2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

# Approximation with Random Shallow ReLU Networks with Applications to Model Reference Adaptive Control

Andrew Lamperski and Tyler Lekang

2024 IEEE 63rd Conference on Decision and Control (CDC) | 979-8-3503-1633-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/CDC56724.2024.10886783

*Abstract*— **Neural networks are regularly employed in adaptive control of nonlinear systems and related methods of reinforcement learning. A common architecture uses a neural network with a single hidden layer (i.e. a shallow network), in which the weights and biases are fixed in advance and only the output layer is trained. While classical results show that there exist neural networks of this type that can approximate arbitrary continuous functions, they are non-constructive, and the networks used in practice have no approximation guarantees. Thus, the approximation properties required for control with neural networks are assumed, rather than proved. In this paper, we aim to fill this gap by showing that for sufficiently smooth functions, ReLU networks with randomly generated weights and biases achieve $L_\infty$ error of $O(m^{-1/2})$ with high probability, where $m$ is the number of neurons. We show how the result can be used to construct approximators of required accuracy in a model reference adaptive control application.**

## I. Introduction

Neural networks have wide applications in control systems, particularly for nonlinear systems with unknown dynamics. In adaptive control they are commonly used to model unknown nonlinearities [1]. In reinforcement learning and dynamic programming, they are used to approximate value functions and to parameterize control strategies [2]–[4].

A theoretical gap arises in the current use of neural networks in adaptive control and reinforcement learning, since the approximation properties are *assumed* rather than proved [2], [5]–[11]. See [12], [13] for discussion. The underlying problem is to use a neural network of the form $\Theta\Phi(Wx+b)$, where $W$ are weights, $b$ are biases, $\Phi$ is a vector of nonlinear functions, and $\Theta$ is a matrix of output parameters, to approximate an unknown function $f(x)$. The specific assumption in the cited work is that $(W,b)$ have been chosen such that $\inf_\Theta \sup_{x\in B} \|f(x)-\Theta\Phi(Wx+b)\|$ is small. While suitable $(W,b)$ are known to exist (see [14]), there has been no practical means to compute them or verify that the deployed $(W,b)$ satisfy the requirements.

The main result of this paper shows that if $\mathbf{W}$ and $\mathbf{b}$ are chosen randomly, then for any smooth $f$, $\inf_\Theta \sup_{x\in B} \|f(x)-\Theta\Phi(\mathbf{W}x+\mathbf{b})\| = O(m^{-1/2})$ holds with high probability, where $m$ is the number of neurons. Here $\Phi$ is constructed from ReLU activation functions and affine terms. This gives a simple algorithm to generate $(W,b)$ which satisfy the required approximation properties by construction.

To prove our approximation theorem, we derive a new integral representation theorem for ReLU activations over bounded domains. Similar integral representations are commonly employed in constructive approximation theory for neural networks [15]–[18]. The advantage of our new integral representation is that the integrand can be precisely bounded.

We apply our approximation method to get guaranteed performance for a neural-network-based controller. In particular, quantify the number of neurons sufficient to achieve the required accuracy for a control algorithm from [1].

Over the last several years, the theoretical properties of neural networks with random initializations have been studied extensively. Well-known results show that as the width of a randomly initialized neural network increases, the behavior approaches a Gaussian process [19]–[21]. Related work shows that with sufficient width, [22], [23], gradient descent reaches near global minima from random initializations.

The closest work on approximation is [24]. In comparison with [24], our error bound is substantially simpler and more explicit. (The error from [24] is a complex expression with unquantified constants.) Additionally, we bound the $L_\infty$ error, which is commonly required in control, while [24] bounds the $L_2$ error. The work in [24] has the advantage of applying to a broader class of functions, and also includes lower bounds that match the achievable approximation error.

Other closely related research includes [25], [26]. In [25], it is shown that learning a non-smooth function with a randomized ReLU network requires a large number of neurons. (We approximate smooth functions in this paper.) Lower bounds on achievable errors for a different class of randomized single-hidden-layer networks are given in [26].

