Deep neural networks, universal approximation, and nonlinear geometric control.

Paulo Tabuada [1] and Bahman Gharesifard [2]

[1] Vijay K. Dhir Professor of Engineering
Cyber-Physical Systems Laboratory
Department of Electrical and Computer Engineering
University of California at Los Angeles

[2] Department of Electrical and Computer Engineering
University of California at Los Angeles

Autonomy and deep learning

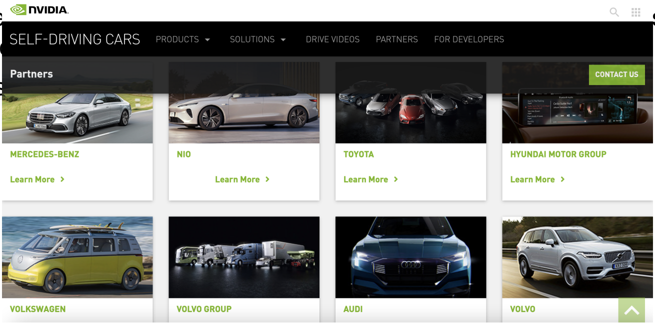
- The need to understand the world through vision and LiDAR data has inexorably linked deep learning and autonomy.

Autonomy and deep learning

- The need to understand the world through vision and LiDAR data has inexorably linked deep learning and autonomy.
- Here is one example from the automotive domain.

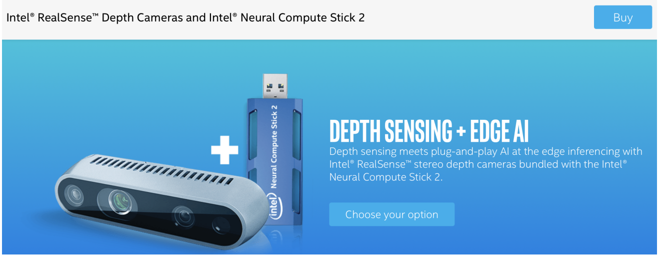
Autonomy and deep learning

- The ne… s inexorably linked …
- Here is…



Autonomy and deep learning

- The need to understand the world through vision and LiDAR data has inexorably linked deep learning and autonomy.
- Here is one example from the robotics domain.



Autonomy and deep learning

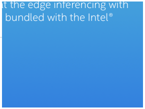
- The need to understand the world through vision and LiDAR data has inexorably linked deep learning and autonomy.
- Here is one example from the robotics domain.

Qualcomm Products Support Company

/ Products / Industrial & Commercial / Robotics / Robotics RB5 Platfor

Robotics RB5 Platform

Overview Documentation Software Hardware Support



The Qualcomm Robotics RB5 Platform supports the development of next generation of high-compute, AI-enabled, low power robots and drones for the consumer, enterprise, defense, industrial and professional service sectors that can be connected by 5G.

The platform's Qualcomm QRB5165 processor, customized for robotics applications, offers a powerful heterogeneous computing architecture coupled with the leading 5th generation Qualcomm® Artificial Intelligence (AI) Engine delivering 15 Trillion Operations Per Second (TOPS) of AI performance to efficiently run complex AI and deep learning workloads and on-device edge inferencing while using lower power , on device machine learning, and accurate edge inferencing. The processor also offers a powerful image signal processor (ISP) with support for seven concurrent cameras, a dedicated computer

Autonomy and deep learning

- Clearly, industry is ahead of academia.

Autonomy and deep learning

- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?

Autonomy and deep learning

- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?
- Today:
  1. What are deep residual neural networks (ResNets)?
  2. Control system models of ResNets.

Autonomy and deep learning

- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?
- Today:
  1. What are deep residual neural networks (ResNets)?
  2. Control system models of ResNets.
  3. Memorization capability of ResNets.
     - Can ResNets memorize finitely many samples?

CyPhyLab UCLA

Autonomy and deep learning

- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?
- Today:
  1. What are deep residual neural networks (ResNets)?
  2. Control system models of ResNets.
  3. Memorization capability of ResNets.
     - Can ResNets memorize finitely many samples?
  4. Approximation capability of ResNets.
     - Can ResNets approximate any continuous function?

Autonomy and deep learning

- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?
- Today:
  1. What are deep residual neural networks (ResNets)?
  2. Control system models of ResNets.
  3. Memorization capability of ResNets.
     - Can ResNets memorize finitely many samples?
  4. Approximation capability of ResNets.
     - Can ResNets approximate any continuous function?
  5. A deterministic generalization bound.
     - Stability guarantees for feedback loops with deep ResNets in the perception pipeline.

Autonomy and deep learning

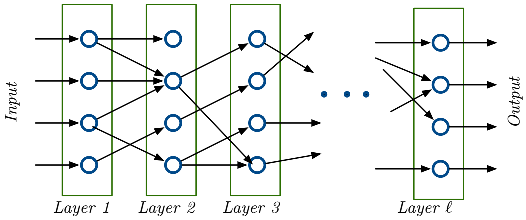
- Clearly, industry is ahead of academia.
- Formal guarantees when deep learning is used within a control loop?
- Today:

  1. What are deep residual neural networks (ResNets)?
  2. Control system models of ResNets.
  3. Memorization capability of ResNets.
     - Can ResNets memorize finitely many samples?
  4. Approximation capability of ResNets.
     - Can ResNets approximate any continuous function?
  5. A deterministic generalization bound.
     - Stability guarantees for feedback loops with deep ResNets in the perception pipeline.
  6. Outlook

Residual Neural Networks

What are ResNets?

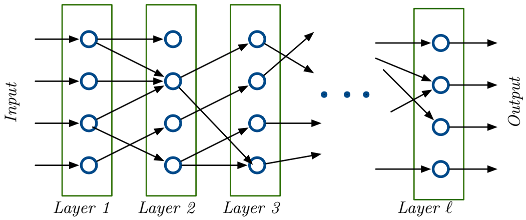
- A diagrammatic depiction of a neural network:



- Let us denote by $x(k) \in \mathbb{R}^4$ the state of each layer with $k = 1, 2, \ldots, \ell$.

What are ResNets?

