Large-time asymptotics in Deep Learning

Borjan Geshkovski

Seminario de Estadística, UAM
October 23rd, 2020

Supervised learning

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).
- High-dimensional interpolation problem.

Classification

Suppose $\vec{x}_i \in \mathbb{R}^2$ and $\vec{y}_i \in \{-1, 1\}$ for $i \leq N$.

- A simple idea:

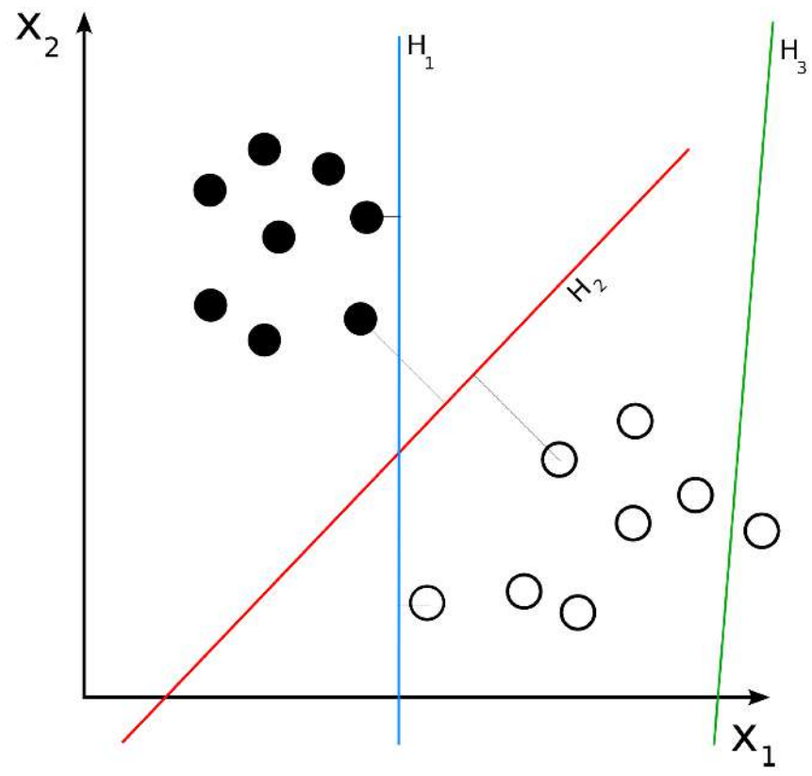
$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \left\| w^\top \vec{x}_i - \vec{y}_i \right\|^2 + \|w\|_{\ell p}^p$$

$p = 1, 2$ and set

$$f(x) = \begin{cases} 1 & \text{if } \ w^\top x > 0 \\ -1 & \text{else.} \end{cases}$$

$\longrightarrow$ discriminative model (other ex: Support Vector Machine)

- Compared to generative models (e.g. Bayes classifier)
- But data is not linearly separable in general.



Neural networks

Neural network: for any $i \leq N$

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

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

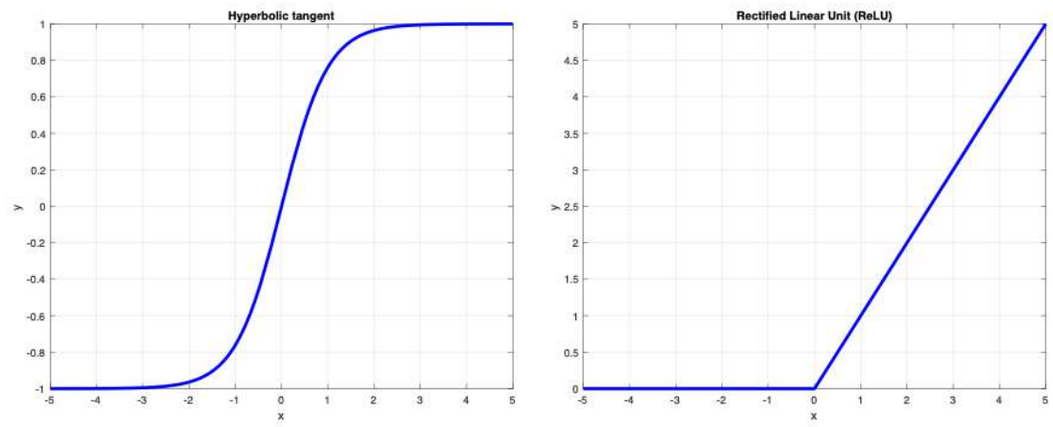


Figure: Sigmoid: $\tanh(x)$ and ReLU: $\max\{x, 0\}$

- ML jargon: multilayer perceptron / fully-connected.

"Training" a NN

Training $\Longleftrightarrow$ Optimization:

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

- $\text{loss}(x, y) = \|x - y\|_{\ell^p}^p$ for $p = 1, 2$ or $\text{loss}(x, y) = \log(1 + e^{x \cdot y})$;
- $\lambda > 0$ fixed.

