \tag{nODE$_1$}$$

For (nODE$_1$), we shall henceforth assume $\sigma(\lambda x) = \lambda\sigma(x)$ for $\lambda > 0$ (positive homogeneity).

[4] Weinan E '17

Residual neural networks

In addition to (nODE_1), one can also consider variants:

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

- Also

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w_1(t)\sigma(w_2\mathbf{x}_i(t) + b_2) + b_1(t) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \end{cases} \qquad (\text{nODE}_3)$$

where $w_1 \in \mathbb{R}^{d \times d_1}$, $w_2 \in \mathbb{R}^{d_1 \times d}$.

Training is optimal control

Given $T, \lambda > 0$:

$$\inf_{[w,b] \in H^k(0,T;\mathbb{R}^{d_u})} \underbrace{\frac{1}{N} \sum_{i=1}^{N} \operatorname{loss}\left(P\mathbf{x}_i(T), \vec{y}_i\right)}_{=:\mathcal{E}(\mathbf{x}(T))} + \lambda \left\| [w, b] \right\|^2_{H^k(0,T;\mathbb{R}^{d_u})}$$

- $k = 0$ for (nODE_2), $k = 1$ for (nODE_1), (nODE_3) (L^2–regularization **may not be enough** for compactness)

Why ODEs?

ODE formulation has been used to great effect..

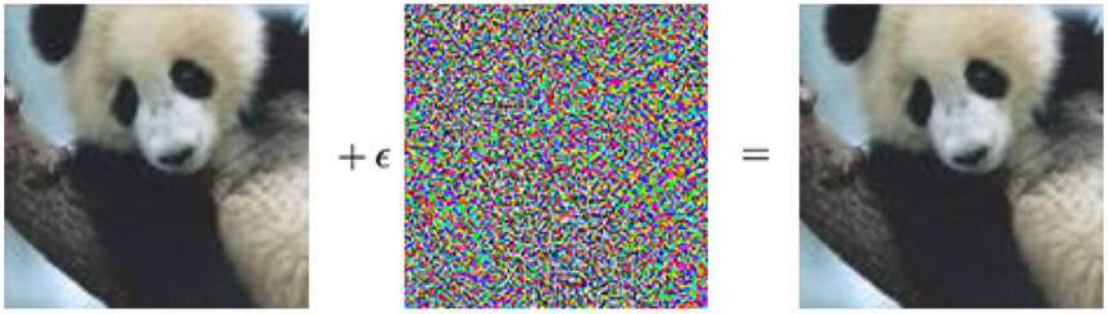
Neural ordinary differential equations [PDF] nips.cc

RTQ Chen, Y Rubanova, J Bettencourt... - Advances in **neural** ..., 2018 - papers.nips.cc

... at 3 Replacing residual networks with **ODEs** for supervised learning In this section, we experimentally investigate the training of **neural ODEs** for supervised learning. Software ...

☆ 99 Cited by 729 Related articles All 20 versions »

- **adaptive schemes, solvers** (Chen et al. '18, Dupont et al. '19, Benning et al. '19)
- **PMP–based training algos** (E et al. '19)
- **Stability to adversarial perturbations** (Haber, Ruthotto et al. '18)



"panda"
57.7% confidence

"gibbon"
99.3% confidence

Artificial intelligence / Machine learning

A radical new neural network design could overcome big challenges in AI

Researchers borrowed equations from calculus to redesign the core machinery of deep learning so it can model continuous processes like changes in health.

by **Karen Hao** December 12, 2018

MIT Tech Review, 2018

Why ODEs?

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

- Set $\mathbf{x}^0 = [\vec{x}_1, \dots, \vec{x}_N]$, $u = [w, b]$, and put both (nODE_1) and (nODE_2) in the form

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

- And so

$$\inf_{\substack{u \in H^k(0,T;\mathbb{R}^{d_u}) \\ \text{subject to (nODE)}}} \mathcal{E}(\mathbf{x}(T)) + \lambda \|u\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \tag{SL_1}$$

Question: What happens to a minimizer u^T solving (SL_1), and corresponding state $\mathbf{x}^T$ to (nODE) when $T \to +\infty$?

Empirical risk minimization

Scaling

$$\inf_{\substack{u \in H^k(0,T;\mathbb{R}^{d_u}) \\ \text{subject to (nODE)}}} \mathcal{E}(\mathbf{x}(T)) + \lambda \|u\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \qquad (\mathrm{SL}_1)$$

Key idea: *Time-Scaling*.

- Assumptions on σ entail $\mathbf{f}(\mathbf{x}, u)$ positively homogeneous w.r.t. u, i.e. $\mathbf{f}(\mathbf{x}, \alpha u) = \alpha \mathbf{f}(\mathbf{x}, u)$ for $\alpha > 0$.
- Hence, given $u^T(t)$ and the solution $\mathbf{x}^T(t)$ to

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

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

Classification

- For simplicity: $\vec{y}_i \in \{-1, +1\}$;

$$\text{loss}(P\mathbf{x}_i(T), \vec{y}_i) := \log\left(1 + e^{-\vec{y}_i P\mathbf{x}_i(T)}\right), \tag{2}$$

- Denote

$$\overline{u}^T := \frac{u^T}{\left\|u^T\right\|_{H^k(0,T;\mathbb{R}^{d_u})}},$$

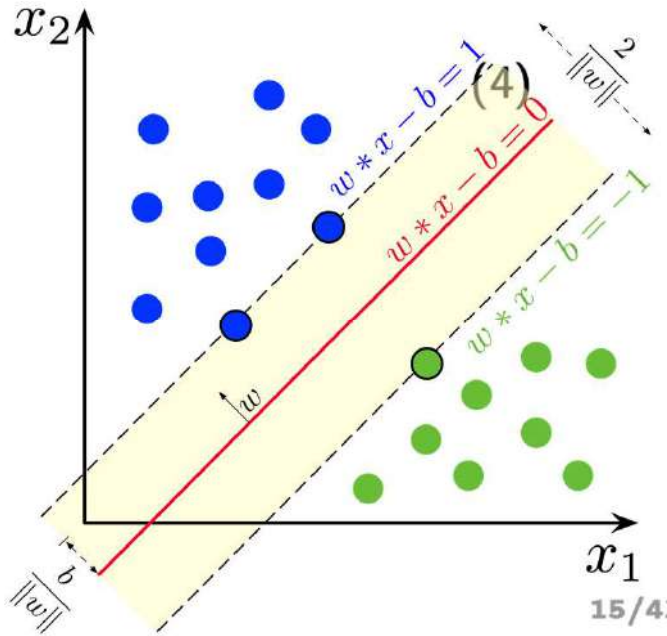
  let $\overline{\mathbf{x}}^T$ denote the associated solution to (1).

- The *margin* of $\overline{u}^T$:

$$\gamma_{\overline{u}^T} := \min_{1 \le i \le N} \vec{y}_i \, P\overline{\mathbf{x}}_i^T(T). \tag{3}$$

- Max-margin:

$$\gamma^* := \sup_{\substack{\|u\|_{H^k(0,1;\mathbb{R}^{d_u})} \le 1 \\ \mathbf{x} \text{ solves (1)}}} \gamma_u. \tag{4}$$



Theorem (Classification): Consider $\mathbf{x}^0 \in \mathbb{R}^{d_x}$ where

$$\mathbf{x}_i^0 := [\vec{x}_{i,1}, \dots, \vec{x}_{i,d}, 0] \in \mathbb{R}^{d+1}$$

for any $i \leq N$. Let $\lambda > 0$ be fixed, and let $P : \mathbb{R}^{d+1} \to \mathbb{R}$ be any non-zero matrix such that $P\mathbf{x}_i^0 = 0$ for $i \leq N$. For any $T > 0$, let $u^T \in H^k(0, T; \mathbb{R}^{d_u})$ be any global minimizer. Assume $\gamma^* > 0$.