- A diagrammatic depiction of a neural network:



- Let us denote by $x(k) \in \mathbb{R}^4$ the state of each layer with $k = 1, 2, \ldots, \ell$.
- The state of layer $k+1$ is computed from the state of layer k according to:

$$x(k+1) = \Sigma(W(k)x(k) + b(k)),$$

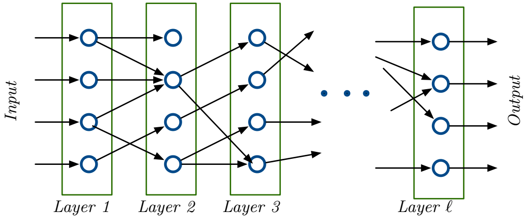
where (W, b) are the weights of the connections (arrows) and Σ is of the form:

$$\Sigma(x) = (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)),$$

for an activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

What are ResNets?

- A diagrammatic depiction of a neural network:



- Let us denote by $x(k) \in \mathbb{R}^4$ the state of each layer with $k = 1, 2, \ldots, \ell$.
- For ResNets, the state of layer $k+1$ is computed from the state of layer k according to:

$$x(k+1) = x(k) + S(k)\Sigma(W(k)x(k) + b(k)),$$

where (S, W, b) are the weights of the connections (arrows) and Σ is of the form:

$$\Sigma(x) = (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)),$$

for an activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

Control system models of ResNets

- It was observed in the last 4 years[1] that the equation:

$$x(k+1) = x(k) + s(k)\Sigma(W(k)x(k) + b(k)), \tag{1}$$

is remarkably similar to the forward Euler discretization of the continuous-time control system:

$$\dot{x} = s\Sigma(Wx + b), \tag{2}$$

with state $x \in \mathbb{R}^n$ and where $(s, W, b) \in \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ are regarded as control inputs.

[1] **A proposal on machine learning via dynamical systems**
E. Weinan, Communications in Mathematics and Statistics, 5, 2017.
Stable architectures for deep neural networks
E. Haber and L. Ruthotto, Inverse Problems, 34(1), 2017.
Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations
Y. Lu, A. Zhong, Q. Li, and B. Dong, International Conference on Machine Learning, 2018.

Control system models of ResNets

- It was observed in the last 4 years[1] that the equation:

$$x(k+1) = x(k) + s(k)\Sigma(W(k)x(k) + b(k)), \tag{1}$$

is remarkably similar to the forward Euler discretization of the continuous-time control system:

$$\dot{x} = s\Sigma(Wx + b), \tag{2}$$

with state $x \in \mathbb{R}^n$ and where $(s, W, b) \in \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ are regarded as control inputs.

- Properties of (2) approximately transfer to (1) by time-discretizing solutions.

[1] **A proposal on machine learning via dynamical systems**
E. Weinan, Communications in Mathematics and Statistics, 5, 2017.
Stable architectures for deep neural networks
E. Haber and L. Ruthotto, Inverse Problems, 34(1), 2017.
Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations
Y. Lu, A. Zhong, Q. Li, and B. Dong, International Conference on Machine Learning, 2018.

Memorization Capabilities of Residual Neural Networks

CyPhyLab UCLA

Memorization capabilities of ResNets

Problem (Memorization)

Given:

- *a function* $f : E \to \mathbb{R}^n$ *defined on a compact set* $E \subset \mathbb{R}^n$*,*
- *a finite set* $E_{samples} \subset E$*,*
- *the evaluation of* f *on* $E_{samples}$*, i.e.,* $f(x)$ *for each* $x \in E_{samples}$*,*

does there exist a ResNet outputing $f(x)$ *for each input* $x \in E_{samples}$*?*

Memorization capabilities of ResNets

Problem (Memorization)

Given:

- *a function* $f : E \to \mathbb{R}^n$ *defined on a compact set* $E \subset \mathbb{R}^n$*,*
- *a finite set* $E_{samples} \subset E$*,*
- *the evaluation of* f *on* $E_{samples}$*, i.e.,* $f(x)$ *for each* $x \in E_{samples}$*,*

does there exist a ResNet outputing $f(x)$ *for each input* $x \in E_{samples}$*?*

Problem (Memorization)

Does there exist a time $\tau \in \mathbb{R}_0^+$ *and an input* $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ *so that the solution* ξ *of:*

$$\dot{x} = s\Sigma(Wx + b),$$

satisfies $\xi(0) = x$ *and* $\xi(\tau) = f(x)$ *for every* $x \in E_{samples}$*.*

CyPhyLab UCLA

Memorization capabilities of ResNets

Problem (Memorization)

Given:

- *a function $f : E \to \mathbb{R}^n$ defined on a compact set $E \subset \mathbb{R}^n$,*
- *a finite set $E_{samples} \subset E$,*
- *the evaluation of f on $E_{samples}$, i.e., $f(x)$ for each $x \in E_{samples}$,*

does there exist a ResNet outputing $f(x)$ for each input $x \in E_{samples}$?

Problem (Memorization)

Does there exist a time $\tau \in \mathbb{R}_0^+$ and an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ so that the solution ξ of:

$$\dot{x} = s\Sigma(Wx + b),$$

satisfies $\xi(0) = x$ and $\xi(\tau) = f(x)$ for every $x \in E_{samples}$.

- Is this a controllability problem?

CyPhyLab UCLA

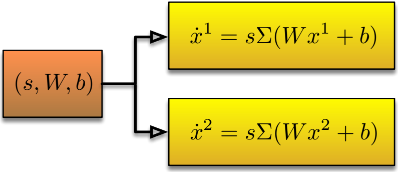
Memorization capabilities of ResNets

Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.

Memorization capabilities of ResNets
Controllability

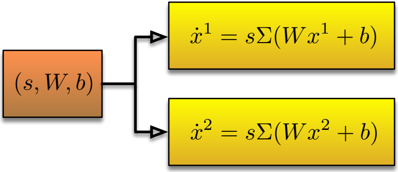
- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.



The same input needs to control two copies of the same system.

Memorization capabilities of ResNets
Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.

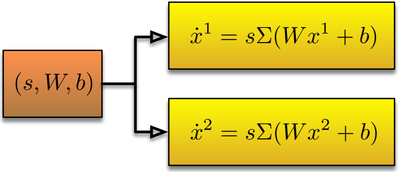


  The same input needs to control two copies of the same system.
- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?

Memorization capabilities of ResNets

Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.



  The same input needs to control two copies of the same system.
- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?

