2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Neural Network Independence Properties with Applications to Adaptive Control

Andrew Lamperski

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992994

***Abstract*— Neural networks form a general purpose architecture for machine learning and parameter identification. The simplest neural network consists of a single hidden layer connected to a linear output layer. It is often assumed that the components of the hidden layer correspond to linearly independent functions, but proofs of this are only known for a few specialized classes of network activation functions. This paper shows that for wide class of activation functions, including most of the commonly used activation functions in neural network libraries, almost all choices of hidden layer parameters lead to linearly independent functions. These linear independence properties are then used to derive sufficient conditions for persistence of excitation, a condition commonly used to ensure parameter convergence in adaptive control.**

I. Introduction

Neural networks are widely employed in machine learning and adaptive control. Common applications include natural language processing and image processing from learning[1] and model reference adaptive control [2].

This simplest neural network architecture consists of a single hidden layer and an output layer. Such a network is a function, $\psi(x,\theta)$, defined by:

$$\psi(x,\theta) = c_0 + \sum_{i=1}^{m} c_i \sigma(w_i^\top x + b_i), \tag{1}$$

where $\theta = (w,b,c)$ contains the parameters of the network, x is the input variable, and σ is a nonlinear function called the *activation function*. Common activation functions include sigmoids, step functions, and ReLUs.

Classical results in neural network theory [3], [4] establish than under minimal assumptions on the activation function, any continuous function can be approximated to arbitrary accuracy by a single hidden layer neural network from (1).

The main contribution of this paper gives sufficient conditions to ensure that $\sigma(w_i^\top x + b_i)$ are linearly independent as functions of x. We prove that for large class of activation functions, which includes most activation functions in the PyTorch library [5], the functions $\sigma(w_i^\top x + b_i)$ are linearly independent for almost all choices of w_i and b_i. The result implies, in particular, that if w_i and b_i are generated independently from a continuous distribution whose support has positive measure (with respect to Lebesgue measure), the functions are linearly independent with probability 1.

This work was supported in part by NSF CMMI-2122856

A. Lamperski is with the department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA alampers@umn.edu

We utilize the linear independence to give sufficient conditions for persistency of excitation, which is commonly assumed without proof in adaptive control.

Beyond persistency of excitation, linear independence of neural networks is assumed in proofs of consistency of function approximation schemes [6], [7]. However, these works do not give conditions for linear independence to hold.

Linear independence properties were showed for classes of analytic activation functions in [8], but the literature directly addressing linear independence of neural networks is sparse. The most closely related topics that have received wide study are neural network interpolation (see [3], [9], [10]) and network identifiability (see [8], [11], [12]).

Section II presents the main results on linear independence and Section III shows how they can be used to derive persistence of excitation conditions for adaptive control. Conclusions are given in Section IV

Notation: The indicator function is denoted by $\mathbb{1}$. If x is a vector, its Euclidean norm is denoted by $\|x\|$. $\mathbb{N}$ denotes the set of non-negative integers, $\mathbb{R}$ denotes the set of real numbers, and $\mathbb{C}$ denotes the set of complex numbers. The Lebesgue measure over $\mathbb{R}^n$ is denoted by μ.

II. Linear Independence of Nonlinear Functions

A. Fourier Transforms of Generalized Functions

Our results will be derived via Fourier transforms. The Fourier transforms of most common activation functions must be interpreted as generalized functions, also known as tempered distributions. The required background is reviewed briefly below. See [13], [14] for more details.

For a differentiable function f and a multi-index $m = (m_1,\dots,m_n) \in \mathbb{N}^n$, let $f^{(m)}(x) = \frac{\partial^{m_1}}{\partial y_1^{m_1}} \cdots \frac{\partial^{m_n}}{\partial y_n^{m_n}} f(y)|_{y=x}$ and let $|m| = \sum_{i=1}^n m_i$. The function f is called a *Schwartz function* if it is infinitely differentiable, and for every $n \geq 1$

$$\max_{|m|\leq n} \sup_{x\in\mathbb{R}} (1+\|x\|^2)^n \left| f^{(m)}(x) \right| < \infty \tag{2}$$

The space of Schwartz functions from $\mathbb{R}^n$ to $\mathbb{C}$ is called the *Schwartz space*, and is denoted by $\mathcal{S}_n$. For simplicity of notation, we denote $\mathcal{S}_1 = \mathcal{S}$. The expression on the left of (2) can be used to define a topology on the space of Schwartz space. See [14]. A *generalized function* is a continuous linear functional on $\mathcal{S}$. Its action on a Schwartz function, f, is denoted by $\langle g, f\rangle$. We often use the term *regular function* to distinguish from generalized functions.

A function $g : \mathbb{R} \to \mathbb{C}$ has *polynomial growth* if there is a constant c and an integer n such that $|g(x)| \leq c(1+|x|)^n$

for all x. If g is measurable and has polynomial growth, it can be identified with a generalized function via

$$\langle g, f\rangle = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{3}$$

As a minor abuse of notation, we denote the "value" of a generalized function as $g(x)$, since the generalized functions in this paper will behave like regular functions, except at a few singular points.