1. There exists a constant $C > 0$ independent of $T > 0$ such that

$$\mathcal{E}(\mathbf{x}_T(T)) \leq \frac{C}{T}.$$

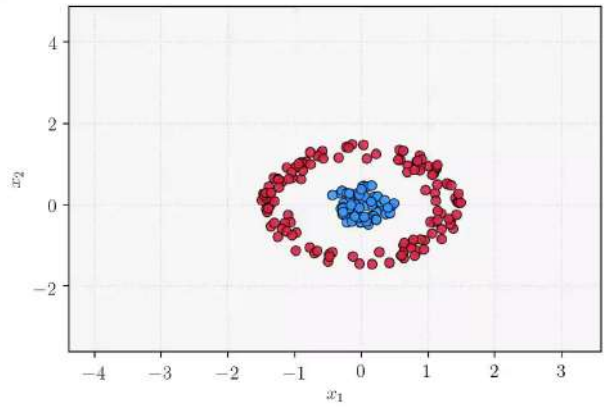
2. Moreover, $\gamma_{\bar{u}_T} \xrightarrow[T \to +\infty]{} \gamma^*$.

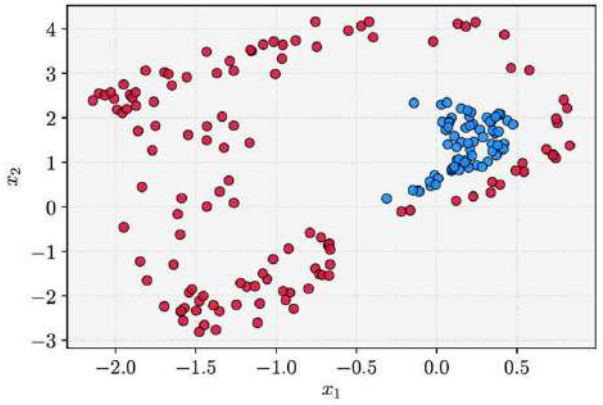
3. $\exists \{T_n\}_{n=1}^{+\infty}$ with $T_n > 0$ and $T_n \to \infty$ such that

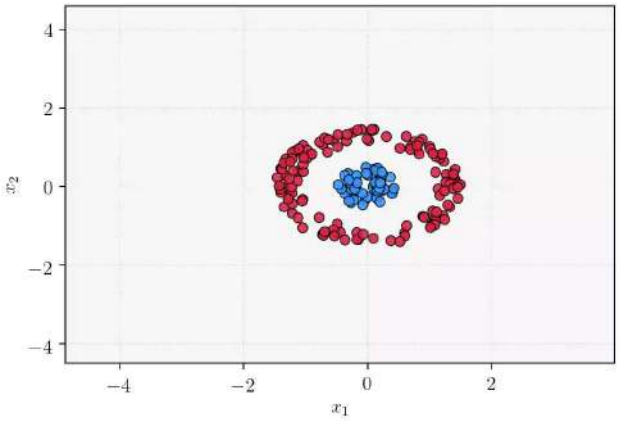
$$\left\| \frac{T_n u^{T_n}(\cdot T_n)}{\left\| T_n u^{T_n}(\cdot T_n) \right\|_{H^k(0,1;\mathbb{R}^{d_u})}} - u^* \right\|_{H^k(0,1;\mathbb{R}^{d_u})} \xrightarrow[n \to +\infty]{} 0$$

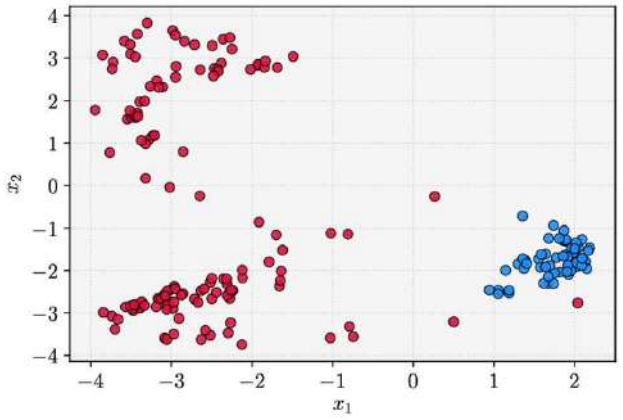
along some subsequence, where $[w^*, b^*] =: u^*$ are such that

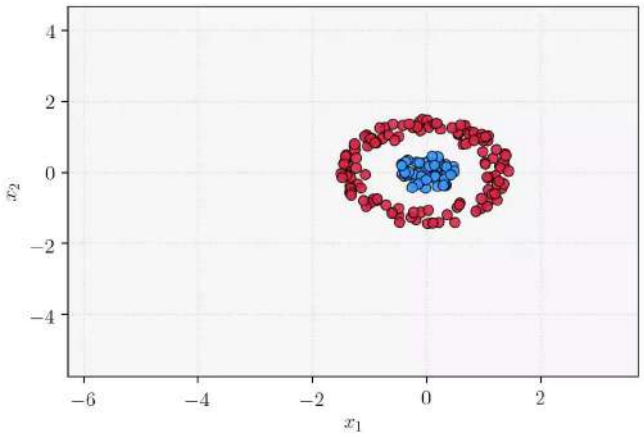
$$\gamma_{u^*} = \sup_{\substack{\|u\|_{H^k(0,1;\mathbb{R}^{d_u})} \leq 1 \\ \mathbf{x} \text{ solves } (1)}} \gamma_u = \gamma^*.$$

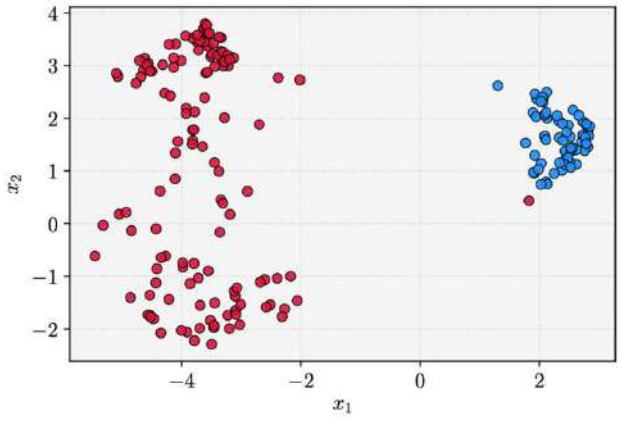

x_1
x_2


x_1
x_2


x_1
x_2


x_1
x_2


x_1
x_2


x_1
x_2

Theorem (Regression[a]): Fix $\lambda > 0$, let $P : \mathbb{R}^d \to \mathbb{R}^m$ be any surjective affine map. For any $T > 0$, let u^T be minimizer in (SL_1), $\mathbf{x}^T$ associated solution to (nODE). Assume that $\{\mathcal{E} = 0\}$ is reachable by (nODE). Then

1. $\exists C > 0$ independent of T such that

$$\mathcal{E}(\mathbf{x}^T(T)) \leq \frac{C}{T}.$$

2. Moreover, $\exists \{T_n\}_{n=1}^{+\infty}$ positive times and $\exists \mathbf{x}_\circ \in \mathbb{R}^{d_x}$, $\mathcal{E}(\mathbf{x}_\circ) = 0$, such that

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

3. Moreover

$$\left\| \frac{1}{T_n} u^{T_n}\left(\frac{\cdot}{T_n}\right) - u^* \right\|_{H^k(0,1;\mathbb{R}^{d_u})} \longrightarrow 0 \qquad \text{as } n \to +\infty$$

where u^* solves

$$\inf_{\substack{u \in H^k(0,1;\mathbb{R}^{d_u}) \\ \text{subject to (nODE) with } T=1 \\ \text{and} \\ \mathcal{E}(\mathbf{x}(1))=0}} \|u\|^2_{H^k(0,1;\mathbb{R}^{d_u})}.$$

[a] Carlos Esteve et al. "Large-time asymptotics in deep learning". In: *arXiv preprint arXiv:2008.02491* (2020).

$$T \to \infty \iff \lambda \to 0$$

Back to

$$\begin{aligned}
\mathcal{E}(\mathbf{x}^T(T)) + \lambda \int_0^T \left\| u^T(t) \right\|^2 dt &= \mathcal{E}(\mathbf{x}^T(T)) + \frac{\lambda}{T} \int_0^1 \left\| T u^T(sT) \right\|^2 ds \\
&= \mathcal{E}(\mathbf{x}^T(T)) + \frac{\lambda}{T} \int_0^1 \left\| u^1(s) \right\|^2 ds \\
&= \mathcal{E}(\mathbf{x}^1(1)) + \frac{\lambda}{T} \int_0^1 \left\| u^1(s) \right\|^2 ds.
\end{aligned}$$

Corollary: All of the conclusions of both Theorems remain true when $T > 0$ is fixed and $\lambda \searrow 0$.