Memorization capabilities of ResNets

Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.

$$(s, W, b) \to \begin{cases} \dot{x}^1 = s\Sigma(Wx^1 + b) \\ \dot{x}^2 = s\Sigma(Wx^2 + b) \end{cases}$$

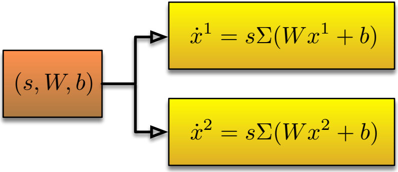
The same input needs to control two copies of the same system.

- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?

CyPhyLab UCLA

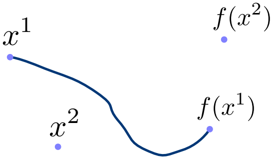
Memorization capabilities of ResNets
Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.



The same input needs to control two copies of the same system.

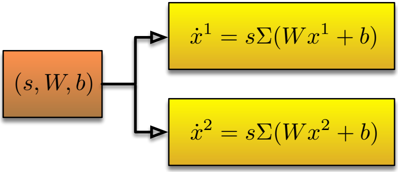
- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?



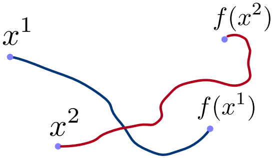
Memorization capabilities of ResNets

Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.



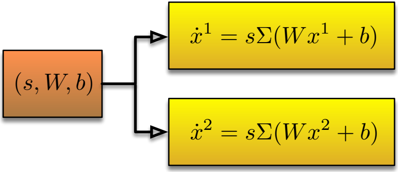
  The same input needs to control two copies of the same system.
- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?



Memorization capabilities of ResNets

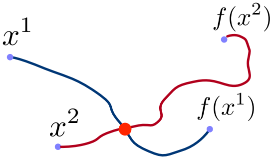
Controllability

- Let's consider the case where $E_{\text{samples}} = \{x^1, x^2\}$.



The same input needs to control two copies of the same system.

- Controllability on $\mathbb{R}^n \times \mathbb{R}^n$?



Memorization capabilities of ResNets

Ensemble controllability

- Given a finite set of samples $E_{\text{samples}} = \{x^1, x^2, \ldots, x^d\}$ we consider the ensemble control system:

$$\dot{X} = \left[s\Sigma(WX_{\bullet 1} + b)|s\Sigma(WX_{\bullet 2} + b)| \ldots |s\Sigma(WX_{\bullet d} + b)\right], \tag{3}$$

where the state $X(t) \in \mathbb{R}^{n\times d}$ is the matrix:

$$X(t) = \left[X_{\bullet 1}(t)|X_{\bullet 2}(t)| \ldots |X_{\bullet d}(t)\right].$$

- Given a finite set of samples $E_{\text{samples}} = \{x^1, x^2, \ldots, x^d\}$ we consider the ensemble control system:

$$\dot{X} = \left[s\Sigma(WX_{\bullet 1} + b)|s\Sigma(WX_{\bullet 2} + b)|\ldots|s\Sigma(WX_{\bullet d} + b)\right], \qquad (3)$$

  where the state $X(t) \in \mathbb{R}^{n \times d}$ is the matrix:

$$X(t) = \left[X_{\bullet 1}(t)|X_{\bullet 2}(t)|\ldots|X_{\bullet d}(t)\right].$$

- We can now ask: does there exist an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ so that the solution X of (3) satisfies:

$$X(0) = \left[x^1|x^2|\ldots|x^d\right] \text{ and } X(\tau) = \left[f(x^1)|f(x^2)|\ldots|f(x^d)\right]?$$

- As typically done in geometric control theory, we work with piecewise constant control inputs so that for each choice of input we obtain a vector field:

$$\mathcal{F} = \{Z_1, Z_2, \ldots, Z_k\}.$$

[2] **Orbits of families of vector fields and integrability of distributions**. Héctor Sussmann. Transactions of the American Mathematical Society, 1973.

- As typically done in geometric control theory, we work with piecewise constant control inputs so that for each choice of input we obtain a vector field:

$$\mathcal{F} = \{Z_1, Z_2, \ldots, Z_k\}.$$

- What is the orbit (the reachable space) of the family $\mathcal{F}$ of vector fields?

[2] **Orbits of families of vector fields and integrability of distributions**. Héctor Sussmann. Transactions of the American Mathematical Society, 1973.

- As typically done in geometric control theory, we work with piecewise constant control inputs so that for each choice of input we obtain a vector field:

$$\mathcal{F} = \{Z_1, Z_2, \ldots, Z_k\}.$$

- What is the orbit (the reachable space) of the family $\mathcal{F}$ of vector fields?
- Difficult problem, in general, that has a simpler answer[2] when $\mathcal{F}$ is symmetric, i.e.:

$$Z \in \mathcal{F} \implies -Z \in \mathcal{F}.$$

[2] **Orbits of families of vector fields and integrability of distributions**. Héctor Sussmann. Transactions of the American Mathematical Society, 1973.

Memorization capabilities of ResNets

Ensemble controllability

- As typically done in geometric control theory, we work with piecewise constant control inputs so that for each choice of input we obtain a vector field:

$$\mathcal{F} = \{Z_1, Z_2, \ldots, Z_k\}.$$

- What is the orbit (the reachable space) of the family $\mathcal{F}$ of vector fields?
- Difficult problem, in general, that has a simpler answer[2] when $\mathcal{F}$ is symmetric, i.e.:

$$Z \in \mathcal{F} \implies -Z \in \mathcal{F}.$$

- It suffices to choose the inputs $s \in \mathbb{R}$ to range in $\{-1, 1\}$ to obtain symmetry.

[2] **Orbits of families of vector fields and integrability of distributions**. Héctor Sussmann. Transactions of the American Mathematical Society, 1973.

- As typically done in geometric control theory, we work with piecewise constant control inputs so that for each choice of input we obtain a vector field:

$$\mathcal{F} = \{Z_1, Z_2, \ldots, Z_k\}.$$

- What is the orbit (the reachable space) of the family $\mathcal{F}$ of vector fields?
- Difficult problem, in general, that has a simpler answer[2] when $\mathcal{F}$ is symmetric, i.e.:

$$Z \in \mathcal{F} \implies -Z \in \mathcal{F}.$$

- It suffices to choose the inputs $s \in \mathbb{R}$ to range in $\{-1, 1\}$ to obtain symmetry.
- The answer is then given by the Lie algebra rank condition:

the dimension of the Lie algebra generated by $\mathcal{F}$ equals nd at every $X \in \mathbb{R}^{n \times d}$.