Comments:

- $x \mapsto w^{N_{layers}} x^{N_{layers}}$ *is the candidate approximation of unknown* f
- $w^{N_{layers}} \mathbf{x}_i^{N_{layers}}$ *can be replaced by* $\varphi\left(w^{N_{layers}} \mathbf{x}_i^{N_{layers}}\right)$ *where* $\varphi : \mathbb{R}^m \to \mathbb{R}^m$ *is nonlinear (softmax).*
- *In practice solved using stochastic gradient descent + backpropagation.*

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

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

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

[1] Weinan E '17

Residual neural networks

- In addition to (NN_1), one can also consider variants:

$$\begin{cases} \mathbf{x}_i^{k+1} = w^k \sigma(\mathbf{x}_i^k) + b^k & \text{for } \ k \in \{0, \dots, N_{layers} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i. \end{cases} \tag{NN_2}$$

$\longrightarrow$ motivates considering

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w(t)\sigma(\mathbf{x}_i(t)) + b(t) & \text{for } \ t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i. \end{cases} \tag{nODE_2}$$

Training is optimal control

Henceforth fix $P: \mathbb{R}^d \to \mathbb{R}^m$ s.t. $P^{-1}(\{\vec{y}_i\}) \neq \varnothing$ for all $i \leq N$.
Given $T, \lambda > 0$:

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

- $k = 0$ for (nODE_2), $k = 1$ for (nODE) (L^2–regularization **may not be enough** for compactness $\longrightarrow$ enhance to Sobolev regularization)
- $d_u := d^2 + d$

Henceforth denote the training error by

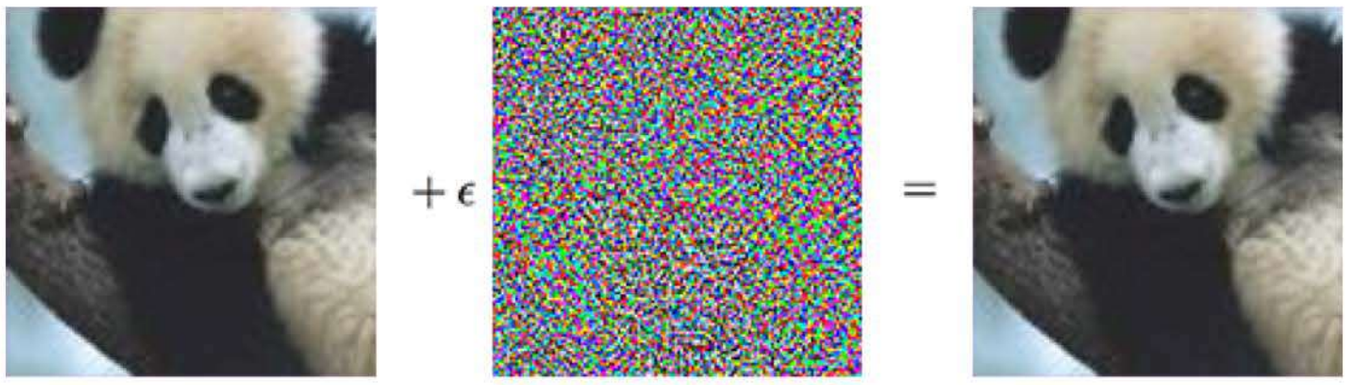
$$\phi(\mathbf{x}(T)) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{loss}\big(P\mathbf{x}_i(T), \vec{y}_i\big)$$

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)



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 $\mathbf{x}^0 = [\vec{x}_1, \dots, \vec{x}_N]^\top$, $u = [w, b]^\top$, and put both (nODE) 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

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

- $T \gg 1$ would imply that $N_{layers} \gg 1 \longrightarrow$ *deep learning regime*. So,

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

Qualitative results

Scaling

Key idea: *Time-Scaling*.

- Assumptions on $\sigma \longrightarrow \mathbf{f}(\mathbf{x}, u)$ is 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

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

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

Then:

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

Theorem (Esteve, G., Pighin, Zuazua, '20): Fix $\lambda > 0$ and suppose $\{\phi = 0\} \neq \varnothing$. For any $T > 0$, let u^T be minimizer in (SL_1), $\mathbf{x}^T$ associated solution to (nODE). Assume that (nODE) is controllable. Then

$$\phi(\mathbf{x}^T(T)) \longrightarrow 0 \qquad \text{as} \quad T \to +\infty.$$