Discussion

- For solution $\widehat{w}^\lambda$ to

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{N} \log\left(1+\exp\left(-\vec{y}_i w^\top \vec{x}_i\right)\right) + \lambda \|w\|_{\ell^p}^p$$

shown[5] that

$$\lim_{\lambda\to 0} \frac{\widehat{w}^\lambda}{\left\|\widehat{w}^\lambda\right\|} = w^*$$

where w^* is maximum margin separator:

$$w^* = \text{argmax}_{\|w\|_p=1} \min_i \vec{y}_i w^\top \vec{x}_i.$$

- Compared to other convergence results of generalization nature: **implicit regularization of gradient descent**[6]:

"In the overparametrized regime, after training a neural network with gradient-based methods until zero training error, with $\lambda = 0$, among the many classifiers which overfit on the training dataset, the algorithm selects the one which performs best on the test dataset."

[5] Rosset, Hu, Hastie '04

[6] Zhang et al. '16, Soudry et al. '18, Gunasekar et al. '18, Chizat & Bach '20

Proof of Theorem (Regression)

For simplicity, suppose $k = 0$.

Part 1). We first show that

$$\mathcal{E}(\mathbf{x}^T(T)) \lesssim T^{-1}.$$

1. By controllability, $\exists u^1 \in L^2(0,1)$ such that $\mathcal{E}(\mathbf{x}^1(1)) = 0$.
2. Since u^T is a minimizer,

$$\begin{aligned}\mathcal{E}\left(\mathbf{x}^T(T)\right) + \lambda \left\|u^T\right\|^2_{L^2(0,T)} &\leq \mathcal{E}\left(\mathbf{x}^1(1)\right) + \lambda \left\|\frac{\cdot}{T} u^1\left(\frac{\cdot}{T}\right)\right\|^2_{L^2(0,T)} \\ &= \frac{\lambda}{T}\left\|u^1\right\|^2_{L^2(0,1)}\end{aligned}$$

Part 2). Now show that $\exists\{T_n\}_{n=1}^{+\infty}$ of positive times and $\exists \mathbf{x}_\circ \in \mathbb{R}^{d_x}$, $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that

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

1. Grönwall + scaling:

$$\begin{aligned} \left\| \mathbf{x}^T(T) - \mathbf{x}^0 \right\| &\lesssim_{N,\sigma} \sqrt{T} \left\| u^T \right\|_{L^2(0,T)} \exp\left(\sqrt{T} \left\| u^T \right\|_{L^2(0,T)} \right) \\ &\lesssim_{N,\sigma} \left\| u^1 \right\|_{L^2(0,1)} \exp\left(\left\| u^1 \right\|_{L^2(0,1)} \right) \end{aligned}$$

   Thus $\left\{\mathbf{x}^T(T)\right\}_{T>0}$ is bounded (subset of $\mathbb{R}^{d_x}$);

2. $\longrightarrow \exists\{T_n\}_{n=1}^{+\infty}$ of positive times and $\exists \mathbf{x}_\circ \in \mathbb{R}^{d_x}$ such that

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

3. By **Part 1)**, $\mathcal{E}\left(\mathbf{x}^{T_n}(T_n)\right) \longrightarrow 0$. We conclude by continuity of $\mathcal{E}$.

Part 3). We finally show that $u_n(t) := \frac{1}{T_n} u^{T_n}(\frac{t}{T_n})$ for $t \in [0, T_n]$ satisfies

$$\|u_n - u^*\|_{L^2(0,1)} \longrightarrow 0 \qquad \text{as } n \to +\infty$$

where u^* solves

$$\inf_{\substack{u \in L^2(0,1) \\ \text{subject to (nODE) with } T=1 \\ \text{and} \\ \mathcal{E}(\mathbf{x}(1))=0}} \|u\|^2_{L^2(0,1)} \cdot$$

1. Assume $\|u_n\|_{L^2(0,1)} \leq \|u^0\|_{L^2(0,1)}$ for every $n \geq 1$; u^0 solution of above.
2. $\exists u^* \in L^2(0,1)$ such that

$$u_n \rightharpoonup u^* \qquad \text{weakly in } L^2(0,1)$$

compactness of ODE:

$$\mathbf{x}_n \longrightarrow \mathbf{x}^* \qquad \text{strongly in } C^0[0,1]$$

3. But $\mathbf{x}^{T_n}(T_n) = \mathbf{x}_n(1)$ thus $\mathbf{x}^*(1) = \mathbf{x}_\circ$ by **Part 1)**, so $\mathcal{E}(\mathbf{x}^*(1)) = 0$.
4. Weak lower semicontinuity of L^2–norm:

$$\|u^0\|^2_{L^2(0,1)} \leq \|u^*\|^2_{L^2(0,1)} \leq \liminf_{n\to\infty} \|u_n\|^2_{L^2(0,1)} \leq \limsup_{n\to\infty} \|u_n\|^2_{L^2(0,1)} \leq \|u^0\|^2_{L^2(0,1)}$$

so strong L^2–convergence and u^* solves the desired problem.

Proof of Theorem (Classification)

$$J_T(u) := \sum_{i=1}^{N} \log\left(1 + e^{-\vec{y}_i P \mathbf{x}_i(T)}\right) + \lambda \|u\|^2_{H^k(0,T)}.$$

We will concentrate on showing

$$\underbrace{\min_i \ \vec{y}_i \, P \, \bar{\mathbf{x}}_i^T(T)}_{:=\gamma_{\bar{u}T}} \longrightarrow \underbrace{\sup_{\|u\|_{H^k(\mathbf{0},\mathbf{1})} \leq 1} \min_i \vec{y}_i P \mathbf{x}_i(1)}_{=:\gamma^*}$$

as $T \to +\infty$.

- Choice of $P \Longrightarrow S_T(\mathbf{x}^0, \alpha u) = \alpha S_T(\mathbf{x}^0, u) = \alpha \mathbf{x}(T)$ for $\alpha > 0$.
- For $\alpha > 0$, u given:

$$J_T(\alpha u) \ \leq \log\left(1 + \exp\left(-\alpha \min_i \vec{y}_i P \mathbf{x}_i(T)\right)\right) + \lambda \alpha^2 \|u\|^2_{H^k(0,T)} \qquad (5)$$

and

$$J_T(\alpha u) \geq \frac{1}{N} \ \log\left(1 + \exp\left(-\alpha \min_i \vec{y}_i P \mathbf{x}_i(T)\right)\right) + \lambda \alpha^2 \|u\|^2_{H^k(0,T)}. \qquad (6)$$

- Let $\|u^*\|_{H^k(0,1)}$ such that $\gamma_{u^*} = \gamma^*$; u_T^* rescaled on $[0, T]$:

$$J_T\left(\sqrt{T}\|u_T\|_{H^k} u_T^*\right) \leq \log\left(1 + \exp\left(\sqrt{T}\|u_T\|_{H^k}\gamma^*\right)\right) + \lambda\|u_T\|_{H^k}^2.$$

  Also

$$J_T(u_T) \geq \frac{1}{N} \log\left(1 + \exp\left(-\|u_T\|_{H^k}\gamma_{\bar{u}T}\right)\right) + \lambda\|u_T\|_{H^k(0,T)}^2.$$

- Thus

$$\log\left(1 + \exp\left(\sqrt{T}\|u_T\|_{H^k}\gamma^*\right)\right) \geq \frac{1}{N} \log\left(1 + \exp\left(-\|u_T\|_{H^k}\gamma_{\bar{u}T}\right)\right)$$

- Since $\sqrt{T}\|u_T\|_{H^k(0,T)} \to +\infty$, we can Taylor and conclude..