Related work by the authors includes [27], which gives sufficient conditions for persistency of excitation of neural network approximators, and [28], which shows that shallow neural networks with randomly generated weights and biases define linearly independent basis functions. Persistency of excitation and linear independence are commonly assumed without proof in the adaptive control literature.

The paper is organized as follows. Section II presents preliminary notation. Section III gives the main result on approximation. Section IV presents an application to Model Reference Adaptive Control. Section V gives conclusions.

## II. Notation

We use $\mathbb{R}, \mathbb{C}$ to denote the real and complex numbers. Random variables are denoted as bold symbols, e.g. $\boldsymbol{x}$. $\mathbb{E}[\boldsymbol{x}]$ denotes the expected value of $\boldsymbol{x}$ and $\mathbb{P}(\boldsymbol{A})$ denotes the probability of event $\boldsymbol{A}$. $B(R) \subset \mathbb{R}^n$ denotes the radius $R$

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

T. Lekang is with Honeywell Aerospace, Plymouth, MN 55441, USA tylerlekang@gmail.com

Euclidean ball centered at $0$. The Euclidean norm is denoted $\|w\|$, while if $M$ is a matrix, then $\|M\|$ denotes the induced 2-norm. If $f$ is a complex-valued function, and $p \in [1, \infty]$, $\|f\|_p$ denotes the corresponding $L_p$ norm.

## III. Approximation by Randomized ReLUs

This section gives our main technical result, which shows that all sufficiently smooth functions can be approximated by an affine function and a single-hidden-layer neural network with ReLU activations, where the weights and biases are generated randomly. The worst-case error over a compact set decays like $O(m^{-1/2})$, where $m$ is the number of neurons.

### A. Background

If $f : \mathbb{R}^n \to \mathbb{C}$, its Fourier transform $\hat{f} : \mathbb{R}^n \to \mathbb{C}$ satisfies

$$\hat{f}(\omega) = \int_{\mathbb{R}^n} e^{-j2\pi\omega^\top x} f(x) dx$$
$$f(x) = \int_{\mathbb{R}^n} e^{j2\pi\omega^\top x} \hat{f}(\omega) d\omega.$$

Let $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n | \|x\| = 1\}$ denote the $(n-1)$-dimensional unit sphere. Let $\mu_{n-1}$ be the area measure over $\mathbb{S}^{n-1}$, with $\mu_0$ the counting measure. The area of $\mathbb{S}^{n-1}$ is

$$A_{n-1} := \int_{\mathbb{S}^{n-1}} \mu_{n-1}(d\alpha) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \tag{1}$$

where $\Gamma$ is the gamma function. The area is maximized at $n = 7$, and decreases geometrically with $n$.

Let $\sigma$ denote the ReLU activation function:

$$\sigma(t) = \max\{0, t\}. \tag{2}$$

### B. Approximation with Random ReLU Networks

Our approximation result below holds for functions $f : \mathbb{R}^n \to \mathbb{R}$ which satisfy the following smoothness assumption:

*Assumption 1:* There exists $k \geqslant n + 3$ and $\rho > 0$ such that $\sup_{\omega \in \mathbb{R}^n} |\hat{f}(\omega)|(1 + \|\omega\|^k) \leqslant \rho$.

This assumption implies, in particular, that $f$ and all of its derivatives up to order $k - 2$ are bounded.

Our main technical result is stated below. It is proved in Subsections III-C, III-D, and III-E.