The *Fourier transform* of a Schwartz function is given by

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} e^{-j2\pi\omega x} f(x)dx,$$

where j is the imaginary unit.

If g is a generalized function, its Fourier transform and its distributional derivative are defined as the generalized functions that respectively satisfy for all $f \in \mathcal{S}$:

$$\langle \hat{g}, f\rangle = \langle g, \hat{f}\rangle$$
$$\langle g', f\rangle = -\langle g, f'\rangle.$$

We will show that the Fourier transforms of many common neural network activation functions can be expressed in terms of Dirac delta functions and generalized functions of the form $\frac{1}{\omega^m}h(\omega)$, where $h(\omega)$ is a regular function.

The Dirac delta and its distributional derivatives of order $m \geq 0$ are given by

$$\langle \delta^{(m)}, f\rangle = (-1)^m f^{(m)}(0).$$

The functions $p_{-m}(x) = \frac{1}{x^m}$ for integers $m \geq 1$ can be viewed as generalized functions with singularities at 0. As discussed in [13], p_{-m} can be defined as distributional derivatives $p_{-m} = \frac{(-1)^{m+1}}{(m-1)!}\ell^{(m+2)}$ of the regular function $\ell(x) = x\log|x| - x$, with action given by

$$\langle p_{-m}, f\rangle = \int_{-\infty}^{\infty} \frac{f(x) - \sum_{i=0}^{m-1} \frac{1}{i!}f^{(i)}(0)x^i}{x^m}dx. \tag{4}$$

So, we can see that $p_{-m}(x)$ behaves exactly like $\frac{1}{x^m}$ on all $f \in \mathcal{S}$ with $f(0) = f^{(1)}(0) = \cdots = f^{(m-1)}(0) = 0$.

In general, the product of a generalized function with a regular function may not be a generalized function. If $h : \mathbb{R} \to \mathbb{C}$ satisfies the following property

$$hf \in \mathcal{S} \quad \forall f \in \mathcal{S}, \tag{5}$$

then gh is the generalized function with action defined by:

$$\langle gh, f\rangle = \langle g, hf\rangle \quad \forall f \in \mathcal{S}.$$

A sufficient condition for (5) is if h is infinitely differentiable and $h^{(k)}$ has polynomial growth for all $k \geq 0$.

If h satisfies (5), then $p_{-m}(x)h(x) = \frac{1}{x^m}h(x)$ is a generalized function. Then since x^m and all its derivatives have polynomial growth, $x^m p_{-m}(x)h(x)$ is a generalized function, and it can be shown that $x^m p_{-m}(x)h(x) = h(x)$.

The following lemma is used to give expressions for Fourier transforms of neural network activation functions.

***Lemma** 1: Say that h is a generalized function such that $\hat{h}$ is a regular function that satisfies (5). If g is a generalized function such that $g^{(m)}(x) = h(x)$, then there are constants, $c_1, \ldots, c_{m-1}$, such that*

$$\hat{g}(\omega) = \frac{1}{(j2\pi\omega)^m}\hat{h}(\omega) + \sum_{i=0}^{m-1} c_i\delta^{(i)}(\omega) \tag{6}$$

Proof: Applying the derivative rule for Fourier transforms gives $\hat{h}(\omega) = (j2\pi\omega)^m\hat{g}(\omega)$.

The assumptions on $\hat{h}(\omega)$ imply that $\hat{h}(\omega) = (j2\pi\omega)^m\left(\frac{1}{(j2\pi\omega)^m}\hat{h}(\omega)\right)$, where $\frac{1}{(j2\pi\omega)^m}\hat{h}(\omega)$ defines a valid generalized function. It follows that

$$(j2\pi\omega)^m\left(\hat{g}(\omega) - \frac{1}{(j2\pi\omega)^m}\hat{h}(\omega)\right) = 0.$$

Exercise 7.23 of [13] shows that there must be constants $c_0, \ldots, c_{m-1}$ such that 6 holds. ∎

***Example** 1:* If $\sigma(t)$ is piecewise polynomial with $m \geq 1$ non-smooth points, then there is an integer n, non-negative integers $k_1, \ldots, k_m \leq n$, and real numbers $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ such that

$$\sigma^{(n)}(t) = \sum_{i=1}^{m} a_i\delta^{(k_i)}(t - b_i).$$

Lemma 1 implies that there exist $c_0, \ldots, c_{n-1}$ such that

$$\hat{\sigma}(\omega) = \sum_{i=1}^{m} a_i(j2\pi\omega)^{k_i-n}e^{-j2\pi\omega b_i} + \sum_{i=0}^{n-1} c_i\delta^{(i)}(\omega).$$

This formula gives expressions for the Fourier transforms the following activation functions from PyTorch [5]: HardShrink, HardSigmoid, HardTanh, HardSwish, LeakyReLU, PReLU, ReLU, ReLU6, RReLU, SoftShrink, Threshold.