Augmented empirical risk minimization

Enhancing the decay rate

Question: Better quantitative estimates for the time T required to approach the zero training error regime $\mathcal{E}(\mathbf{x}(T)) = 0$?

- Consider $\mathrm{loss}(x, y) = \|x - y\|^2$ and so we recall

$$\mathcal{E}(\mathbf{x}(T)) := \frac{1}{N} \sum_{i=1}^{N} \|P\mathbf{x}_i(T) - \vec{y}_i\|^2$$

- We shall suppose $P : \mathbb{R}^d \to \mathbb{R}^m$ surjective, Lipschitz, but arbitrary
- and let $\bar{\mathbf{x}} \in \mathbb{R}^{d_x}$ s.t. $\bar{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ for $i \leq N$ be fixed.
- Augmented problem:

$$\inf_{\substack{u \in H^k(0,T;\mathbb{R}^{d_u}) \\ \text{subject to (nODE)}}} \mathcal{E}(\mathbf{x}(T)) + \int_0^T \|\mathbf{x}(t) - \bar{\mathbf{x}}\|^2 dt + \lambda \|u\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \qquad \text{(SL*)}$$

Exponential decay

Theorem[a]: Fix $\lambda > 0$, and suppose that (nODE) is controllable with linear estimate of the cost. There exist $T^* > 0$ such that for any $T \geq T^*$, any solution $(u^T, \mathbf{x}^T)$ to (SL*)–(nODE) satisfies

$$\|\mathbf{x}^T(t) - \bar{\mathbf{x}}\| \leq C_1 e^{-\mu t} \qquad \forall t \in [0, T]$$

and

$$\mathcal{E}\left(\mathbf{x}^T(t)\right) \leq C_2 e^{-\mu t} \qquad \forall t \in [0, T]$$

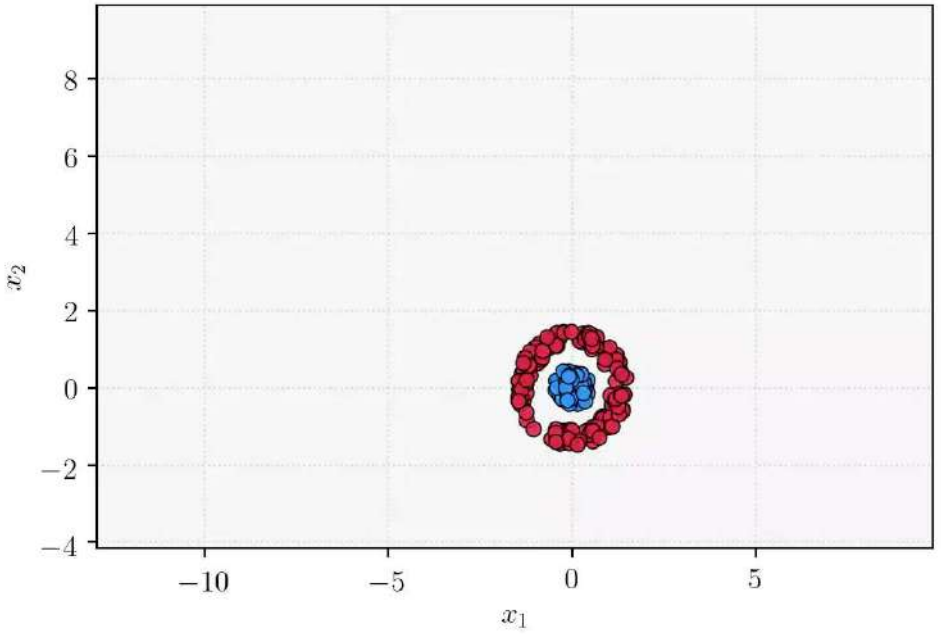
and

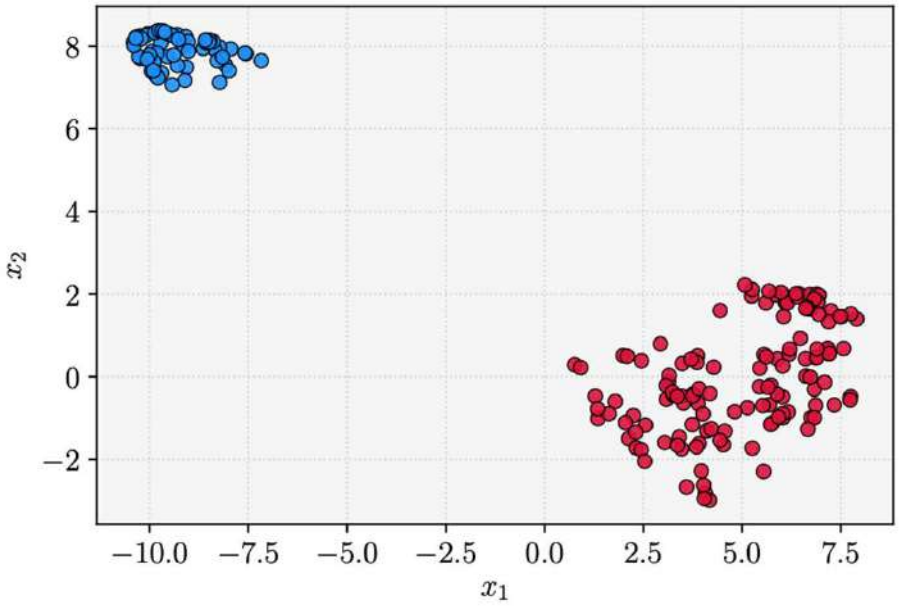
$$\left\|u^T(t)\right\| \leq C_3 e^{-\mu t} \qquad \text{for a.e. } t \in [0, T]$$

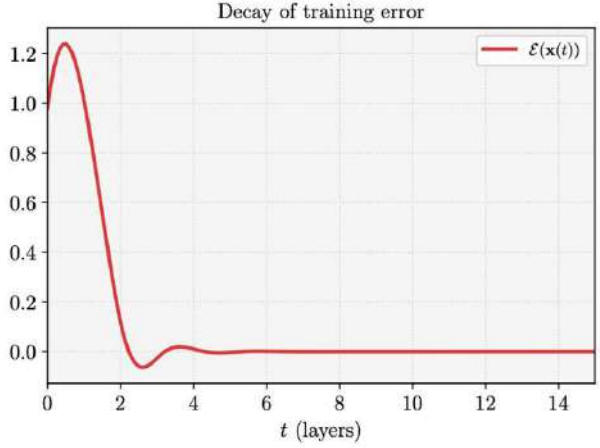
for some $C_1, C_2, C_3, \mu > 0$, all independent of T.