*Theorem 1: Let $R > 0$ and let $m$ and $n$ be positive integer. Let $P$ be a probability density function over $\mathbb{S}^{n-1} \times [-R, R]$ with $\inf_{(\alpha,t) \in \mathbb{S}^{n-1} \times [-R,R]} P(\alpha, t) = P_{\min} > 0$. Let $(\boldsymbol{\alpha}_1, \boldsymbol{t}_1), \dots, (\boldsymbol{\alpha}_m, \boldsymbol{t}_m)$ be independent, identically distributed samples from $P$. If $f$ satisfies Assumption (1), then there is a vector $a \in \mathbb{R}^n$, a number $b \in \mathbb{R}$, and coefficients $\boldsymbol{c}_1, \dots, \boldsymbol{c}_m$ with*

$$\|a\| \leqslant 4\pi A_{n-1}\rho$$
$$|b| \leqslant (1 + (2\pi R)) A_{n-1}\rho$$
$$|\boldsymbol{c}_i| \leqslant \frac{8\pi^2\rho}{m P_{\min}}$$

*such that for all $\nu \in (0, 1)$, with probability at least $1 - \nu$, the neural network approximation*

$$\boldsymbol{f}_N(x) = a^\top x + b + \sum_{i=1}^{m} \boldsymbol{c}_i \sigma(\boldsymbol{\alpha}_i^\top x - \boldsymbol{t}_i) \tag{3}$$

*satisfies*

$$\sup_{x \in B(R)} |\boldsymbol{f}_N(x) - f(x)| \leqslant \frac{1}{\sqrt{m}} \left(\kappa_0 + \kappa_1 \sqrt{\log(4/\nu)}\right).$$

*Here*

$$\kappa_0 = 800 n^{1/2} \pi^{5/2} R\rho \left(\frac{\pi}{P_{\min}} + A_{n-1}\right)$$
$$\kappa_1 = \frac{264\pi^2\rho R}{P_{\min}} + \rho A_{n-1}(4 + 256 R\pi).$$

The uniform distribution has $P(\alpha, t) = \frac{1}{2RA_{n-1}}$, so that $\frac{1}{P_{\min}} = 2RA_{n-1}$. In this case the bounds simplify:

*Corollary 1: If $P$ is the uniform distribution over $\mathbb{S}^{n-1} \times [-R, R]$, then the coefficients depending on $P_{\min}$ satisfy:*

$$|\boldsymbol{c}_i| \leqslant \frac{16\pi^2}{m} \rho A_{n-1}$$
$$\kappa_0 \leqslant 800 n^{1/2} \pi^{5/2} R\rho A_{n-1}(1 + 2\pi R)$$
$$\kappa_1 \leqslant \rho A_{n-1}\left(528(\pi R)^2 + 256(\pi R) + 4\right).$$

*Remark 1:* The coefficient bounds depend on $\rho A_{n-1}$. More work is needed to quantify $\rho A_{n-1}$ in practice.

Theorem 1 implies that randomly generated weights and biases suffice to approximate smooth functions by appropriate choice of output coefficients $a$, $b$, and $\boldsymbol{c}_i$.

Samples over $\mathbb{S}^{n-1}$ can be generated by normalizing samples over $\mathbb{R}^n$. In particular, uniform samples can be generated by normalizing standard Gaussians.

### C. Integral Representation via the ReLU Activation

Here we derive an integral representation for smooth real-valued functions. Related representations were derived in [15], [16]. The main advantage of our representation is the explicit bound on the integrand.

*Lemma 1: Let $f$ be a real-valued function satisfying Assumption 1. There is a function $g : \mathbb{S}^{n-1} \times [-R, R] \to \mathbb{R}$ such that $\|g\|_\infty \leqslant 8\pi^2\rho$, a vector $a \in \mathbb{R}^n$ with $\|a\| \leqslant 4\pi\rho A_{n-1}$, and a scalar $b \in \mathbb{R}$ with $|b| \leqslant (2 + 4\pi R)\rho A_{n-1}$ such that for all $\|x\| \leqslant R$*

$$f(x) = \int_{\mathbb{S}^{n-1}} \int_{-R}^{R} g(\alpha, t)\sigma(\alpha^\top x - t) dt \mu_{n-1}(d\alpha) + a^\top x + b. \tag{4}$$

*Proof:* First, note that for all $n \geqslant 1$, we can express $\omega = r\alpha$ for $r \geqslant 0$ and $\alpha \in \mathbb{S}^{n-1}$. Then, the volume element satisfies $d\omega = r^{n-1} dr \mu_{n-1}(d\alpha)$. For $n \geqslant 2$, this follows from the $n$-dimensional spherical coordinate representation from [29], while for $n = 1$, it is direct calculation since $\mu_0$ is the counting measure.