[2] **Orbits of families of vector fields and integrability of distributions**. Héctor Sussmann. Transactions of the American Mathematical Society, 1973.

- The Lie algebra generated by $\mathcal{F}$ is the smallest vector subspace of $T\mathbb{R}^{n\times d}$ containing $\mathcal{F}$ and closed under Lie brackets:

$$[Z_1, Z_2] = \frac{\partial Z_2}{\partial A} Z_1 - \frac{\partial Z_1}{\partial A} Z_2.$$

Memorization capabilities of ResNets

Ensemble controllability

- The Lie algebra generated by $\mathcal{F}$ is the smallest vector subspace of $T\mathbb{R}^{n\times d}$ containing $\mathcal{F}$ and closed under Lie brackets:

$$[Z_1, Z_2] = \frac{\partial Z_2}{\partial A} Z_1 - \frac{\partial Z_1}{\partial A} Z_2.$$

- Since the activation function is not known, these brackets are not known.

Memorization capabilities of ResNets

Ensemble controllability

- The Lie algebra generated by $\mathcal{F}$ is the smallest vector subspace of $T\mathbb{R}^{n\times d}$ containing $\mathcal{F}$ and closed under Lie brackets:

$$[Z_1, Z_2] = \frac{\partial Z_2}{\partial A} Z_1 - \frac{\partial Z_1}{\partial A} Z_2.$$

- Since the activation function is not known, these brackets are not known.
- Choose the inputs W and b so that the vector fields and their Lie brackets only contain:

$$\sigma, D\sigma, D^2\sigma, \ldots$$

Memorization capabilities of ResNets

Ensemble controllability

- The Lie algebra generated by $\mathcal{F}$ is the smallest vector subspace of $T\mathbb{R}^{n\times d}$ containing $\mathcal{F}$ and closed under Lie brackets:

$$[Z_1, Z_2] = \frac{\partial Z_2}{\partial A} Z_1 - \frac{\partial Z_1}{\partial A} Z_2.$$

- Since the activation function is not known, these brackets are not known.
- Choose the inputs W and b so that the vector fields and their Lie brackets only contain:

$$\sigma, D\sigma, D^2\sigma, \ldots$$

- With such choice, and after judicious (and tedious) manipulations, the rank of the Lie algebra is nd provided the rank of the following matrix is n:

$$\begin{bmatrix} 1 & \sigma(A_{1\ell}) & D\sigma(A_{1\ell}) & \cdots & D^{n-2}\sigma(A_{1\ell}) \\ 1 & \sigma(A_{2\ell}) & D\sigma(A_{2\ell}) & \cdots & D^{n-2}\sigma(A_{2\ell}) \\ \vdots & \vdots & & \vdots & \\ 1 & \sigma(A_{n\ell}) & D\sigma(A_{n\ell}) & \cdots & D^{n-2}\sigma(A_{n\ell}) \end{bmatrix}.$$

- When is the determinant of this matrix nonzero:

$$\begin{bmatrix} 1 & \sigma(A_{1\ell}) & D\sigma(A_{1\ell}) & \cdots & D^{n-2}\sigma(A_{1\ell}) \\ 1 & \sigma(A_{2\ell}) & D\sigma(A_{2\ell}) & \cdots & D^{n-2}\sigma(A_{2\ell}) \\ \vdots & \vdots & & \vdots & \\ 1 & \sigma(A_{n\ell}) & D\sigma(A_{n\ell}) & \cdots & D^{n-2}\sigma(A_{n\ell}) \end{bmatrix}?$$

- When is the determinant of this matrix nonzero:

$$\begin{bmatrix} 1 & \sigma(A_{1\ell}) & D\sigma(A_{1\ell}) & \cdots & D^{n-2}\sigma(A_{1\ell}) \\ 1 & \sigma(A_{2\ell}) & D\sigma(A_{2\ell}) & \cdots & D^{n-2}\sigma(A_{2\ell}) \\ \vdots & \vdots & & \vdots & \\ 1 & \sigma(A_{n\ell}) & D\sigma(A_{n\ell}) & \cdots & D^{n-2}\sigma(A_{n\ell}) \end{bmatrix}?$$

- Key idea: relation to Vandermonde matrices?

Lemma

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a function that satisfies the quadratic differential equation:

$$D\sigma(x) = a_0 + a_1\sigma(x) + a_2\sigma^2(x),$$

where $a_0, a_1, a_2 \in \mathbb{R}$. Suppose that derivatives of σ of up to order $(\ell - 2)$ exist at ℓ points $x_1, \ldots, x_\ell \in \mathbb{R}$. Then, the determinant of the matrix:

$$L(x_1, x_2, \ldots, x_\ell) = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \sigma(x_1) & \sigma(x_2) & \ldots & \sigma(x_\ell) \\ D\sigma(x_1) & D\sigma(x_2) & \ldots & D\sigma(x_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ D^{\ell-2}\sigma(x_1) & D^{\ell-2}\sigma(x_2) & \ldots & D^{\ell-2}\sigma(x_\ell) \end{bmatrix},$$

is given by:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i!\, a_2^i \prod_{1 \leq i < j \leq \ell} (\sigma(x_i) - \sigma(x_j)).$$

Memorization capabilities of ResNets

Ensemble controllability

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1 \leq i < j \leq \ell} (\sigma(x_i) - \sigma(x_j))?$$

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1 \leq i < j \leq \ell} (\sigma(x_i) - \sigma(x_j))?$$

- If $a_2 \neq 0$ and σ is injective, $\prod_{1 \leq i < j \leq \ell}(x_i - x_j) \neq 0$ implies $\det L \neq 0$.

Memorization capabilities of ResNets

Ensemble controllability

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1 \leq i < j \leq \ell} (\sigma(x_i) - \sigma(x_j))?$$

- If $a_2 \neq 0$ and σ is injective, $\prod_{1 \leq i < j \leq \ell}(x_i - x_j) \neq 0$ implies $\det L \neq 0$.
- Two different ensemble elements i and j cannot be in states $X_{\bullet i}$ and $X_{\bullet j}$ that share an entry, i.e., for any ℓ: $X_{\ell i} \neq X_{\ell j}$.