Also note that $\widehat{\sigma^{(n)}}(\omega) = \sum_{i=1}^{m} a_i(j2\pi\omega)^{k_i}e^{-j2\pi\omega b_i}$ must be non-zero for almost all $\omega \in \mathbb{R}$ with respect to the Lebesgue measure. Arguing as in Theorem 1, it can be shown that if $\widehat{\sigma^{(n)}}(\omega)$ were zero on a set of positive measure, then it must be identically zero. But then $\sigma(t)$ must be a polynomial, contradicting the assumption that $\sigma(t)$ is non-smooth.

***Example** 2:* If σ is one of the PyTorch functions LogSigmoid, Sigmoid, SoftPlus, Tanh, and TanShrink then there is some $n \geq 1$, and non-zero numbers a and b, $\sigma^{(n)}(t) = a\cdot\mathrm{sech}^2(bt)$. Thus, there are constants $c_0, \ldots, c_{n-1}$ such that

$$\hat{\sigma}(\omega) = \frac{1}{(j2\pi\omega)^n}2\pi\frac{a\omega}{|b|}\mathrm{csch}(\pi^2\omega/b) + \sum_{i=0}^{n-1} c_i\delta^{(i)}(\omega).$$

It can be shown that the function $\omega\mathrm{csch}(\omega)$ is infinitely differentiable and all of its derivatives have polynomial growth, so the multiplication is well-defined. Also note that $\omega\mathrm{csch}(\omega) > 0$ for all $\omega \neq 0$.

***Example** 3:* The PyTorch functions CELU, ELU, and SELU all have the form

$$\sigma(t) = a\left(e^{bt} - 1\right)\mathbb{1}(t \leq 0) + ct\mathbb{1}(t > 0),$$

with $b > 0$ and a and c are not both zero. It follows that

$$\sigma^{(2)}(t) = ab^2e^{bt}\mathbb{1}(t \leq 0) + (c - ab)\delta(t).$$

Lemma 1 shows that there are numbers, r_0 and r_1 such that

$$\hat{\sigma}(\omega) = \frac{1}{(j2\pi\omega)^2}\left(\frac{cb - j2\pi(c-ab)\omega}{b - j2\pi\omega}\right) + \sum_{i=0}^{1} r_i \delta^{(i)}(\omega).$$

Also note that the fraction is non-zero for all $\omega \neq 0$ since $b > 0$ and a and c are not both zero.

In each of the activation function examples, we have seen that the Fourier transform, $\hat{\sigma}(\omega)$, behaves like a regular function at all $\omega \neq 0$, and the corresponding regular function is continuous at all $\omega \neq 0$ and non-zero for almost all $\omega \in \mathbb{R}$.

B. An Intermediate Result on Linear Independence

This subsection presents a lemma on the linear independence of functions which will be used to prove Theorem 1 on independence of neural network activation functions.

A collection of generalized functions $g_1, \ldots, g_m$ is *linearly dependent* if there are numbers, $v_1, \ldots, v_m \in \mathbb{C}$, not all zero such that for all $f \in \mathcal{S}_n$,

$$\left\langle \sum_{i=1}^{m} v_i g_i, f \right\rangle = 0.$$

The generalized functions, $g_1, \ldots, g_m$ are *linearly independent* if they are not linearly dependent.

Say now that $g_1, \ldots, g_m$ are regular functions that have polynomial growth which are continuous almost everywhere. If we regard these functions as generalized functions, then they are linearly dependent if and only if there is a collection of numbers, $v_1, \ldots, v_m \in \mathbb{C}$, not all zero such that

$$\sum_{i=1}^{m} v_i g_i(x) = 0 \text{ for almost all } x \in \mathbb{R}^n. \tag{7}$$

This definition rules out pathologies that arise from functions for which (7) fails on a set of measure zero.

The following lemma gives a condition that can be used for checking linear independence of a collection of functions. It builds on an argument from [3] for continuous functions.

***Lemma** 2: Let $g_1, \ldots, g_m$ be a collection of functions $g_i : \mathbb{R}^n \to \mathbb{C}$ with polynomial growth that are continuous almost everywhere. The collection is linearly independent (as generalized functions) if and only if there are points $x^1, \ldots, x^m \in \mathbb{R}^n$ such that g_i are continuous at x^j for $i, j = 1, \ldots, m$ and the following matrix is invertible:*

$$M(x) = \begin{bmatrix} g_1(x^1) & \cdots & g_m(x^1) \\ \vdots & & \vdots \\ g_1(x^m) & \cdots & g_m(x^m) \end{bmatrix}.$$

Proof: Let $x = (x^1, \ldots, x^m)$ be a collection of points such that each g_i is continuous at each x^j and $M(x)$ is invertible. Thus, $M(y)$ must be invertible for all y in a neighborhood U of x.