Assumption 1 implies that for $i = 0, 1, 2$:

$$\begin{aligned}\int_{\mathbb{R}^n} |\hat{f}(\omega)| \|2\pi\omega\|^i d\omega &\leqslant (2\pi)^i \rho \int_{\mathbb{R}^n} \frac{\|\omega\|^i}{1 + \|\omega\|^k} d\omega \\ &= (2\pi)^i \rho \int_{\mathbb{S}^{n-1}} \int_0^\infty \frac{r^{n+i-1}}{1 + r^k} dr \mu_{n-1}(d\alpha) \\ &\leqslant 2(2\pi)^i \rho A_{n-1},\end{aligned} \tag{5}$$

where $A_{n-1}$ is the area of $\mathbb{S}^{n-1}$, from (1). Thus,

$$\|\hat{f}\|_1 \leqslant 2\rho A_{n-1} \tag{6a}$$

$$\int_{\mathbb{R}^n} |\hat{f}(\omega)|\|2\pi\omega\| d\omega \leqslant 2(2\pi)\rho A_{n-1} \tag{6b}$$

$$Z := \int_{\mathbb{R}^n} |\hat{f}(\omega)|\|2\pi\omega\|^2 d\omega \leqslant 2(2\pi)^2 \rho A_{n-1}. \tag{6c}$$

Set $p(\omega) := \frac{1}{Z}|\hat{f}(\omega)|\|2\pi\omega\|^2$, which is a probability density over $\mathbb{R}^n$ with $p(0) = 0$.

Let $\hat{f}(\omega) = e^{j2\pi\theta(\omega)}|\hat{f}(\omega)|$ be the magnitude and phase representation of the Fourier transform. Then

$$\begin{aligned} f(x) &= \int_{\mathbb{R}^n} e^{j2\pi(\omega^\top x + \theta(\omega))}|\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^n} \frac{Z}{\|2\pi\omega\|^2} \cos(2\pi(\omega^\top x + \theta(\omega))p(\omega) d\omega. \end{aligned} \tag{7}$$

The second equality uses that $f$ is real-valued.

Define $\psi : \mathbb{R} \times (\mathbb{R}^n \backslash \{0\}) \to \mathbb{R}$ by

$$\psi(t, \omega) = \frac{Z}{\|2\pi\omega\|^2} \cos(2\pi(\|\omega\| t + \theta(\omega))). \tag{8}$$

Direct calculation shows that for all $y \in [-R, R]$

$$\begin{aligned} \psi(y, \omega) = \int_{-R}^{R} &\frac{\partial^2 \psi(t, \omega)}{\partial t^2} \sigma(y - t) dt \\ &+ \frac{\partial \psi(-R, \omega)}{\partial t}(y + R) + \psi(-R, \omega). \end{aligned} \tag{9}$$

A related identity was used in [30], but not written explicitly.

Combining (7), (8), and (9) with $y = \left(\frac{\omega}{\|\omega\|}\right)^\top x$ gives:

$$\begin{aligned} f(x) = &\left( \int_{\mathbb{R}^n} \frac{\partial \psi(-R, \omega)}{\partial t} \frac{\omega}{\|\omega\|} p(\omega) d\omega \right)^\top x \\ &+ \int_{\mathbb{R}^n} \left( \frac{\partial \psi(-R, \omega)}{\partial t} R + \psi(-R, \omega) \right) p(\omega) d\omega \\ &+ \int_{\mathbb{R}^n} \int_{-R}^{R} \frac{\partial^2 \psi(t, \omega)}{\partial t^2} \sigma\left( \left( \frac{\omega}{\|\omega\|} \right)^\top x - t \right) p(\omega) dt d\omega \end{aligned} \tag{10}$$

for $\|x\| \leqslant R$. The first two lines define $a$ and $b$.