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

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

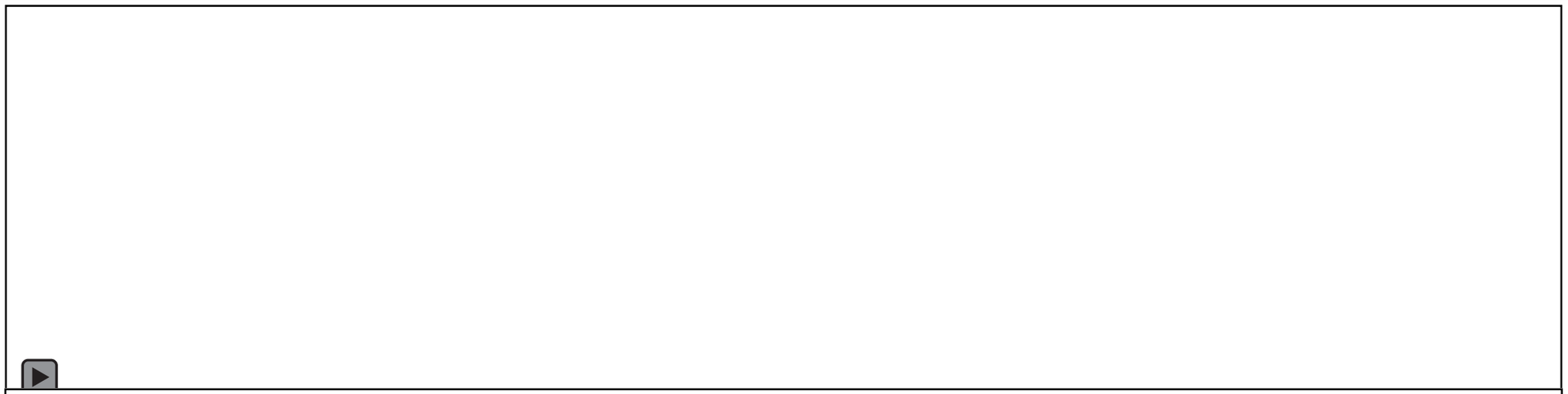
Moreover

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

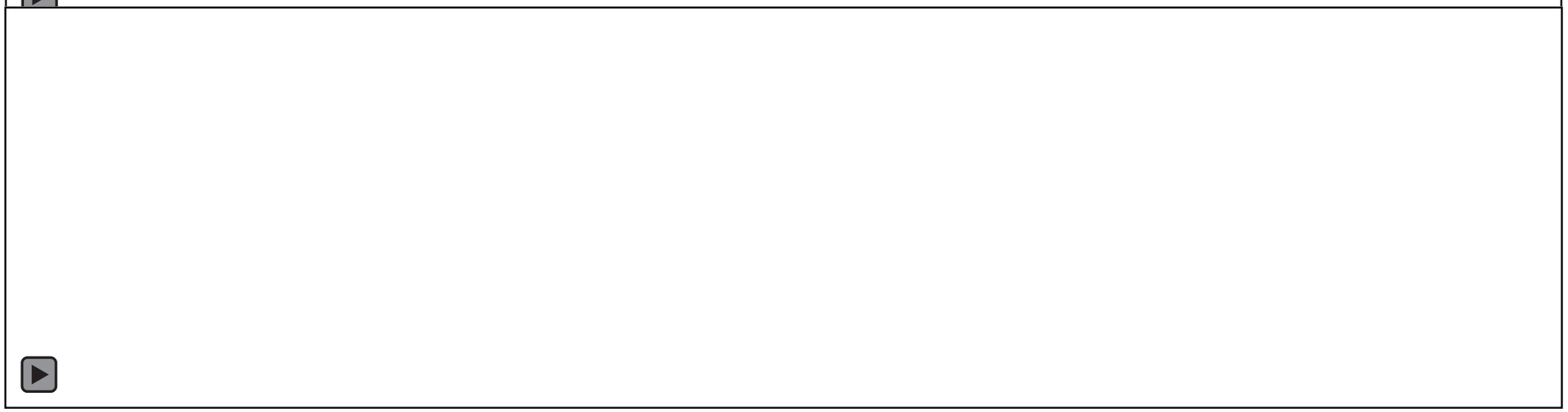
where u^* solves

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

T=16



T=36



T=81

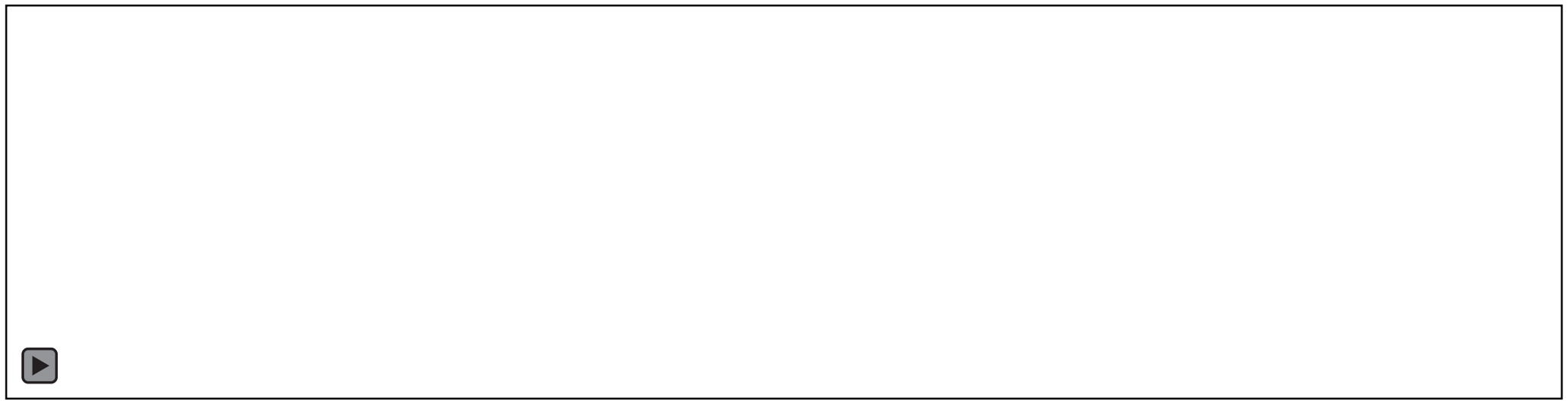


Figure: Here $N_{\text{layers}} = \left\lfloor T^{\frac{3}{2}} \right\rfloor$ and thus $h = \frac{1}{\sqrt{T}}$, and we consider $\alpha = 1$.