Memorization capabilities of ResNets

Ensemble controllability

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1 \le i < j \le \ell} (\sigma(x_i) - \sigma(x_j))?$$

- If $a_2 \neq 0$ and σ is injective, $\prod_{1 \le i < j \le \ell}(x_i - x_j) \neq 0$ implies $\det L \neq 0$.
- Two different ensemble elements i and j cannot be in states $X_{\bullet i}$ and $X_{\bullet j}$ that share an entry, i.e., for any ℓ: $X_{\ell i} \neq X_{\ell j}$.
- Are there injective activation functions σ satisfying $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$?

Memorization capabilities of ResNets

Ensemble controllability

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1\leq i<j\leq \ell} (\sigma(x_i) - \sigma(x_j))?$$

- If $a_2 \neq 0$ and σ is injective, $\prod_{1\leq i<j\leq \ell}(x_i - x_j) \neq 0$ implies $\det L \neq 0$.
- Two different ensemble elements i and j cannot be in states $X_{\bullet i}$ and $X_{\bullet j}$ that share an entry, i.e., for any ℓ: $X_{\ell i} \neq X_{\ell j}$.
- Are there injective activation functions σ satisfying $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$?

Function name	**Definition**	**Satisfied differential equation**
Logistic function	$\sigma(x) = \frac{1}{1+e^{-x}}$	$D\sigma - \sigma + \sigma^2 = 0$
Hyperbolic tangent	$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$D\sigma - 1 + \sigma^2 = 0$
Soft plus	$\sigma(x) = \frac{1}{r}\log(1 + e^{rx})$	$D^2\sigma - rD\sigma + r(D\sigma)^2 = 0$

CyPhyLab UCLA

Memorization capabilities of ResNets

Ensemble controllability

- When is this expression non-zero:

$$\det L(x_1, x_2, \ldots, x_\ell) = \prod_{i=1}^{\ell-2} i! a_2^i \prod_{1 \leq i < j \leq \ell} (\sigma(x_i) - \sigma(x_j))?$$

- If $a_2 \neq 0$ and σ is injective, $\prod_{1 \leq i < j \leq \ell}(x_i - x_j) \neq 0$ implies $\det L \neq 0$.
- Two different ensemble elements i and j cannot be in states $X_{\bullet i}$ and $X_{\bullet j}$ that share an entry, i.e., for any ℓ: $X_{\ell i} \neq X_{\ell j}$.
- Are there injective activation functions σ satisfying $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$?

Function name	**Definition**	**Satisfied differential equation**
Logistic function	$\sigma(x) = \frac{1}{1+e^{-x}}$	$D\sigma - \sigma + \sigma^2 = 0$
Hyperbolic tangent	$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$D\sigma - 1 + \sigma^2 = 0$
Soft plus	$\sigma(x) = \frac{1}{r}\log(1 + e^{rx})$	$D^2\sigma - rD\sigma + r(D\sigma)^2 = 0$

Moreover, $\lim_{r\to\infty} \frac{1}{r}\log(1 + e^{rx}) = \mathrm{ReLU}(x) = \max\{0, x\}$.

Memorization capabilities of ResNets

Ensemble controllability

Theorem

Let $N \subset \mathbb{R}^{n\times d}$ be the set defined by:

$$N = \left\{ A \in \mathbb{R}^{n\times d} \mid \prod_{1\leq i<j\leq d} (A_{\ell i} - A_{\ell j}) = 0,\ \ell \in \{1,\ldots,n\} \right\}.$$

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable on the submanifold $M = \mathbb{R}^{n\times d}\backslash N$.

CyPhyLab UCLA

Memorization capabilities of ResNets

Ensemble controllability

Theorem

Let $N \subset \mathbb{R}^{n\times d}$ be the set defined by:

$$N = \left\{ A \in \mathbb{R}^{n\times d} \mid \prod_{1\le i<j\le d} (A_{\ell i} - A_{\ell j}) = 0,\ \ell \in \{1,\ldots,n\} \right\}.$$

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable on the submanifold $M = \mathbb{R}^{n\times d}\backslash N$.

- When $n > 1$, M is connected, open, and dense in $\mathbb{R}^{n\times d}$.

CyPhyLab UCLA

Memorization capabilities of ResNets

Ensemble controllability

Theorem

Let $N \subset \mathbb{R}^{n \times d}$ be the set defined by:

$$N = \left\{ A \in \mathbb{R}^{n \times d} \mid \prod_{1 \leq i < j \leq d} (A_{\ell i} - A_{\ell j}) = 0,\ \ell \in \{1, \dots, n\} \right\}.$$

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable on the submanifold $M = \mathbb{R}^{n \times d} \backslash N$.

- When $n > 1$, M is connected, open, and dense in $\mathbb{R}^{n \times d}$.
- If E_{samples} and $f(E_{\text{samples}})$ are subsets of M, a ResNet can memorize them exactly.

Memorization capabilities of ResNets

Ensemble controllability

Theorem

Let $N \subset \mathbb{R}^{n\times d}$ be the set defined by:

$$N = \left\{ A \in \mathbb{R}^{n\times d} \mid \prod_{1\leq i<j\leq d} (A_{\ell i} - A_{\ell j}) = 0,\ \ell \in \{1,\ldots,n\} \right\}.$$

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable on the submanifold $M = \mathbb{R}^{n\times d}\backslash N$.

- When $n > 1$, M is connected, open, and dense in $\mathbb{R}^{n\times d}$.
- If E_{samples} and $f(E_{\text{samples}})$ are subsets of M, a ResNet can memorize them exactly.
- Otherwise we can perturb E_{samples} or/and $f(E_{\text{samples}})$ to make them subsets of M.

CyPhyLab UCLA

Approximation Capabilities of Residual Neural Networks

Approximation capabilities of ResNets

- We established controllability on the finite dimensional state space $\mathbb{R}^{n\times d}$.
- But we would really like to establish controllability on some infinite dimensional space of functions[3].

[3] **Control on the manifold of mappings as a setting for deep learning.**
A. Agrachev and A. Sarychev. arXiv preprint arXiv:2008.12702, 2020.

Approximation capabilities of ResNets

- We established controllability on the finite dimensional state space $\mathbb{R}^{n\times d}$.
- But we would really like to establish controllability on some infinite dimensional space of functions[3].
- Let $\phi^t : \mathbb{R}^n \to \mathbb{R}^n$ be the flow defined by the solution of the control system:

$$\dot{x} = s\Sigma(Wx + b),$$

i.e., $\phi^t(x) = \xi(t)$ where ξ is the solution satisfying $\xi(0) = x$.