The required derivatives of $\psi$ are given by:

$$\begin{aligned} \frac{\partial \psi(t, \omega)}{\partial t} &= -\frac{Z}{\|2\pi\omega\|} \sin(2\pi(\|\omega\| t + \theta(\omega))) \\ \frac{\partial^2 \psi(t, \omega)}{\partial t^2} &= -Z \cos(2\pi(\|\omega\| t + \theta(\omega))). \end{aligned}$$

The bounds on $\|a\|$ and $|b|$ now follow from (6).

The double integral in (10) converges absolutely because $\left|\frac{\partial^2 \psi(t,\omega)}{\partial t^2}\right| \leqslant Z$. Thus, Fubini's theorem implies the order of integration can be switched.

Now we derive the expression for $g$ from (4). Let $\omega = r\alpha$ with $r \geqslant 0$ and $\alpha \in \mathbb{S}^{n-1}$. Since $p(\omega)$ is a probability density over $\mathbb{R}^n$, $p(\omega)d\omega = p(r\alpha)r^{n-1}dr\mu_{n-1}(d\alpha)$ defines a probability measure over $[0, \infty) \times \mathbb{S}^{n-1}$. Bayes rule then implies that we can factorize $p(r\alpha)r^{n-1} = q(r|\alpha)q(\alpha)$, where $q(\cdot|\alpha)$ is a conditional density, and

$$q(\alpha) = \int_0^\infty p(r\alpha) r^{n-1} dr. \tag{11}$$

We can now express the third term on the right of (10) as:

$$\int_{-R}^{R} \int_{\mathbb{S}^{n-1}} g(\alpha, t)\sigma(\alpha^\top x - t)\mu_{n-1}(d\alpha) dr$$

where

$$g(\alpha, t) = q(\alpha) \int_0^\infty \frac{\partial^2 \psi(t, r\alpha)}{\partial t^2} q(r|\alpha) dr.$$

The bound $\left\|\frac{\partial^2 \psi}{\partial t^2}\right\|_\infty \leqslant Z$ and the fact that $q(\cdot|\alpha)$ is a probability density over $[0, \infty)$ shows that $\|g\|_\infty \leqslant Z\|q\|_\infty$. The proof is completed by bounding $\|q\|_\infty$:

$$\begin{aligned} 0 \leqslant q(\alpha) &= \int_0^\infty \frac{1}{Z}|\hat{f}(r\alpha)|(2\pi)^2 r^{n+1} dr \\ &\leqslant \frac{4\pi^2 \rho}{Z} \int_0^\infty \frac{r^{n+1}}{1 + r^k} dr \leqslant \frac{8\pi^2 \rho}{Z}. \end{aligned}$$

■

### D. Importance Sampling

Let $P$ be a probability density function over $\mathbb{S}^n \times [-R, R]$ with $\inf_{\mathbb{S}^n \times [-R,R]} P(\alpha, t) = P_{\min} > 0$. When $(\boldsymbol{\alpha}, \boldsymbol{t})$ are distributed according to $P$:

$$\begin{aligned} &f(x) - a^\top x - b \\ &= \int_{\mathbb{S}^{n-1}} \int_{-R}^{R} \frac{g(\alpha, t)}{P(\alpha, t)} \sigma(\alpha^\top x - t) P(\alpha, t) dt \mu_{n-1}(d\alpha) \\ &= \mathbb{E}\left[ \frac{g(\boldsymbol{\alpha}, \boldsymbol{t})}{P(\boldsymbol{\alpha}, \boldsymbol{t})} \sigma(\boldsymbol{\alpha}^\top x - \boldsymbol{t}) \right], \end{aligned} \tag{12}$$

where $\mathbb{E}$ denotes the expected value over $(\boldsymbol{\alpha}, \boldsymbol{t})$.