$$T \to \infty \Longleftrightarrow \lambda \to 0$$

Back to

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

Corollary: Fix $T > 0$. Under the assumptions of the Theorem,

$$\phi(\mathbf{x}^T(T)) \longrightarrow 0 \qquad \text{as} \quad \lambda \to 0.$$

Moreover

$$\left\|u^T\right\|_{H^k(0,T;\mathbb{R}^{d_u})} \longrightarrow \|u^*\|_{H^k(0,1;\mathbb{R}^{d_u})} \qquad \text{as} \quad \lambda \to 0$$

where u^* solves

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

Discussion

- Back to

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

where e.g. $\text{loss}(x, y) = \log(1 + e^{x\cdot y})$; shown[2] that

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

where w^* is maximum margin separator:

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

- Compared to other convergence results of generalization nature: **implicit bias property of gradient descent**[3]:

  "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."

- We clearly exhibit the explicit L^2–regularization of the control parameters.

---

[2] Rosset, Hu, Hastie '04

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

Proof of Theorem

For simplicity, suppose $k = 0$.

Part 1). We first show that

$$\phi(\mathbf{x}^T(T)) \longrightarrow 0 \qquad \text{as } T \to \infty.$$

Recall that $\min_{\mathbb{R}^{d_x}} \phi = 0$.

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

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

3. Since $\phi(\mathbf{x}^1(1)) = 0$,

$$0 \leq \phi\left(\mathbf{x}^T(T)\right) \leq \frac{\lambda}{T} \left\|u^1\right\|^2_{L^2(0,1)}.$$

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

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

1. Integral formulation of ODE + Grönwall + Cauchy Schwarz + 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 $\exists \mathbf{x}^\dagger \in \mathbb{R}^{d_x}$ such that

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

3. By **Part 1)**, $\phi\left(\mathbf{x}^{T_n}(T_n)\right) \longrightarrow 0$. By lower semicontinuity of ϕ,

$$\phi(\mathbf{x}^\dagger) \leq \liminf_{n \to +\infty} \phi\left(\mathbf{x}^{T_n}(T_n) \right) = 0.$$

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

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

where u^* solves

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

1. Let $u^0 \in L^2(0,1)$ solution to above minimization problem;
2. By contradiction: $\|u_n\|_{L^2(0,1)} \leq \|u^0\|_{L^2(0,1)}$ for every $n \geq 1$;
3. Banach-Alaoglu: $\exists u^* \in L^2(0,1)$ such that

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

   $\mathbf{x}^*$ trajectory associated to u^*; compactness properties of ODE:

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

4. But $\mathbf{x}^{T_n}(T_n) = \mathbf{x}_n(1)$ thus $\mathbf{x}^*(1) = \mathbf{x}^\dagger$ by **Part 1)**, so $\phi(\mathbf{x}^*(1)) = 0$.
5. 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.

Quantitative results

Setting

Question: Quantitative estimates for the time T required to reach the zero training error regime $\phi(\mathbf{x}(T)) = 0$?

- We might need to change the cost function!
- Consider $\text{loss}(x, y) = \|x - y\|^2$ and so we recall

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

- Modified supervised learning problem:

$$\min_{u \in H^k(0,T;\mathbb{R}^{d_u})} \left[J_T(u) := \int_0^T \phi(\mathbf{x}(t))dt + \lambda \|u\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \right] \quad \text{(SL*)}$$

subject to (nODE)

Exponential stabilization

Theorem (Esteve, G., Pighin, Zuazua '20): Fix $\lambda > 0$ and suppose that (nODE) is controllable. There exist $T^* > 0$ such that for any $T \geq T^*$, any solution $(u^T, \mathbf{x}^T)$ to (SL*)–(nODE) satisfies