Let v be such that (7) holds for almost all $x \in \mathbb{R}^n$ and let $\mathcal{Y}$ be the set of $x \in \mathbb{R}^n$ such that (7) holds. Then, we must have that $\mu(U \cap \mathcal{Y}^m) = \mu(U) > 0$, so that $U \cap \mathcal{Y}^m \neq \emptyset$. For any $y \in U \cap \mathcal{Y}^m$, we must have that $M(y)$ is invertible and $M(y)v = 0$. Thus, we must have that $v = 0$.

Conversely, say that $g_1, \ldots, g_m$ are linearly independent. We will show how to construct the desired $x^1, \ldots, x^m$ so that M is invertible and each g_i is continuous at each x^j. Throughout, let $\mathcal{C}$ denote the set of $x \in \mathbb{R}^n$ such that all g_i are continuous at x. We must have that $\mu(\mathbb{R}^n \setminus \mathcal{C}) = 0$.

In the case that $m = 1$, $M(x) = g_1(x^1)$. There must be a set of non-zero measure such that $M(x) \neq 0$, and so at least some points in this set must belong to $\mathcal{C}$. So, any point in this intersection suffices. So, assume that $m \geq 2$.

Fix any $0 \neq v^1 \in \mathbb{C}^m$. The set of vectors, x^1, such that

$$\underbrace{\begin{bmatrix} g_1(x^1) & \cdots & g_m(x^1) \end{bmatrix}}_{M_1} v^1 \neq 0$$

has non-zero measure. So, we can choose an $x^1 \in \mathcal{C}$ such that $M_1 v^1 = d_1 \neq 0$.

Now assume that for some $i < m$, $x^1, \ldots, x^i$ and $v^1, \ldots, v^i$ have been chosen so that

$$\underbrace{\begin{bmatrix} g_1(x^1) & \cdots & g_m(x^1) \\ \vdots & & \vdots \\ g_1(x^i) & \cdots & g_m(x^i) \end{bmatrix}}_{M_i} \underbrace{\begin{bmatrix} v^1 & \cdots & v^i \end{bmatrix}}_{V_i} = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ \star & & d_i \end{bmatrix}, \tag{8}$$

where the right side is lower triangular and invertible.

Take any $v^{i+1} \neq 0$ in the nullspace of M_i. Such a vector must exist because M_i has i rows and $i < m$.

By linear independence, the set of $x^{i+1} \in \mathbb{R}^n$ such that

$$\begin{bmatrix} g_1(x^{i+1}) & \cdots & g_m(x^{i+1}) \end{bmatrix} v^{i+1} \neq 0$$

has non-zero measure. Thus, we can take $x^{i+1} \in \mathcal{C}$ that satisfies this inequality. Then $M_{i+1}V_{i+1}$ is again lower triangular and invertible.

By induction, there are vectors $x^1, \ldots, x^m \in \mathcal{C}$ and a matrix, $V \in \mathbb{C}^{m \times m}$, such that $M(x)V$ is invertible. Since $M(x)$ is square, $M(x)$ must be invertible. ■

C. Linear Independence of Neural Network Activations

This section presents the main result of the paper. It implies that for *all* of the activation functions described in Subsection II-A, the functions formed by randomly generating weights and biases will be linearly independent with probability 1 whenever the support of the distributions have positive Lebesgue measure.

***Theorem** 1: Let $\sigma : \mathbb{R} \to \mathbb{C}$ be a regular function with polynomial growth which is continuous almost everywhere. Assume there is a regular function h which is continuous almost everywhere and non-zero almost everywhere, and there is a number $n \geq 1$ such that for all $f \in \mathcal{S}$ with $f(0) = f^{(1)}(0) = \cdots = f^{(n-1)}(0)$, we have that*

$$\langle \hat{\sigma}, f \rangle = \int_{-\infty}^{\infty} h(x) f(x) dx.$$

Then, for almost all $w_1, \ldots, w_m \in \mathbb{R}^n$ and almost all $b_1, \ldots, b_m \in \mathbb{R}$, the functions defined by $\sigma(w_1^\top x + b_1), \ldots, \sigma(w_m^\top x + b_m)$ are linearly independent.

Proof: Denote the coordinates of w_i by w_{ij}. Assume that $w_{i1} \neq 0$, which holds almost everywhere. Let $\ell_i(x_1)$ be

the functions of x_1 defined by $\ell_i(x_1) = \sigma(w_i^\top x + b_1)$. For almost all $\zeta \in \mathbb{R}$, $\hat{\ell}_i(\zeta)$ behaves like:

$$\hat{\ell}_i(\zeta) = \frac{1}{|w_{i1}|} \exp\left(j2\pi \frac{\zeta\left(b_i + \sum_{k=2}^n w_{ik}x_k\right)}{w_{i1}}\right) h\left(\frac{\zeta}{w_{i1}}\right)$$
$$=: f_i(y), \text{ where } y = \begin{bmatrix}\zeta & x_2 & \cdots & x_m\end{bmatrix}^\top.$$

Note that if $f_1, \ldots, f_m$ are linearly independent, then the original functions $\sigma(w_1^\top x + b_1), \ldots, \sigma(w_m^\top x + b_m)$ must be linearly independent. Indeed, taking Fourier transforms of $\sum_{i=1}^m v_i \sigma(w_i^\top x + b_i) = 0$ with respect to x_1 shows that if the original functions are linearly dependent, then so are $f_1, \ldots, f_m$.