Let $(\boldsymbol{\alpha}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{\alpha}_m, \boldsymbol{t}_m)$ be independent, identically distributed samples from $P$. The *importance sampling* estimate of $f$ is defined by:

$$\boldsymbol{f}_I(x) = a^\top x + b + \frac{1}{m} \sum_{i=1}^{m} \frac{g(\boldsymbol{\alpha}_i, \boldsymbol{t}_i)}{P(\boldsymbol{\alpha}_i, \boldsymbol{t}_i)} \sigma(\boldsymbol{\alpha}_i^\top x - \boldsymbol{t}_i).$$

*Lemma 2: If Assumption 1 holds, then for all $\nu \in (0, 1)$, the following bound holds with probability at least $1 - \nu$:*

$$\sup_{x \in B(R)} |\boldsymbol{f}_I(x) - f(x)| \leqslant \frac{1}{\sqrt{m}} \left( 100\sqrt{n\pi} LR + (\gamma + 32LR)\sqrt{\log(4/\nu)} \right),$$

*where $\gamma = \frac{8\pi^2 \rho R}{P_{\min}} + 4\rho A_{n-1}(1 + R\pi)$ and $L = \frac{8\pi^2 \rho}{P_{\min}} + 8\pi A_{n-1}\rho$.*

*Proof:* Define the random functions $\boldsymbol{\xi}_i$ and $\boldsymbol{\theta}$ by

$$\begin{aligned} \boldsymbol{\xi}_i(x) &= \frac{g(\boldsymbol{\alpha}_i, \boldsymbol{t}_i)}{P(\boldsymbol{\alpha}_i, \boldsymbol{t}_i)} \sigma(\boldsymbol{\alpha}_i^\top x - \boldsymbol{t}_i) + a^\top x + b - f(x) \\ \boldsymbol{\theta}(x) &= \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\xi}_i(x) = \boldsymbol{f}_I(x) - f)(x) \end{aligned}$$

Lemma 1 implies that $\boldsymbol{\xi}_i(x)$ have zero mean for all $\|x\| \leqslant R$.

In order to bound $\sup_{x \in B(R)} |\boldsymbol{\theta}(x)|$, we utilize:

$$\begin{aligned} \sup_{x \in B(R)} |\boldsymbol{\theta}(x)| &= \sup_{x \in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(0) + \boldsymbol{\theta}(0)| \\ &\leqslant |\boldsymbol{\theta}(0)| + \sup_{x, y \in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)|. \end{aligned} \tag{13}$$

We will bound each term on the right with high probability.

To bound $|\boldsymbol{\theta}(0)|$, we first bound $|\boldsymbol{\xi}_i(0)|$: The triangle inequality, followed by the bounds from Lemma 1 and $|f(0)| \leqslant \|\hat{f}\|_1 \leqslant 2\rho A_{n-1}$ gives

$$|\boldsymbol{\xi}_i(0)| \leqslant \left|\frac{g(\boldsymbol{\alpha}_i,\boldsymbol{t}_i)}{P(\boldsymbol{\alpha}_i,\boldsymbol{t}_i)} t_i\right| + |b| + |f(0)| \leqslant \gamma.$$