[3] **Control on the manifold of mappings as a setting for deep learning.** A. Agrachev and A. Sarychev. arXiv preprint arXiv:2008.12702, 2020.

Approximation capabilities of ResNets

- We established controllability on the finite dimensional state space $\mathbb{R}^{n\times d}$.
- But we would really like to establish controllability on some infinite dimensional space of functions[3].
- Let $\phi^t : \mathbb{R}^n \to \mathbb{R}^n$ be the flow defined by the solution of the control system:

$$\dot{x} = s\Sigma(Wx + b),$$

  i.e., $\phi^t(x) = \xi(t)$ where ξ is the solution satisfying $\xi(0) = x$.
- Do there exist inputs $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ resulting in a flow ϕ^t satisfying:

$$\phi^0(x) = x \text{ and } \phi^\tau(x) = f(x)?$$

[3] **Control on the manifold of mappings as a setting for deep learning.**
A. Agrachev and A. Sarychev. arXiv preprint arXiv:2008.12702, 2020.

Approximation capabilities of ResNets

- We established controllability on the finite dimensional state space $\mathbb{R}^{n\times d}$.
- But we would really like to establish controllability on some infinite dimensional space of functions[3].
- Let $\phi^t : \mathbb{R}^n \to \mathbb{R}^n$ be the flow defined by the solution of the control system:

$$\dot{x} = s\Sigma(Wx + b),$$

i.e., $\phi^t(x) = \xi(t)$ where ξ is the solution satisfying $\xi(0) = x$.

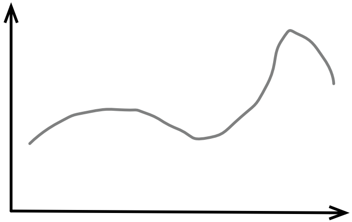
- Do there exist inputs $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ resulting in a flow ϕ^t satisfying:

$$\phi^0(x) = x \text{ and } \phi^\tau(x) = f(x)?$$

- Can we use the previous controllability result as a stepping stone?
  - If we map finitely many points to the right location, can things go wrong for the points we leave out?

[3] **Control on the manifold of mappings as a setting for deep learning.** A. Agrachev and A. Sarychev. arXiv preprint arXiv:2008.12702, 2020.

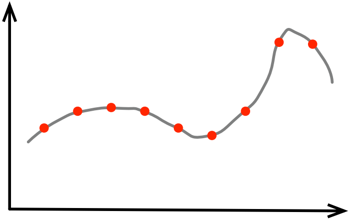
- Consider the function interpolation problem.



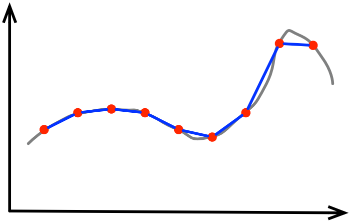
Approximation capabilities of ResNets

Inspiration from function interpolation

- Consider the function interpolation problem.



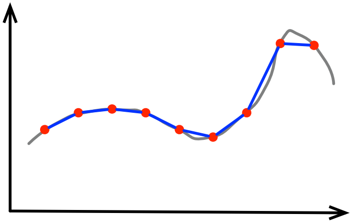
- Consider the function interpolation problem.



Approximation capabilities of ResNets

Inspiration from function interpolation

- Consider the function interpolation problem.

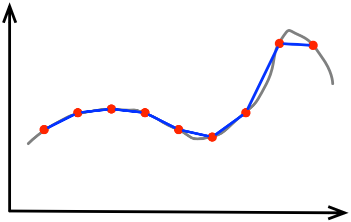


- How to control the behavior of the interpolating function between the interpolation points?

Approximation capabilities of ResNets

Inspiration from function interpolation

- Consider the function interpolation problem.



- How to control the behavior of the interpolating function between the interpolation points?
- Key idea: monotonicity.

Approximation capabilities of ResNets
Monotonicity

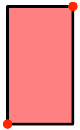
- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x'_i$ for all $i = 1, 2, \ldots, n$.

- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x_i'$ for all $i = 1, 2, \ldots, n$.
  - A flow $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if:

$$x \preceq x' \implies \phi(x) \preceq \phi(x').$$

- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x_i'$ for all $i = 1, 2, \ldots, n$.
  - A flow $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if:

$$x \preceq x' \implies \phi(x) \preceq \phi(x').$$

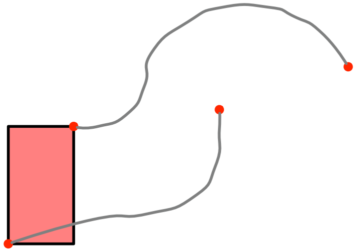


Approximation capabilities of ResNets

Monotonicity

- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x_i'$ for all $i = 1, 2, \ldots, n$.
  - A flow $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if:

$$x \preceq x' \implies \phi(x) \preceq \phi(x').$$

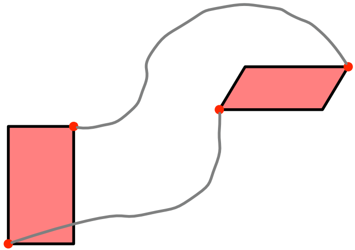


Approximation capabilities of ResNets

Monotonicity

- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x_i'$ for all $i = 1, 2, \ldots, n$.
  - A flow $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if:

$$x \preceq x' \implies \phi(x) \preceq \phi(x').$$

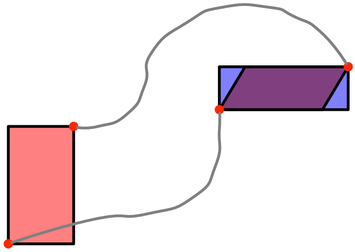


Approximation capabilities of ResNets

Monotonicity

- What is monotonicity?
  - Define the ordering $\preceq$ on $\mathbb{R}^n$ by $x \preceq x'$ iff $x_i \leq x'_i$ for all $i = 1, 2, \ldots, n$.
  - A flow $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if:

$$x \preceq x' \implies \phi(x) \preceq \phi(x').$$



- When the function f to be learned is monotone, we can construct a monotone flow ϕ^t, by using the previous controllability result, approximating f on E_{samples}.