So, to prove the theorem, we will show that $f_1, \ldots, f_m$ are linearly independent by application of Lemma 2.

Note that each f_i is continuous and non-zero at almost all y. Indeed, the exponential function is continuous and non-zero at all y, and the h is continuous and non-zero at almost all ζ by assumption.

Let $y_1, \ldots, y_m \in \mathbb{R}^n$, with entries $y_k = \begin{bmatrix}\zeta_k & x_{k2} & \cdots & x_{km}\end{bmatrix}^\top$. We will give conditions on y_k that ensure that the associated matrix with entries $M_{ki} = f_i(y_k)$ is invertible.

Set $c_{ki} = j2\pi\frac{\zeta_k}{w_i}$ and

$$z_{ki} = \frac{1}{|w_{i1}|} \exp\left(j2\pi \frac{\zeta_k\left(b_i + \sum_{p=2}^n w_{ip}x_{kp}\right)}{w_{i1}}\right) h\left(\frac{\zeta_k}{w_{i1}}\right)$$

We set the terms x_{kp} arbitrarily, and we choose ζ_k such that $h\left(\frac{\zeta_k}{w_{i1}}\right) \neq 0$ for all $i, k = 1, \ldots, m$ and all c_{ki} are unique and non-zero. To see that almost all ζ_k satisfy these assumptions, note that $h\left(\frac{\zeta}{w_{i1}}\right) \neq 0$ for almost all $\zeta \in \mathbb{R}$. Any $\bar{\zeta} = \begin{bmatrix}\zeta_1 & \cdots & \zeta_m\end{bmatrix}^\top$ such that $\zeta_i \neq 0$ and $\frac{\zeta_k}{w_{i1}} \neq \frac{\zeta_p}{w_{q1}}$ for all i, k, p, q. Each of these inequalities hold for almost all $\bar{\zeta}$. Thus, almost any choice of $y_1, \ldots, y_m$ leads to non-zero z_{ki}, non-zero c_{ki}, distinct c_{ki}, and continuity of all f_i at y_k. For the rest of the proof, assume that such $y_1, \ldots, y_m$ have been chosen.

In the notation above, we can write M from Lemma 2 as

$$M(b, c, z) = \begin{bmatrix} z_{11}e^{c_{11}b_1} & \cdots & z_{1m}e^{c_{1m}b_m} \\ \vdots & & \vdots \\ z_{m1}e^{c_{m1}b_1} & \cdots & z_{mm}e^{c_{mm}b_m} \end{bmatrix}.$$

We will prove that $M(b, c, z)$ is invertible for almost all $b \in \mathbb{R}^m$ via induction on m.

For $m = 1$, $M(b, c, z) = z_{11}e^{c_{11}b_1} \neq 0$ for all b_1.

Assume inductively that $M(b, c, z)$ is invertible for almost all $b \in \mathbb{R}^{m-1}$, provided that the entries of $c \in \mathbb{C}^{(m-1)\times(m-1)}$ and $z \in \mathbb{C}^{(m-1)\times(m-1)}$ are all non-zero and distinct.

Now consider $M(b, c, z) \in \mathbb{C}^{m\times m}$. The Laplace expansion formula implies that:

$$\det M(b, c, z) = \sum_{i=1}^m (-1)^{i+1} e^{c_{i1}b_1} z_{i1} \det M(\tilde{b}, \tilde{c}^i, \tilde{z}^i).$$

Here, $\tilde{b} = \begin{bmatrix}b_2 & \cdots & b_m\end{bmatrix}^\top$, $\tilde{c}^i$ corresponds to the matrix c with column 1 and row i removed, and $\tilde{z}^i$ corresponds to the matrix z with column 1 and row i removed. By the inductive assumption, $\det M(\tilde{b}, \tilde{c}^i, \tilde{z}^i) \neq 0$ for almost all $\tilde{b}$. So, fix $\tilde{b}$ such that all of these numbers are non-zero.

Let $v_i = (-1)^{i+1} z_{i1} \det M(\tilde{b}, \tilde{c}^i, \tilde{z}^i) \neq 0$, $r_i = c_{i1}$, and $t = b_1$. In this notation, we have that

$$g(t) = \sum_{i=1}^m e^{r_i t} v_i = \det M(b, c, z).$$

The proof will be completed by showing that $g(t) \neq 0$ for almost all $t \in \mathbb{R}$.