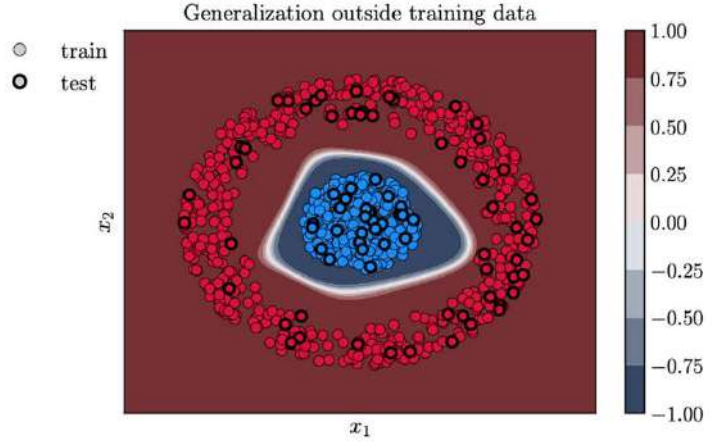
[a] Carlos Esteve et al. "Large-time asymptotics in deep learning". In: *arXiv preprint arXiv:2008.02491* (2020), Carlos Esteve et al. "Turnpike in Lipschitz-nonlinear optimal control". In: *arXiv preprint arXiv:2011.11091* (2020).

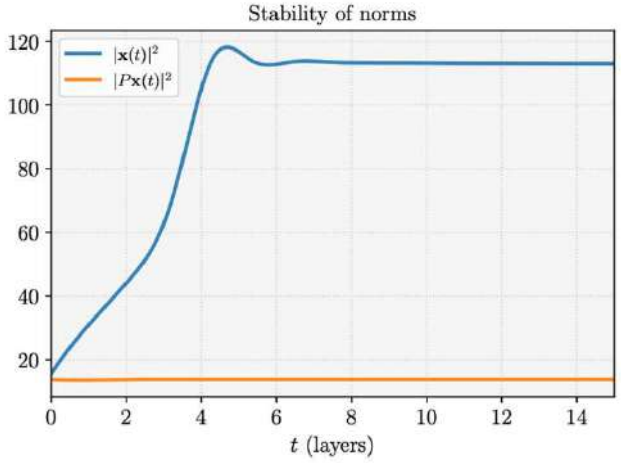
- Akin to *universal approximation*: given tolerance $\varepsilon > 0$, there exists $T_\varepsilon > 0$ (number of layers) and control parameters u^ε such that the neural network output is ε–close to the desired target.
- One difference with universal approximation is that our parameters may be computed explicitly via a training procedure.


8
6
4
2
0
−2
−4
x_2
−10
−5
0
5
x_1


8
6
4
2
0
−2
x_2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
x_1


Decay of training error
$\mathcal{E}(\mathbf{x}(t))$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
t (layers)


Generalization outside training data
train
test
x_2
x_1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Stability of norms
$|\mathbf{x}(t)|^2$
$|P\mathbf{x}(t)|^2$
120
100
80
60
40
20
0
2
4
6
8
10
12
14
t (layers)

A related problem

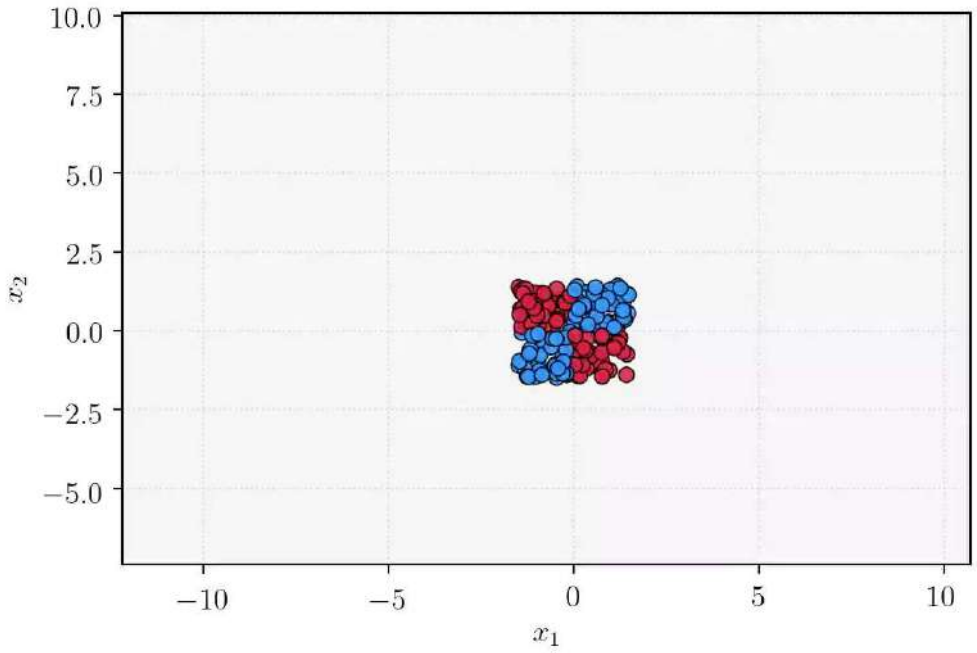
Take any $\mathrm{loss}(\cdot,\cdot)$ and consider

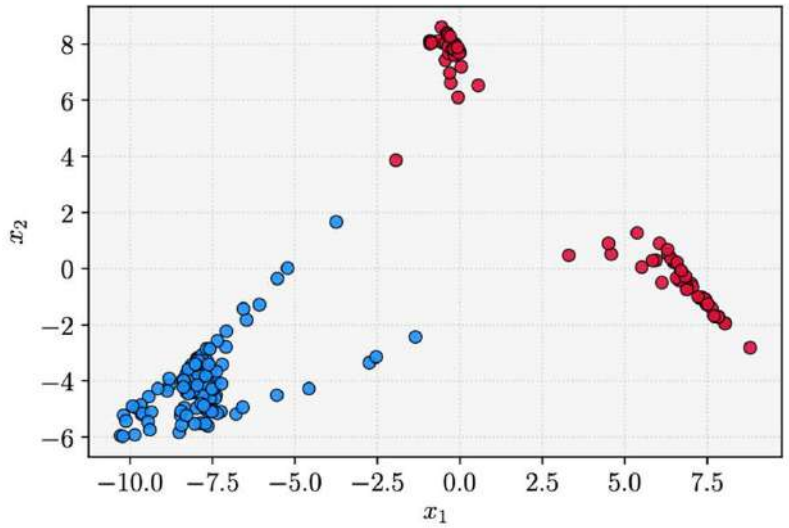
$$\min_{u\in H^k(0,T;\mathbb{R}^{d_u})} \int_0^T \mathcal{E}(\mathbf{x}(t))dt + \lambda \|u\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \quad \text{subject to (nODE)} \qquad \text{(SL*)}$$

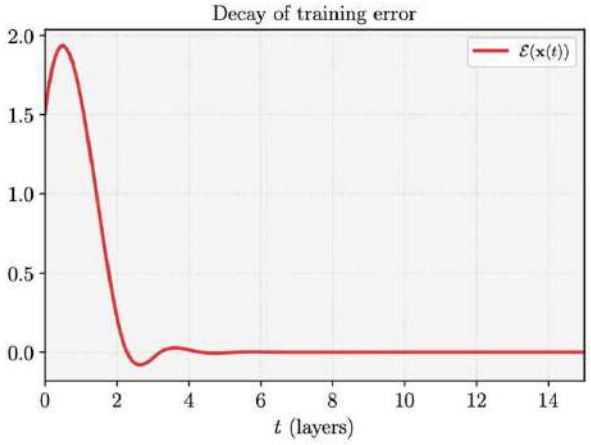
- For instance, if $\vec{y}_i \in \{1,\dots,m\}$, then $P\mathbf{x}_i(T) \in \mathbb{R}^m$ and consider cross-entropy loss