Theorem

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every monotone analytic function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exist a time $\tau \in \mathbb{R}^+$ and an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ so that the flow $\phi^\tau : \mathbb{R}^n \to \mathbb{R}^n$ defined by the solution of $\dot{x} = s\Sigma(Wx + b)$ under the said input satisfies:

$$\|f - \phi^\tau\|_{L^\infty(E)} \leq \varepsilon.$$

Approximation capabilities of ResNets

Main result

- When the function f to be learned is monotone, we can construct a monotone flow ϕ^t, by using the previous controllability result, approximating f on E_{samples}.

Theorem

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every monotone analytic function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exist a time $\tau \in \mathbb{R}^+$ and an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ so that the flow $\phi^\tau : \mathbb{R}^n \to \mathbb{R}^n$ defined by the solution of $\dot{x} = s\Sigma(Wx + b)$ under the said input satisfies:

$$\|f - \phi^\tau\|_{L^\infty(E)} \leq \varepsilon.$$

- What happens when f is not monotone?

- When the function f to be learned is monotone, we can construct a monotone flow ϕ^t, by using the previous controllability result, approximating f on E_{samples}.

Theorem

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every monotone analytic function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exist a time $\tau \in \mathbb{R}^+$ and an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ so that the flow $\phi^\tau : \mathbb{R}^n \to \mathbb{R}^n$ defined by the solution of $\dot{x} = s\Sigma(Wx + b)$ under the said input satisfies:

$$\|f - \phi^\tau\|_{L^\infty(E)} \leq \varepsilon.$$

- What happens when f is not monotone?
- Key idea: monotone embedding.

CyPhyLab UCLA

Approximation capabilities of ResNets

Main result

- We seek:
  - a linear injection $\alpha : \mathbb{R}^n \to \mathbb{R}^{n+1}$,
  - a linear projection $\beta : \mathbb{R}^{n+1} \to \mathbb{R}^n$,
  - a monotone function $\widetilde{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$,

  so that:

$$f = \beta \circ \widetilde{f} \circ \alpha.$$

Approximation capabilities of ResNets

Main result

- We seek:
  - a linear injection $\alpha : \mathbb{R}^n \to \mathbb{R}^{n+1}$,
  - a linear projection $\beta : \mathbb{R}^{n+1} \to \mathbb{R}^n$,
  - a monotone function $\widetilde{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$,

  so that:

$$f = \beta \circ \widetilde{f} \circ \alpha.$$

- The functions α and β can be implemented by the first and last layers of a ResNet.
- The intermediate ResNet approximates the monotone function $\widetilde{f}$.

Approximation capabilities of ResNets

Main result

- We seek:
  - a linear injection $\alpha : \mathbb{R}^n \to \mathbb{R}^{n+1}$,
  - a linear projection $\beta : \mathbb{R}^{n+1} \to \mathbb{R}^n$,
  - a monotone function $\widetilde{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$,

  so that:

$$f = \beta \circ \widetilde{f} \circ \alpha.$$

- The functions α and β can be implemented by the first and last layers of a ResNet.
- The intermediate ResNet approximates the monotone function $\widetilde{f}$.
- This can be accomplished with:
  - $\alpha(x) = (x, \mathbf{1}^T x) = (x, x_1 + x_2 + \ldots + x_n)$,
  - $\beta(x, y) = x - \kappa y$,
  - $\widetilde{f}(x, y) = (f(x) + \kappa \mathbf{1} y, y)$.

Corollary

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exist a time $\tau \in \mathbb{R}^+$, an injection $\alpha : \mathbb{R}^n \to \mathbb{R}^{n+1}$, a projection $\beta : \mathbb{R}^{n+1} \to \mathbb{R}^n$, and an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{(n+1)\times(n+1)} \times \mathbb{R}^{n+1}$ so that the flow $\phi^\tau : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defined by the solution of $\dot{x} = s\Sigma(Wx + b)$ under the said input satisfies:

$$\|f - \beta \circ \phi^\tau \circ \alpha\|_{L^\infty(E)} \leq \varepsilon.$$

A Deterministic Generalization Bound

CyPhyLab UCLA

A deterministic generalization bound

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous map defined on a compact set $E \subset \mathbb{R}^n$. Suppose $E_{samples} \subset \mathbb{R}^n$ is a finite set satisfying:

$$\forall x \in E \quad \exists \underline{x}, \overline{x} \in E_{samples}, \qquad |\underline{x} - \overline{x}|_\infty \leq \delta \quad \wedge \quad \underline{x} \preceq x \preceq \overline{x}, \tag{4}$$

with $\delta \in \mathbb{R}^+$. For any monotone map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ we have:

$$\|f - \phi\|_{L^\infty(E)} \leq 2\omega_f(\delta) + 3\|f - \phi\|_{L^\infty(E_{samples})},$$

where ω_f is the modulus[a] of continuity of f.

[a] Note that f, being continuous, is uniformly continuous on any compact set.

CyPhyLab UCLA

A deterministic generalization bound

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous map defined on a compact set $E \subset \mathbb{R}^n$. Suppose $E_{samples} \subset \mathbb{R}^n$ is a finite set satisfying:

$$\forall x \in E \quad \exists \underline{x}, \overline{x} \in E_{samples}, \qquad |\underline{x} - \overline{x}|_\infty \leq \delta \quad \wedge \quad \underline{x} \preceq x \preceq \overline{x}, \tag{4}$$

with $\delta \in \mathbb{R}^+$. For any monotone map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ we have:

$$\|f - \phi\|_{L^\infty(E)} \leq 2\omega_f(\delta) + 3\|f - \phi\|_{L^\infty(E_{samples})},$$

where ω_f is the modulus[a] of continuity of f.

[a] Note that f, being continuous, is uniformly continuous on any compact set.

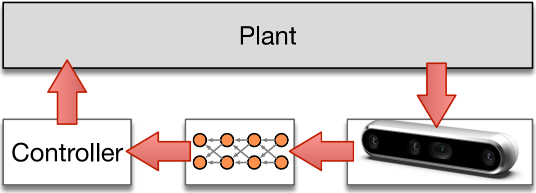
- Can we use such bound in a control context?