Let $\mathcal{Y} = \{t \in \mathbb{R} | g(t) = 0\}$. Assume for the sake of contradiction that $\mu(\mathcal{Y}) > 0$. There must be a compact subset $K \subset \mathcal{Y}$ such that $\mu(K) > 0$. See Theorem 2.17 of [15]. The set K, and thus $\mathcal{Y}$, must have an accumulation point. Otherwise, there is an open cover of K consisting of disjoint open sets around each of its elements. But then compactness implies that this cover has a finite subcover, and so K must be a finite set, contradicting that $\mu(K) > 0$.

Note that g can be extended to analytic function $g : \mathbb{C} \to \mathbb{C}$. Since $g(t) = 0$ for all $t \in \mathcal{Y}$ and $\mathcal{Y}$ has an accumulation point, g must be identically zero. See Section 3.2 of [16].

So, if g is identically zero and $v_i \neq 0$, it must be that the functions $e^{r_i t}$ are linearly dependent. However, this contradicts a classical result that exponential functions with distinct r_i are linearly independent. ■

A case of Theorem 1 with $n = 1$ and $w_i = 1$ for all i is sketched in [8].

Remark 1: A similar linear independence claim for $\sigma = \frac{1}{1+e^{-t}}$ is made in [17] based on an incomplete argument from [9]. It is claimed in [9] that differentiability would imply that if functions were linearly dependent, then

$$\sum_{i=1}^m v_i w_i^k \sigma^{(k)}(w_i t + b_i) = 0$$

for all $t \geq 0$ and all $k \geq 0$. They claim that no non-zero solution for v_i can exist, since there are an infinite number of equations and only m unknowns. However, they did not prove that these equations cannot share a common null-space.

Remark 2: The proof of Theorem 1 indicates that the points chosen to make $M(x)$ invertible can be chosen as $x^i = \begin{bmatrix}x_1^i & z_2 & \cdots & z_m\end{bmatrix}^\top$. Thus, if $i \neq j$, then x^i and x^j only differ in their first coordinate. By first applying a change of basis to x, then applying the manipulations in the proof of Theorem 1, we see that $x^1, \ldots, x^m$ can be chosen along an arbitrary line in $\mathbb{R}^n$. This fact will be utilized when deriving sufficient conditions for persistence of excitation below.

Remark 3: Theorem 1 does not readily extend to multi-layer networks, and linear independence will fail with positive probability for multi-layer ReLU networks with weights and biases chosen via Gaussians.

III. Application to Persistence of Excitation in Control

A. General Result

Adaptive control is often applied to systems of the form:

$$\frac{dx_t}{dt} = f(x_t) + h(x_t)u_t + \Theta\Phi(x_t),$$

where f and h are known functions, x_t is the state, u_t is the input, $\Phi(x) = \begin{bmatrix} g_1(x) & \cdots & g_m(x) \end{bmatrix}^\top$, $g_i : \mathbb{R}^n \to \mathbb{R}$, and Θ is a matrix of unknown parameters.

Adaptive control gives methods for simultaneously estimating Θ and choosing inputs in control the system. In order to estimate Θ, the state needs to satisfy a condition known as persistence of excitation [18].

A trajectory $x : [0, \infty) \to \mathbb{R}^n$ is called *persistently exciting* if there is a time horizon T and constants $0 < \alpha \leq \beta$ such that for any $t \geq 0$:

$$\alpha I \preceq \int_t^{T+t} \Phi(x_\tau)\Phi(x_\tau)^\top d\tau \preceq \beta I.$$

Here $\succeq$ denotes the semidefinite partial ordering.

In many cases, $\Phi(x_t)$ is bounded and so an upper bound must exist. The lower bound is more of a challenge and few explicit sufficient conditions for persistence of excitation exist in the literature. The most famous condition, related to the notion of *sufficient richness*, arises when $\Phi(x_t)$ is the output of a collection of m linear time-invariant filters driven by an input with non-zero energy on at least $m/2$ distinct frequencies. See [19]–[23] for details and extensions.

The sufficient conditions for persistence of excitation described above do not cover the common case that x_t is the state of a dynamic system and $g_i(x) = \sigma(x^\top w_i + b_i)$ come from a neural network. The proposition below implies that in this case, persistence of excitation will hold, provided that g_i are linearly independent and x_t traverses through specific regions of the state space.

***Proposition** 1: Assume that $g_1, \ldots, g_m$ with $g_i : \mathbb{R}^n \to \mathbb{R}$ are continuous almost everywhere, linearly independent functions. Let $\Phi(x) = \begin{bmatrix} g_1(x) & \cdots & g_m(x) \end{bmatrix}^\top$. Fix any $c > 0$. There are vectors $x^1, \ldots, x^m \in \mathbb{R}^n$ and positive numbers ϵ and δ such that the following holds. Let $x_t \in \mathbb{R}^n$ be any trajectory over $[\tau, \tau + T]$ such that:*

- *$\|\dot{x}_t\| \leq c$ for all $t \in [\tau, \tau + T]$*
- *There are times, $t_i \in [\tau, \tau + T]$ such that $\|x_{t_i} - x^i\| \leq \epsilon$ for $i = 1, \ldots, m$.*

Then,

$$\int_\tau^{\tau+T} \Phi(x_t)\Phi(x_t)^\top dt \succeq \delta I \tag{9}$$

Proof: Let $x^1, \ldots, x^m \in \mathbb{R}^n$ be the vectors from Lemma 2 and let M be the corresponding matrix. In this notation, we have that $M^\top = \begin{bmatrix} \Phi(x^1) & \cdots & \Phi(x^m) \end{bmatrix}$ and

$$M^\top M = \sum_{i=1}^m \Phi(x^i)\Phi(x^i)^\top \succeq \sigma_{\min}(M)^2 I,$$

where $\sigma_{\min}(M) > 0$ is the minimum singular value of M.