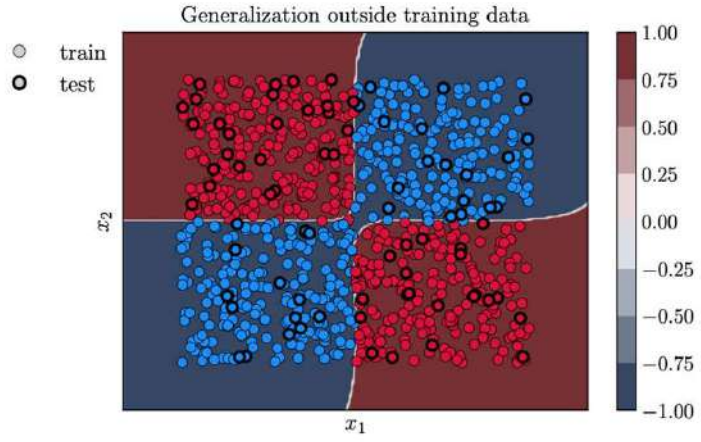
$$\mathrm{loss}(P\mathbf{x}_i(T), \vec{y}_i) := -\log\left(\frac{e^{-P\mathbf{x}_i(T)_{\vec{y}_i}}}{\sum_{j=1}^m e^{-P\mathbf{x}_i(T)_j}}\right).$$

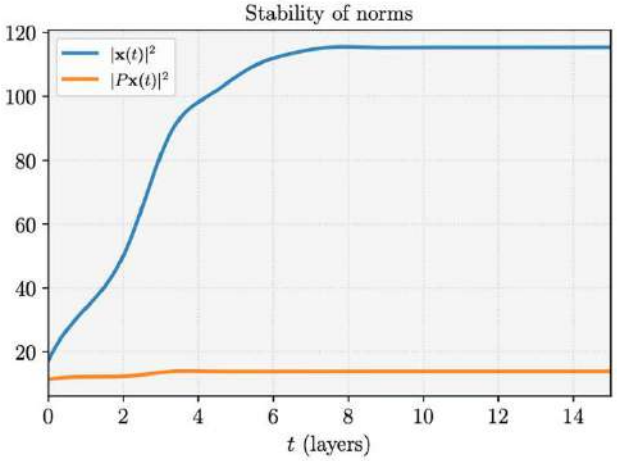
- Do we still have stabilizatiion/turnpike?

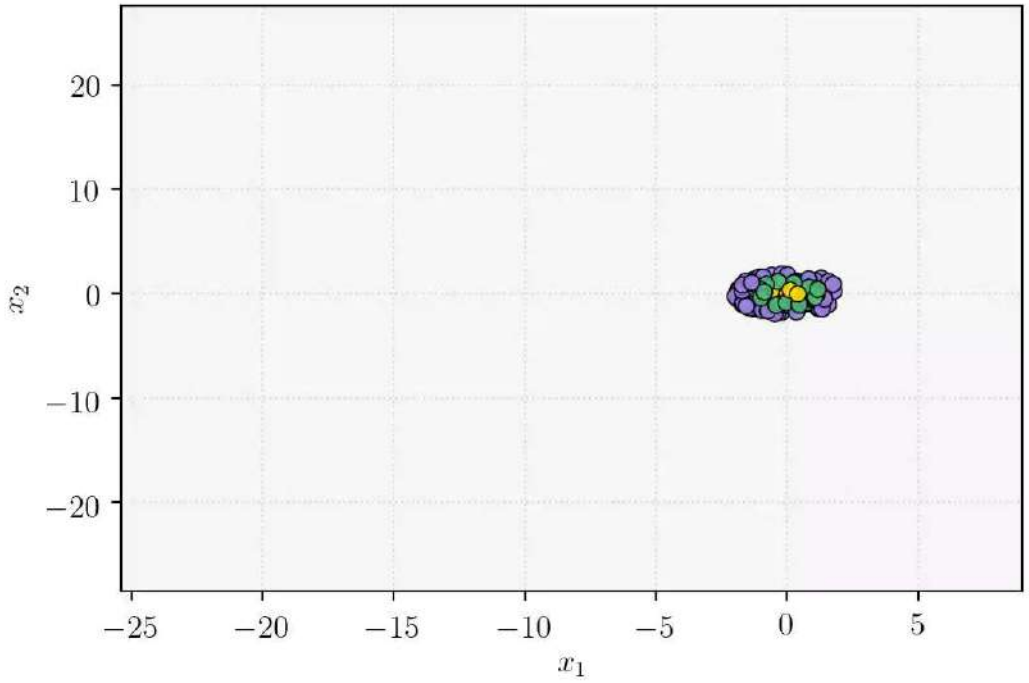

10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
x_2
−10
−5
0
5
10
x_1

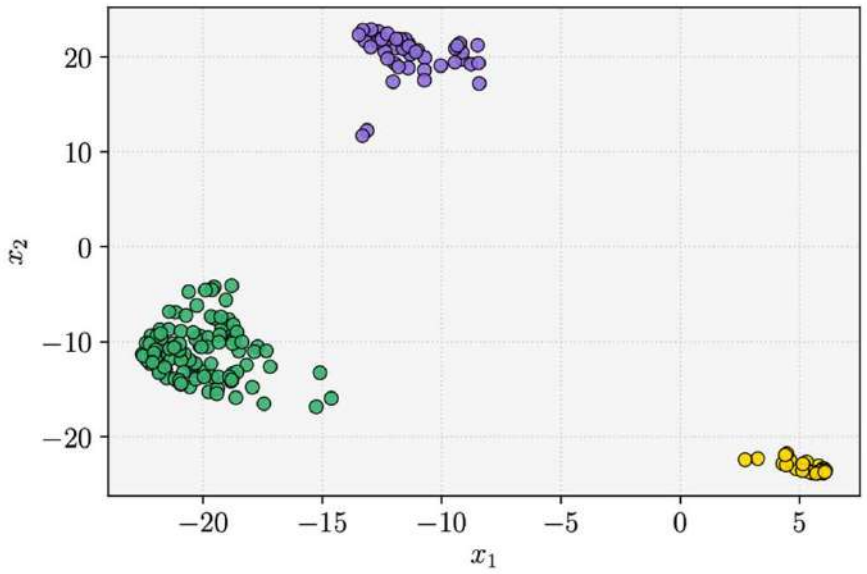

8
6
4
2
0
−2
−4
−6
x_2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
x_1

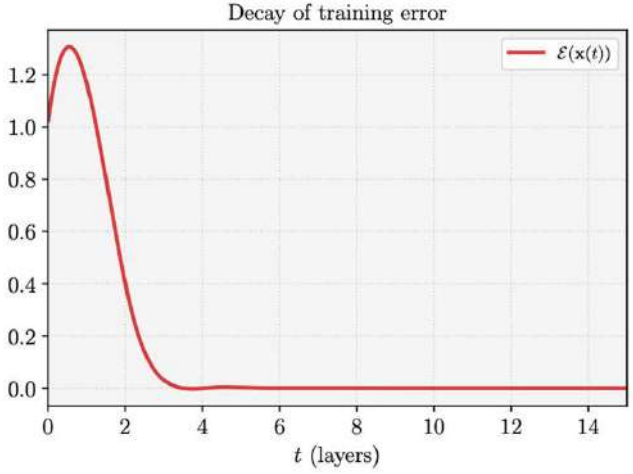

Decay of training error
$\mathcal{E}(\mathbf{x}(t))$
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
14
t (layers)