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

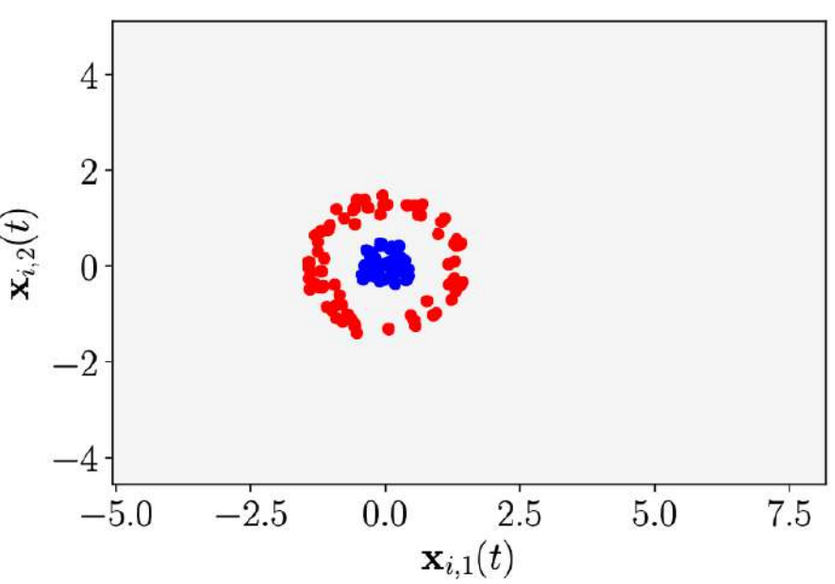
and

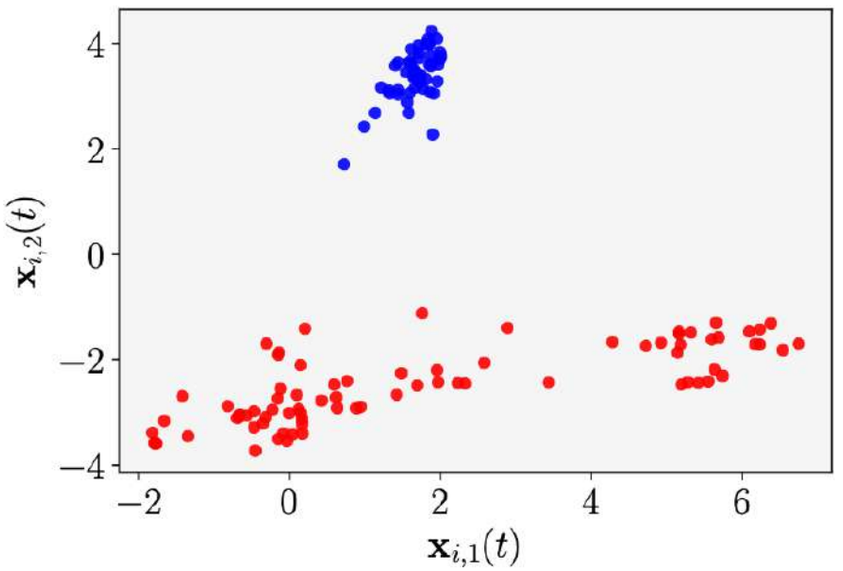
$$\left\|u^T(t)\right\| \leq C_2 e^{-\mu t} \qquad \text{for a.e. } t \in [0, T]$$

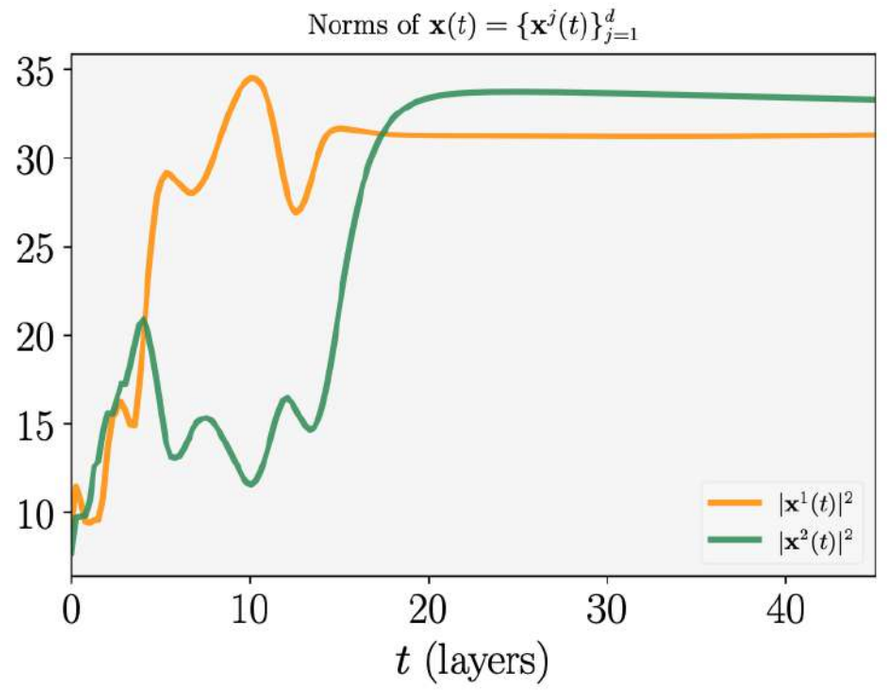
for some $C_1, C_2, \mu > 0$, all independent of T.