Outlook

Outlook

Stability guarantees with deep perception pipelines

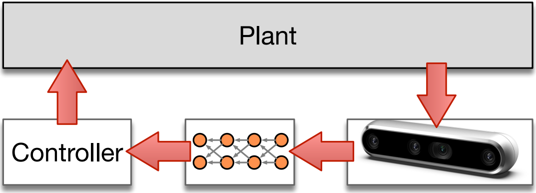
- Consider a closed-loop system with a ResNet in the perception pipeline.



Outlook

Stability guarantees with deep perception pipelines

- Consider a closed-loop system with a ResNet in the perception pipeline.



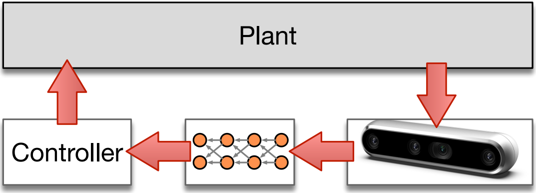
- Assume the controller $u = k(\widehat{x}) = k(x + e)$ renders the closed-loop system ISS with respect to estimation errors e, i.e.:

$$\|x(t)\| \leq \beta(\|x(0)\|, t) + \gamma(\|e\|_{L^\infty}).$$

Outlook

Stability guarantees with deep perception pipelines

- Consider a closed-loop system with a ResNet in the perception pipeline.



- Assume the controller $u = k(\widehat{x}) = k(x + e)$ renders the closed-loop system ISS with respect to estimation errors e, i.e.:

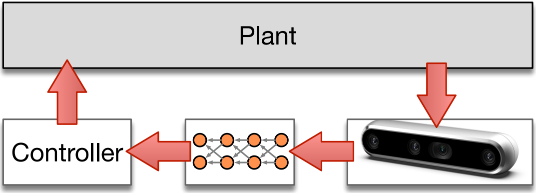
$$\|x(t)\| \leq \beta(\|x(0)\|, t) + \gamma(\|e\|_{L^\infty}).$$

- Train a ResNet to learn the map from output measurements y to the state x, i.e., to act as an observer $\phi(y) = \widehat{x}$.

Outlook

Stability guarantees with deep perception pipelines

- Consider a closed-loop system with a ResNet in the perception pipeline.



- Assume the controller $u = k(\widehat{x}) = k(x + e)$ renders the closed-loop system ISS with respect to estimation errors e, i.e.:

$$\|x(t)\| \leq \beta(\|x(0)\|, t) + \gamma(\|e\|_{L^\infty}).$$

- Train a ResNet to learn the map from output measurements y to the state x, i.e., to act as an observer $\phi(y) = \widehat{x}$.
- By the generalization lemma, $e = \widehat{x} - x$ is bounded by a constant $c \in \mathbb{R}^+$ and we directly obtain practical stability:

$$\|x(t)\| \leq \beta(\|x(0)\|, t) + \gamma(c).$$

Summary

- Ideas from control can play a big role in understanding deep neural networks.

Summary

- Ideas from control can play a big role in understanding deep neural networks.
- Approximation and generalization guarantees are possible in deterministic settings.

Summary

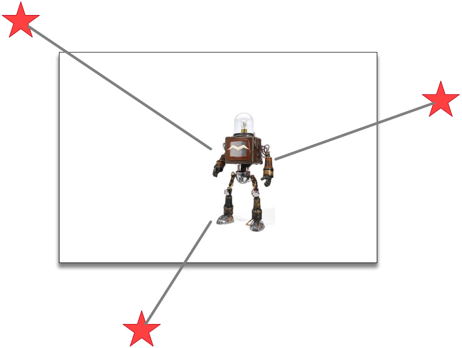
- Ideas from control can play a big role in understanding deep neural networks.
- Approximation and generalization guarantees are possible in deterministic settings.
- We can start to imagine control loops with learning components designed to satisfy formal safety and performance guarantees.

Summary

- However, there are still many challenges.
  - Learning observers from vision/LiDAR data require us to address extrapolation (the generalization lemma is about interpolation).

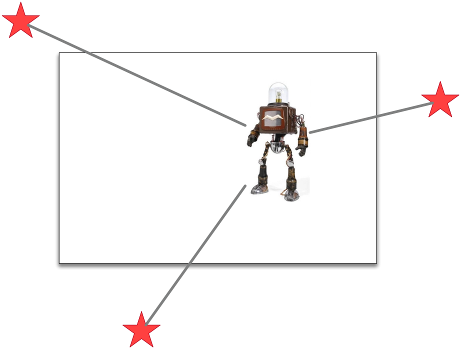
Summary

- However, there are still many challenges.
  - Learning observers from vision/LiDAR data require us to address extrapolation (the generalization lemma is about interpolation).



Summary

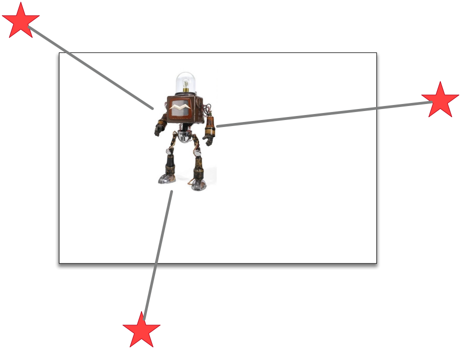
- However, there are still many challenges.
  - Learning observers from vision/LiDAR data require us to address extrapolation (the generalization lemma is about interpolation).



CyPhyLab UCLA

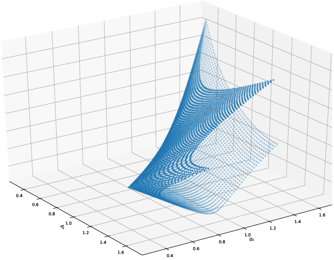
Summary

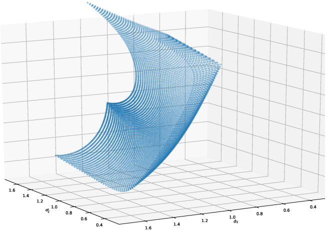
- However, there are still many challenges.
  - Learning observers from vision/LiDAR data require us to address extrapolation (the generalization lemma is about interpolation).



Summary

- However, there are still many challenges.
  - Learning observers from vision/LiDAR data require us to address extrapolation (the generalization lemma is about interpolation).





Acknowledgements

- Students and collaborators;
- SRC, DARPA,
- Prof. Mourão.

For more information:
http://www.cyphylab.ee.ucla.edu/
http://www.ee.ucla.edu/∼tabuada