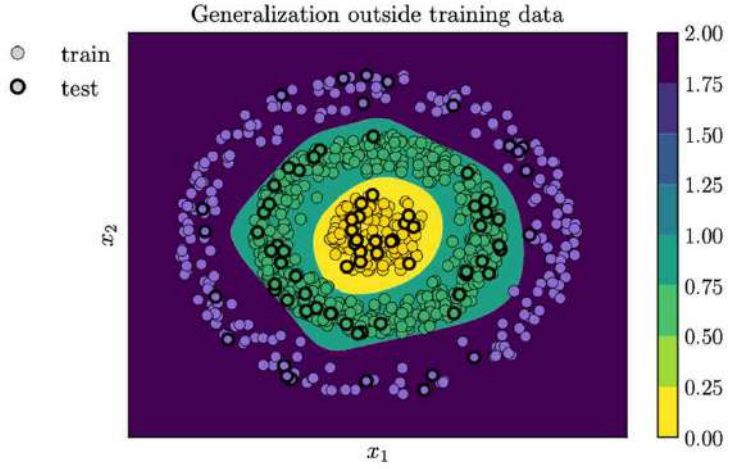

Generalization outside training data
train
test
x_2
x_1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

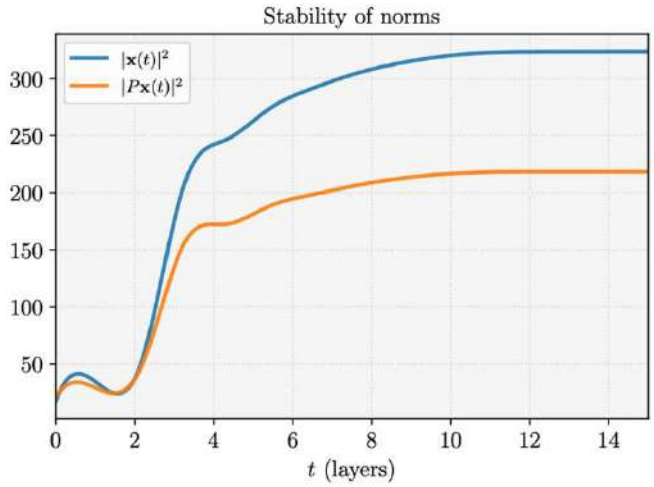

Stability of norms
$|\mathbf{x}(t)|^2$
$|P\mathbf{x}(t)|^2$
120
100
80
60
40
20
0
2
4
6
8
10
12
14
t (layers)


20
10
0
−10
−20
x_2
−25
−20
−15
−10
−5
0
5
x_1


20
10
0
−10
−20
x_2
−20
−15
−10
−5
0
5
x_1


Decay of training error
$\mathcal{E}(\mathbf{x}(t))$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
t (layers)


Generalization outside training data
train
test
x_2
x_1
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


Stability of norms
$|\mathbf{x}(t)|^2$
$|P\mathbf{x}(t)|^2$
300
250
200
150
100
50
0
2
4
6
8
10
12
14
t (layers)

Turnpike property

- Theorem is a special manifestation of the well-known **turnpike property** in optimal control and economics.
- *For suitable optimal control problems in a sufficiently large T, any optimal solution $(u^T, \mathbf{x}^T)$ remains, during most of the time, $\mathcal{O}(e^{-t} + e^{-(T-t)})$–close to the optimal solution of a corresponding "static" optimal control problem.*
  Optimal static solution is referred to as the turnpike – the name stems from the idea that a turnpike is the fastest route between two points which are far apart, even if it is not the most direct route.
- Since $\mathbf{f}(\mathbf{x}, 0) = 0$ for all $\mathbf{x}$, $\bar{\mathbf{x}}_i$ may be seen as the turnpike for $P\mathbf{x}_i$. Since this is a steady state, we do not see an exit from the turnpike and we stabilize.

Proof of $\mathcal{E}(\mathbf{x}^T(t)) + \|\mathbf{x}^T(t) - \bar{\mathbf{x}}\| \lesssim e^{-t}$

$k = 0$, $N = 1$ for simplicity. Then $\mathcal{E}(\mathbf{x}^T(t)) = \left\|P\mathbf{x}^T(t) - \vec{y}\right\|^2$.

Part 1). For $T \geq 1$, we first prove that

$$\left\|\mathbf{x}^T(t) - \bar{\mathbf{x}}\right\|^2 + \left\|\mathbf{x}^T - \bar{\mathbf{x}}\right\|^2_{L^2(0,T)} + \left\|u^T\right\|^2_{L^2(0,T)} \lesssim_\sigma \left\|\mathbf{x}^0 - \bar{\mathbf{x}}\right\|^2 \tag{7}$$

for all $t \in [0, T]$ uniformly in T.

1. $\exists u^* \in L^2(0,1)$ such that $\mathbf{x}^*(1) = \bar{\mathbf{x}}$ and $\|u^*\|_{L^2} \lesssim \left\|\mathbf{x}^0 - \bar{\mathbf{x}}\right\|$.
2. Grönwall: $\|\mathbf{x}^*(t) - \bar{\mathbf{x}}\| \lesssim_\sigma \left\|\mathbf{x}^0 - \bar{\mathbf{x}}\right\|$
3. Set
$$u^{\text{aux}}(t) := \begin{cases} u^*(t) & \text{for } t \in (0,1) \\ 0 & \text{for } t \in (1,T). \end{cases}$$
Then $\mathbf{x}^{\text{aux}}(t) = \bar{\mathbf{x}}$ for $t \in [1, T]$.
4. u^T minimizer, so
$$\begin{aligned} \left\|\mathbf{x}^T - \bar{\mathbf{x}}\right\|^2_{L^2(0,T)} + \left\|u^T\right\|^2_{L^2(0,T)} &\leq \|\mathbf{x}^* - \bar{\mathbf{x}}\|^2_{L^2(0,1)} + \|u^*\|^2_{L^2(0,1)} \\ &\lesssim_\sigma \left\|\mathbf{x}^0 - \bar{\mathbf{x}}\right\|. \end{aligned}$$
5. Conclude by Grönwall.

Part 2). Fix $\tau > C_\sigma^4 + 1$, and let $T \geq 2\tau + 1$.

1. For $t \in [0, \tau + 1]$, desired estimate follows from (7):

$$\left\| \mathbf{x}^T(t) - \bar{\mathbf{x}} \right\| \lesssim_\sigma \|\mathbf{x}^0 - \bar{\mathbf{x}}\| \lesssim_{\sigma,\tau} e^{-t} \|\mathbf{x}^0 - \bar{\mathbf{x}}\|.$$

2. (7) + contradiction argument:

$$\|\mathbf{x}^T(t) - \vec{y}\| \leq \frac{C_\sigma^2}{\sqrt{\tau}} \|\vec{x} - \vec{y}\|$$

for $t \in [\tau, T]$.

3. Bootstrap: for $n \leq \frac{T}{2\tau}$

$$\|\mathbf{x}^T(t) - \bar{\mathbf{x}}\| \leq \left(\frac{C_\sigma^2}{\sqrt{\tau}} \right)^n \|\mathbf{x}^0 - \bar{\mathbf{x}}\|$$

for $t \in [n\tau, T]$.

4. Suppose $t \in [\tau + 1, T]$. Set $n(t) = \lfloor \frac{t}{\tau+1} \rfloor$. Then $t \in [n(t)\tau, T]$, so

$$\begin{aligned} \left\| \mathbf{x}^T(t) - \bar{\mathbf{x}} \right\| &\leq \exp\left(-n(t) \log\left(\frac{\sqrt{\tau}}{C_\sigma^4} \right) \right) \|\mathbf{x}^0 - \bar{\mathbf{x}}\| \\ &\lesssim_{\tau,\sigma} \exp\left(-\frac{\log\left(\frac{\sqrt{\tau}}{C_\sigma^4} \right)}{\tau + 1} t \right) \|\mathbf{x}^0 - \bar{\mathbf{x}}\|. \end{aligned}$$

Proof of $\left\|u^T(t)\right\| \lesssim e^{-t}$

Let $t \in [0, T)$ and $0 < h \ll 1$ s.t. $t + 2h \in [0, T]$.