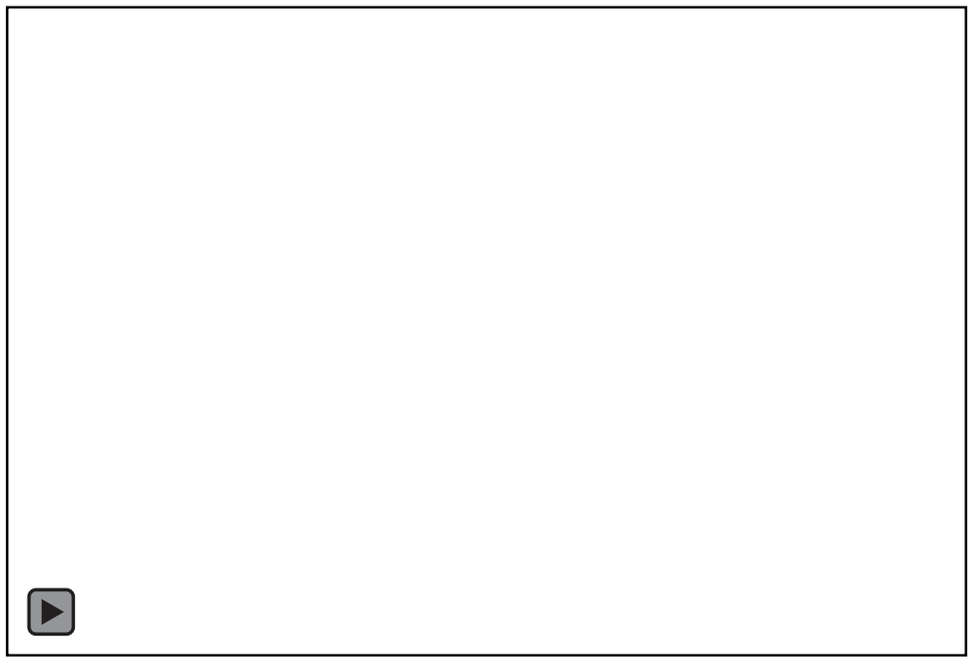
Remarks:

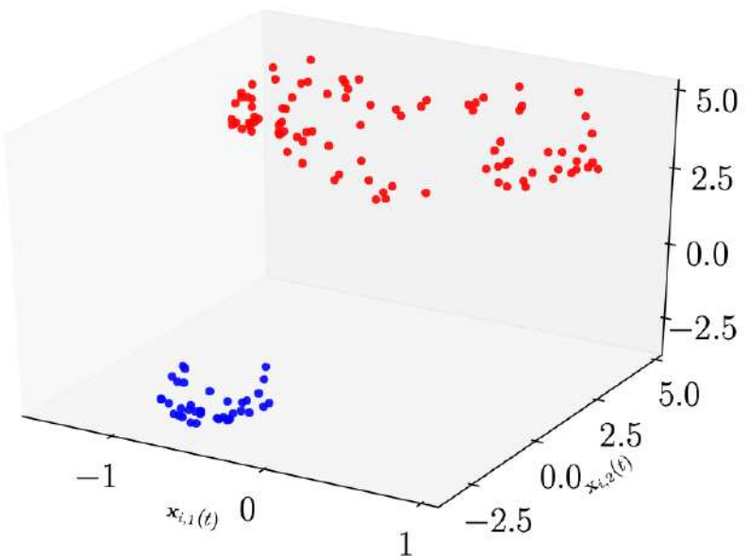
- Akin to *universal approximation*: given tolerance $\varepsilon > 0$, there exists $T_\varepsilon > 0$ (number of layers) and control parameters u^ε such that the neural network output is ε–close to the desired target.
- The difference with universal approximation is that our parameters may be computed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

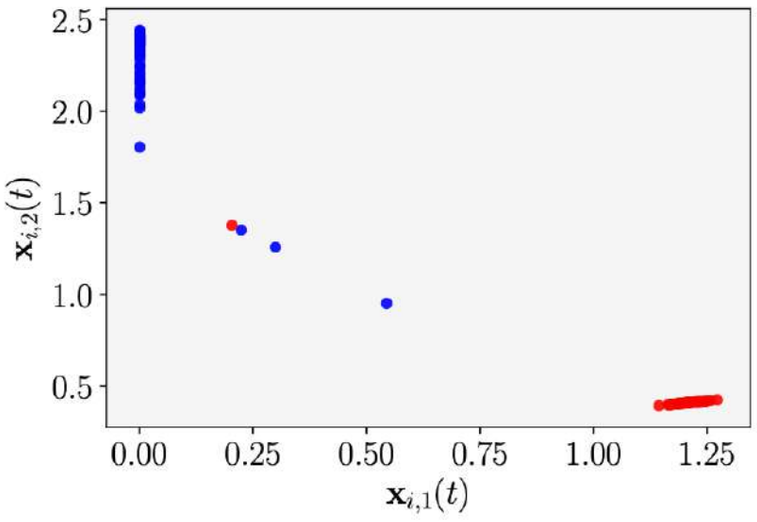

$\mathbf{x}_{i,2}(t)$
$\mathbf{x}_{i,1}(t)$

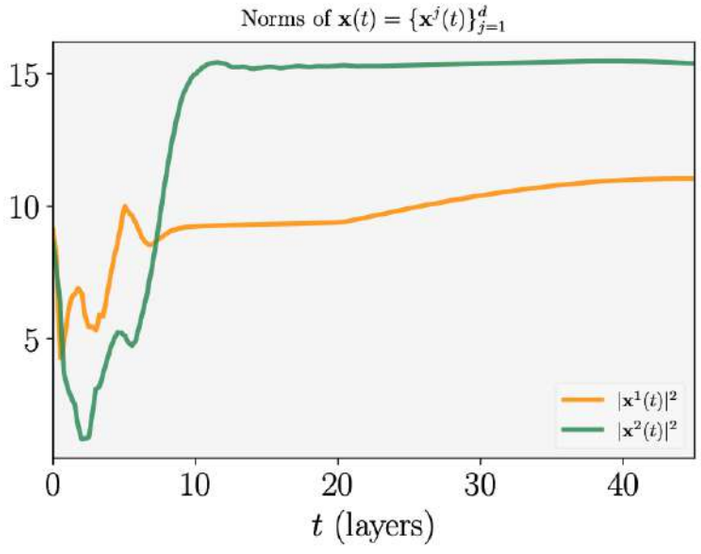

$\mathbf{x}_{i,2}(t)$
$\mathbf{x}_{i,1}(t)$


Norms of $\mathbf{x}(t) = \{\mathbf{x}^j(t)\}_{j=1}^d$
$|\mathbf{x}^1(t)|^2$
$|\mathbf{x}^2(t)|^2$
t (layers)