Continuity of g_j functions at the x^i values implies that there is a radius $r > 0$ such that if $\|y^i - x^i\| \leq r$ for $i = 1, \ldots, m$, then

$$\sum_{i=1}^m \Phi(y^i)\Phi(y^i)^\top \succeq \frac{1}{2}\sigma_{\min}(M)^2 I.$$

For $0 < a < T/2$, let $\mathcal{T}_i \subset [\tau, \tau + T]$ be an interval of length a containing t_i. The assumption that $0 < a < T/2$ implies that such intervals must exist. By construction, $|t - t_i| \leq a$ for all $t \in \mathcal{T}_i$.

The speed bound implies that $\|x_t - x_{t_i}\| \leq c|t - t_i|$. Thus, for all $t \in \mathcal{T}_i$, we have that

$$\|x_t - x^i\| \leq \|x_t - x_{t_i}\| + \|x_{t_i} - x^i\| \leq ca + \epsilon. \tag{10}$$

In particular, if $ca + \epsilon \leq r$, we have that $\|x_t - x^i\| \leq r$.

We now have that:

$$\begin{aligned}\int_\tau^{\tau+T} \Phi(x_t)\Phi(x_t)^\top dt &\succeq \sum_{i=1}^m \int_{\mathcal{T}_i} \Phi(x_t)\Phi(x_t)^\top dt \\ &\succeq \frac{a}{2}\sigma_{\min}(M)^2 I.\end{aligned}$$

To prove the second inequality, we do Riemann sum approximations of the integral as follows. For any integer $N > 0$, let $h = a/N$ and let $\tau_{i,0}, \ldots, \tau_{i,N}$ be a uniform gridding of $\mathcal{T}_i$ with $\tau_{i,j+1} - \tau_{i,j} = h$.

$$\begin{aligned}\sum_{i=1}^m \int_{\mathcal{T}_i} \Phi(x_t)\Phi(x_t)^\top dt &\approx \sum_{i=1}^m \sum_{j=0}^{N-1} h\Phi(x_{\tau_{i,j}})\Phi(x_{\tau_{i,j}})^\top \\ &\succeq hN\frac{1}{2}\sigma_{\min}(M)^2 I.\end{aligned}$$

The inequality follows because $hN = a$ and the Riemman sum converges to the integral as $N \to \infty$.

So, we finish the proof by setting $\delta = \frac{a}{2}\sigma_{\min}(M)^2$ where a and ϵ are any positive numbers such that $ca + \epsilon \leq r$. ■

***Corollary** 1: Let σ satisfy the conditions of Theorem 1. Let $y \in \mathbb{R}^n$ and $0 \neq z \in \mathbb{R}^n$ be arbitrary vectors and assume that $g_i(x) = \sigma(w_i^\top x + b_i)$ for $i = 1, \ldots, m$. For almost all choices of $w_1, \ldots, w_m$ and $b_1, \ldots, b_m$ there are numbers $p < q$ such that x^i can be chosen to have the form $x^i = y + \alpha_i z$ for $\alpha_i \in [p, q]$. Consequently, as long as x_t traverses the line segment $\{y + \alpha z | p \leq \alpha \leq q\}$ on each interval of the form $[\tau, \tau + T]$, it is persistently exciting.*

Proof: This follows from a combination of Proposition 1, Theorem 1, and Remark 2. ■

***Remark** 4:* For ReLUs and step functions, explicit constructions of $x^1, \ldots, x^m$ can be given in terms of polyhedra. (This is discussed in another paper at this conference [24].)

B. Discussion on Reference Tracking

Ensuring the conditions of Proposition 1 or Corollary 1 may be impossible without precise control of x_t. Here we describe a scenario in which asymptotic estimation of the parameters can be guaranteed, even if the persistence of excitation conditions do not hold precisely for the state.

A common goal in adaptive control is make the state of the system track the state of a known reference system:

$$\frac{d\hat{x}_t}{dt} = A\hat{x}_t + Br_t,$$

where $\hat{x}_t$ is the reference state, r_t is the reference input, and A and B are known matrices.