1. Set

$$u^{\text{aux}}(s) := \begin{cases} u^T(s) & \text{for } s \in (0, t) \\ \frac{1}{2} u^T\left(t + \frac{s-t}{2}\right) & \text{for } s \in (t, t+2h) \\ u^T(s-h) & \text{for } s \in (t+2h, T). \end{cases}$$

2. Since u^T minimizer, by $J_T(u^T) \leq J_T(u^{\text{aux}})$, we will find

$$\frac{1}{2} \int_t^{t+h} \left\|u^T(s)\right\|^2 ds \leq \int_t^{t+h} \left\|x^T(s) - \vec{y}\right\|^2 ds.$$

3. Combined with $\left\|x^T(s) - \vec{y}\right\|^2 \lesssim e^{-t}$,

$$\int_t^{t+h} \left\|u^T(s)\right\|^2 ds \lesssim \int_t^{t+h} e^{-s} ds \lesssim h e^{-t}$$

4. Lebesgue differentiation theorem: for a.e. $t \in [0, T]$,

$$\left\|u^T(t)\right\|^2 = \lim_{h \to 0^+} \frac{1}{h} \int_t^{t+h} \left\|u^T(s)\right\|^2 ds \lesssim e^{-t}.$$

Extensions

Variable width

Variable width ResNets via **integro-differential equation**: for $i \leq N$

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

- e.g. $\Omega = \text{image} \times (0,1) \subset \mathbb{R}^3$;
- All previous asymptotics theorems apply here;
- Variable width ResNets can be obtained by semi-discretizing via time-dependent mesh.

Switched systems: Changing widths over layers as switched systems over time:

$$\dot{x}(t) = f_{\rho(t)}(x(t), u(t))$$

given M vector fields $f_1, \ldots, f_M$ and switching signal $\rho : [0, T] \to \{1, \ldots, M\}$;

Quasi-turnpike strategy:

#1 increase the dimension to the "optimal system" f_{j^*},

#2 use the stabilization/turnpike for fixed width

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

Outlook

Open problems..

- Asymptotics remain to be proven when $P : \mathbb{R}^d \to \mathbb{R}^m$ is optimizable variable
- Proof of turnpike for functional integrating $\mathcal{E}(\mathbf{x}(t))$ over $[0, T]$
- Statistical complexity bounds for asymptotic limits?

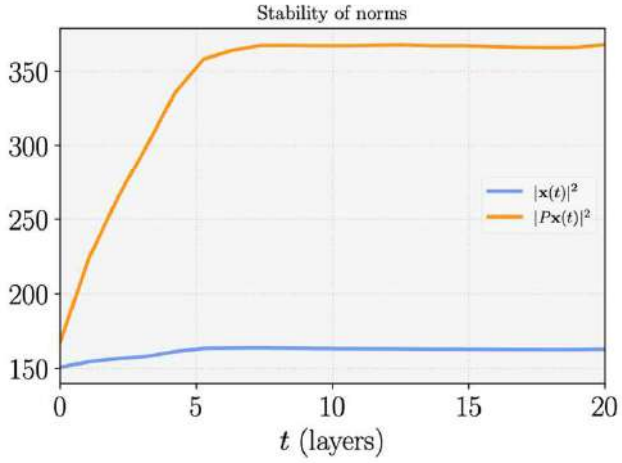
Extensive bibliography can be found in

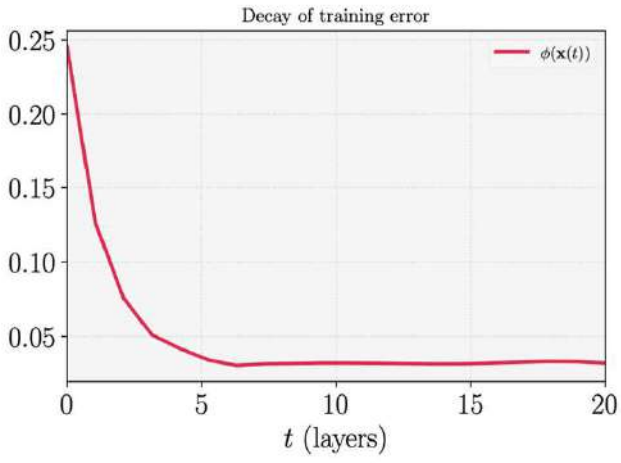
LARGE-TIME ASYMPTOTICS IN DEEP LEARNING

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

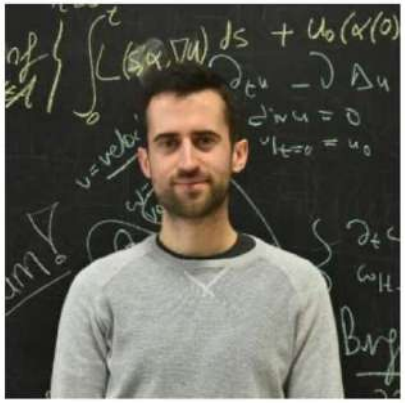
https://arxiv.org/abs/2008.02491





Thank you for your attention!

Collaborators:



- C. Esteve (UAM/Deusto), D. Pighin (PhD @ UAM, 2020), E. Zuazua (FAU/ Deusto/UAM).

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.

Alexander von Humboldt Stiftung/Foundation

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

DyCon DYNAMIC CONTROL erc

L^1–regularization

Theorem (Esteve, G., Pighin, Zuazua, '20): Fix $M > 0$ and assume $\{\mathcal{E} = 0\} \neq \varnothing$. Suppose ($\text{nODE}_2$) is controllable. Consider

$$\inf_{\substack{u \in L^1(0,T;\mathbb{R}^{d_u}) \\ \text{esssup}\|u\| \leq M \\ \text{subject to } (\text{nODE}_2)}} \int_0^T \mathcal{E}(\mathbf{x}(t))dt + \lambda \|u\|_{L^1(0,T;\mathbb{R}^{d_u})}.$$

Then there exists $T_M > 0$ such that for any $T > T_M$, any optimal u^T and corresponding state $\mathbf{x}^T$, unique solution to (nODE_2), satisfy

$$\mathcal{E}(\mathbf{x}^T(t)) = 0, \qquad \text{for all } \; t \in [T^*, T]$$

and

$$\begin{aligned} \left\| u^T(t) \right\| &= M, \qquad \text{for a.e. } t \in (0, T^*) \\ \left\| u^T(t) \right\| &= 0, \qquad \text{for a.e. } t \in (T^*, T). \end{aligned}$$

for some $0 < T^* \leq T_M$.

Controllability

Theorem (Esteve et al. '20): Let $T > 0$ and assume that $N \leq d$. Fix $\mathbf{x}^1 \in \mathbb{R}^{d_x}$ and assume that $\sigma \in C^1(\mathbb{R})$ is such that

$$\left\{\sigma\left(\mathbf{x}_1^1\right), \ldots, \sigma\left(\mathbf{x}_i^1\right), \ldots, \sigma\left(\mathbf{x}_N^1\right)\right\}$$

is a system of linearly independent vectors in $\mathbb{R}^d$.
There exists $r > 0$ such that for any $\mathbf{x}^0 \in \mathbb{R}^{d_x}$ satisfying $\|\mathbf{x}^0 - \mathbf{x}^1\| \leq r$, there exists weights $w \in L^\infty(0, T; \mathbb{R}^{d^2})$ s.t. the solution $\mathbf{x}$ to

$$\begin{cases} \dot{\mathbf{x}}(t) = \operatorname{diag}(w(t))\sigma(\mathbf{x}(t)) & \text{in } (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases}$$

satisfies

$$\mathbf{x}(T) = \mathbf{x}^1,$$

and the estimate

$$\|w\|_{L^\infty(0,T;\mathbb{R}^{d^2})} \leq \frac{C}{T}\left\|\mathbf{x}^0 - \mathbf{x}^1\right\|,$$

holds for some $C > 0$ independent of T.