5.0
2.5
0.0
−2.5
5.0
2.5
0.0
−2.5
−1
0
1
$x_{i,1}(t)$
$x_{i,2}(t)$


2.5
2.0
1.5
1.0
0.5
0.00
0.25
0.50
0.75
1.00
1.25
$\mathbf{x}_{i,2}(t)$
$\mathbf{x}_{i,1}(t)$


Norms of $\mathbf{x}(t) = \{\mathbf{x}^j(t)\}_{j=1}^d$
15
10
5
0
10
20
30
40
t (layers)
$|\mathbf{x}^1(t)|^2$
$|\mathbf{x}^2(t)|^2$

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 $\mathbf{x}$, $\overline{y}_i$ 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 $\phi(\mathbf{x}^T(t)) \lesssim e^{-t}$

Suppose $k = 0$, $d = m$ and $N = 1$ for simplicity. Then $\phi(\mathbf{x}^T(t)) = \left\|x^T(t) - \vec{y}\right\|^2$.

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

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

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

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

Part 2). Fix $\tau > C_\sigma^4 + 1$ ($C_\sigma > 0$ appears in (2)) and let $T \geq 2\tau + 1$.

1. For $t \in [0, \tau + 1]$, desired estimate follows from (2):

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

2. (2) + 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 $n \leq \frac{T}{2\tau}$

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

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

4. Suppose $t \in [\tau + 1, T]$. Set $n(t) = \lfloor \frac{t}{\tau+1} \rfloor$. Then $n(t) \leq \frac{T}{2\tau}$, $t \in [n(t)\tau, T]$ and $n(t) \geq \frac{t}{\tau+1} - 1$, so

$$\left\| \mathbf{x}^T(t) - \vec{y} \right\| \leq \exp\left(-n(t) \log\left(\frac{\sqrt{\tau}}{C_\sigma^4} \right) \right) \|\vec{x} - \vec{y}\|$$

$$\lesssim_{\tau,\sigma} \exp\left(-\frac{\log\left(\frac{\sqrt{\tau}}{C_\sigma^4} \right)}{\tau + 1} t \right) \|\vec{x} - \vec{y}\|.$$

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. $t \in [0, T]$,

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

L^1–regularization

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

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

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

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

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 $0 < T^* \leq T_M$.

Proof

1. **Part 1:** Show that for

$$T^* := \min\left\{ t \in [0, T] : \ \phi\left(\mathbf{x}^T(t)\right) = \min_{s \in [0,T]} \phi\left(\mathbf{x}^T(s)\right) \right\}$$

it holds

$$\phi\left(\mathbf{x}^T(t)\right) > \phi\left(\mathbf{x}^T(T^*)\right) \quad \text{for all } t \in [0, T^*),$$
$$\text{and} \quad \phi\left(\mathbf{x}^T(t)\right) = \phi\left(\mathbf{x}^T(T^*)\right) \quad \text{for all } t \in [T^*, T].$$

2. **Part 2:** Show that the optimal control parameters u^T satisfy

$$\left\| u^T(t) \right\| = M \quad \text{for a.e. } t \in (0, T'),$$
$$\text{and} \quad \left\| u^T(t) \right\| = 0 \quad \text{for a.e. } t \in (T', T).$$

This relies on the fact that the L^1–norm is invariant to the time-scaling and Lebesgue measure theory.

3. **Part 3:** Show that there exists $T_M > 0$ such that if $T > T_M$, then

$$T^* \leq T_M \quad \text{and} \quad \phi(\mathbf{x}^T(T^*)) = 0.$$

Zero training error regime

Recall:

- We are given a training dataset $\{\vec{x}_i, \vec{y}_i\}$;
- the training error $\phi \in C^0(\mathbb{R}^{d\times N}; \mathbb{R}_+)$ is defined by

$$\phi(x) = \frac{1}{N} \sum_{i=1}^{N} \text{loss}(Px_i, \vec{y}_i)$$

Our results stipulate: **when depth N_{layers} is increased, trained trajectories of neural networks approach the zero training error regime $\phi = 0$.**

Question: Can we reach exactly the zero training error regime, and how big are the control parameters in this case?

A lower bound

Weights need to be at least of a certain size! $\longrightarrow$ depend on the way the dataset is "spread out".