It is shown in [25] that if $\lim_{t\to\infty}(x_t - \hat{x}_t) = 0$ and $\hat{x}_t$ is persistently exciting, then x_t inherits the persistence of excitation properties of $\hat{x}_t$. Building upon this insight gives:

***Corollary** 2: Let $x^1, \ldots, x^m$ be the vectors from Proposition 1 and let c, ϵ and δ be the corresponding numbers. Assume that $\left\| \frac{d\hat{x}(t)}{dt} \right\| \leq c$ for all $t \geq 0$ and for all $\tau \geq 0$, there are $t_i \in [\tau, \tau + T]$ such that $\|\hat{x}_{t_i} - x^i\| \leq \epsilon/2$ for $i = 1, \ldots, m$. If $\lim_{t\to\infty}(x_t - \hat{x}_t) = 0$, then (9) holds for x_t for all sufficiently large τ.*

Proof: Let $\mathcal{T}_i$ be the intervals constructed in the proof of Proposition 1, applied to $\hat{x}_t$. Then for t sufficiently large we have that $\|x_t - \hat{x}_t\| \leq \frac{\epsilon}{2}$, so the triangle inequality gives:

$$\begin{aligned}\|x_t - x^i\| &\leq \|x_t - \hat{x}_t\| + \|\hat{x}_t - \hat{x}_{t_i}\| + \|\hat{x}_{t_i} - x^i\| \\ &\leq \frac{\epsilon}{2} + ca + \frac{\epsilon}{2} = ca + \epsilon \leq r.\end{aligned}$$

Note that this bound recovers (10). The steps of rest of the proof of Proposition 1 can be followed. ∎

IV. Conclusion

This paper proves that neural networks with a single hidden layer give rise to linearly independent functions for nearly all choices of weights and biases in the hidden layer. Such properties are tacitly assumed in many works on adaptive control and function approximation. Future work will investigate applications to network identifiability and function approximation, and also aim to get more explicit conditions for persistence of excitation.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] E. Lavretsky and K. A. Wise, *Robust and Adaptive Control: with Aerospace Applications*. Springer, 2013.

[3] A. Pinkus, "Approximation theory of the mlp model in neural networks," *Acta numerica*, vol. 8, pp. 143–195, 1999.

[4] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[6] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized hamilton-jacobi-bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[7] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. IET, 2013, vol. 2.

[8] F. Albertini, E. D. Sontag, and V. Maillot, "Uniqueness of weights for neural networks," *Artificial Neural Networks for Speech and Vision*, pp. 115–125, 1993.

[9] S. Tamura and M. Tateishi, "Capabilities of a four-layered feedforward neural network: Four layers versus three," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 251–255, 1997.

[10] C. Yun, S. Sra, and A. Jadbabaie, "Small relu networks are powerful memorizers: A tight analysis of memorization capacity," *arXiv preprint arXiv:1810.07770*, 2018.

[11] V. Vlačić and H. Bölcskei, "Neural network identifiability for a family of sigmoidal nonlinearities," *Constructive Approximation*, pp. 1–52, 2021.

[12] C. Fefferman and S. Markel, "Recovering a feed-forward net from its output," in *Advances in Neural Information Processing Systems*, 1994, pp. 335–342.

[13] D. W. Kammler, *A first course in Fourier analysis*. Cambridge University Press, 2007.

[14] W. Rudin, *Functional Analysis*, Second. McGraw-Hill, 1991.

[15] ——, *Real and Complex Analysis*, Third. McGraw-Hill, 1987.

[16] L. V. Ahlfors, *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable*, Third. McGraw-Hill, Inc., 1979.

[17] G.-B. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," *IEEE transactions on neural networks*, vol. 14, no. 2, pp. 274–281, 2003.

[18] S. Sastry and M. Bodson, *Adaptive control: stability, convergence and robustness*. Prentice Hall, Inc., 1989.

[19] P. A. Ioannou and J. Sun, *Robust adaptive control*. Courier Corporation, 2012.

[20] I. M. Mareels and M. Gevers, "Persistency of excitation criteria for linear, multivariable, time-varying systems," *Mathematics of Control, Signals and Systems*, vol. 1, no. 3, pp. 203–226, 1988.

[21] G. Kreisselmeier and G. Rietze-Augst, "Richness and excitation on an interval-with application to continuous-time adaptive control," *IEEE transactions on automatic control*, vol. 35, no. 2, pp. 165–171, 1990.

[22] J.-S. Lin and I. Kanellakopoulos, "Nonlinearities enhance parameter convergence in strict feedback systems," *IEEE Transactions on Automatic Control*, vol. 44, no. 1, pp. 89–94, 1999.

[23] P. Karg, F. Köpf, C. A. Braun, and S. Hohmann, "Excitation for adaptive optimal control of nonlinear systems in differential games," *arXiv preprint arXiv:2105.02260*, 2021.

[24] T. Lekang and A. Lamperski, "Sufficient conditions for persistency of excitation with step and relu activation functions," in *IEEE Conference on Decision and Control (CDC)*, 2022.

[25] E. Panteley, A. Loria, and A. Teel, "Relaxed persistency of excitation for uniform asymptotic stability," *IEEE Transactions on Automatic Control*, vol. 46, no. 12, pp. 1874–1886, 2001.