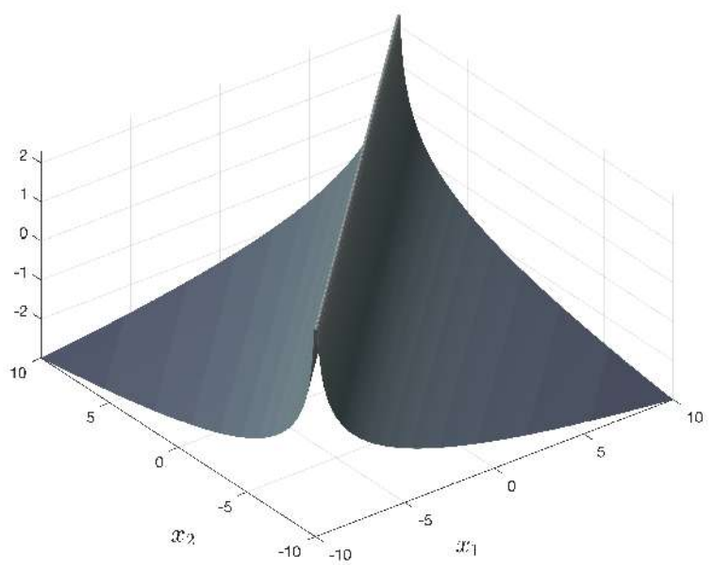
Theorem (Esteve et al. '20): Let $T > 0$. Assume that for some control parameters $u := [w, b]^\top$, the solutions $\mathbf{x}_i(t)$ to either (nODE) or (nODE_2) satisfies

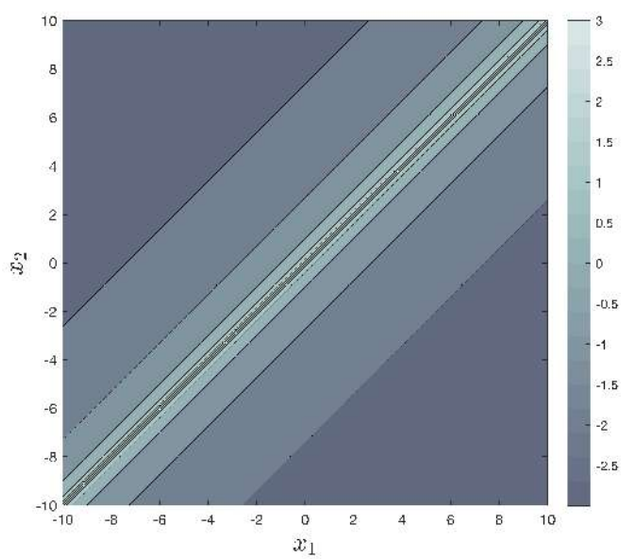
$$P\mathbf{x}_i(T) = \vec{y}_i \qquad \text{for all } i \in \{1, \dots, N\}.$$

Then

$$\|w\|_{L^1(0,T;\mathbb{R}^{d_u})} \geq L_\sigma \max_{\substack{(i,j)\in\{1,\dots,N\}^2 \\ i\neq j}} \inf_{\substack{\mathbf{x}_i^1 \in P^{-1}(\{\vec{y}_i\}) \\ \mathbf{x}_j^1 \in P^{-1}(\{\vec{y}_j\})}} \log\left(\frac{\left\|\mathbf{x}_i^1 - \mathbf{x}_j^1\right\|}{\left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\|}\right)$$

where $L_\sigma > 0$ is the Lipschitz constant of σ.





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

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

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.

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

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..)

Some additional problems:

- Asymptotics remain to be proven when $P : \mathbb{R}^d \to \mathbb{R}^m$ is optimizable variable?
- Stabilization for non L^2–loss with L^2–regularization?
- Further characterization of limit control u^* in first theorem?
- Extensive bibliography can be found in

LARGE-TIME ASYMPTOTICS IN DEEP LEARNING

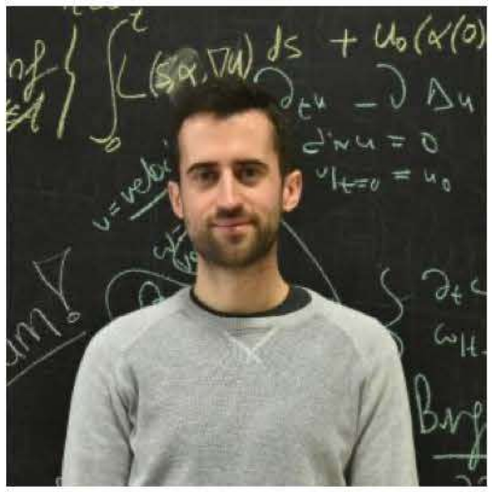
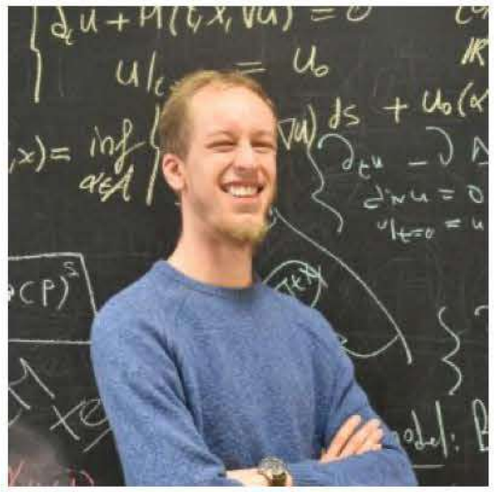
CARLOS ESTEVE, BORJAN GESHKOVSKI, DARIO PIGHIN, AND ENRIQUE ZUAZUA

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