A random variable, $\boldsymbol{v}$, is called $\sigma$-sub-Gaussian if $\mathbb{E}[e^{\lambda \boldsymbol{v}}] \leqslant e^{\frac{\lambda^2\sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$. Hoeffding's lemma implies that that $\boldsymbol{\xi}_i(0)$ is $\gamma$-sub-Gaussian. Then, the Hoeffding bound applied to $\boldsymbol{\theta}(0)$ and $-\boldsymbol{\theta}(0)$ gives for all $t \geqslant 0$: $\mathbb{P}(|\boldsymbol{\theta}(0)| \geqslant t) \leqslant 2\exp\left(-\frac{mt^2}{2\gamma}\right)$. Setting $2\exp\left(-\frac{mt^2}{2\gamma^2}\right) = \frac{\nu}{2}$ gives

$$\mathbb{P}\left(|\boldsymbol{\theta}(0)| \geqslant \gamma\sqrt{\frac{2\log(4/\nu)}{m}}\right) \leqslant \frac{\nu}{2}. \tag{14}$$

Now we will bound $\sup_{x,y\in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)|$ via the Dudley entropy integral.

We show that $\boldsymbol{\xi}_i$ are $L$-Lipschitz, where $L$ was defined above. By (6b), $f$ is $4\pi A_{n-1}\rho$-Lipschitz. The bound on $L$ now follows via the triangle inequality, using that $\sigma$ is 1-Lipschitz and the bounds on $\|a\|_2$ and $\|g\|_\infty$.

Let $\psi_2(t) = e^{t^2} - 1$. The corresponding *Orlicz* norm for a zero-mean random scalar variable, $\boldsymbol{v}$, is defined by

$$\|\boldsymbol{v}\|_{\psi_2} = \inf\{\lambda > 0 | \mathbb{E}[\psi_2(\boldsymbol{v}/\lambda)] \leqslant 1\}.$$

If $\boldsymbol{v}$ is $\sigma$-sub-Gaussian, then $\|\boldsymbol{v}\|_{\psi_2} \leqslant 2\sigma$. (Lemma 7 of [31].)

The Lipschitz property and Hoeffding's Lemma imply that $\boldsymbol{\xi}_i(x) - \boldsymbol{\xi}_i(y)$ is $L\|x-y\|$-sub-Gaussian. By Exercise 2.13 of [32], $\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)$ is $\frac{L\|x-y\|}{\sqrt{m}}$-sub-Gaussian. Thus, $\|\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)\|_{\psi_2} \leqslant \frac{2L}{\sqrt{m}}\|x-y\|$. Thus, $\boldsymbol{\theta}$ is an *Orlicz process* with respect to the metric $d(x,y) = \frac{2L}{\sqrt{m}}\|x-y\|$. The diameter of $B(R)$ with respect to $d$ is $D := \frac{4LR}{\sqrt{m}}$.

Let $N(\epsilon, B(R), d)$ denote the $\epsilon$-*covering number* of $B(R)$ under the metric $d$, which is the minimal cardinality of a covering of $B(R)$ with $d$-balls of radius $\epsilon > 0$. Using that $N(\epsilon, B(1), \|\cdot\|) \leqslant \left(1+\frac{2}{\epsilon}\right)^n$ gives $N(\epsilon, B(R), d) \leqslant \left(1+\frac{D}{\epsilon}\right)^n$ after rescaling.

Let $\mathcal{J} = \int_0^D \sqrt{\log(1+N(\epsilon,B(R),d))}d\epsilon$. Theorem 5.36 of [32] implies that for all $t > 0$:

$$\mathbb{P}\left(\sup_{x,y\in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)| \geqslant 8(t+\mathcal{J})\right) \leqslant 2e^{-t^2/D^2}.$$

(Theorem 5.36 in [32] is more general, and has an unspecified constant. In this case, the constant is 8.)

Setting $2e^{-t^2/D^2} = \nu/2$ gives

$$\mathbb{P}\left(\sup_{x,y\in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)| \geqslant 8\left(\mathcal{J} + D\sqrt{\log(4/\nu)}\right)\right) \leqslant \frac{\nu}{2}. \tag{15}$$

Using $1 + N(\epsilon, B(R), d) \leqslant 2\left(1+\frac{D}{\epsilon}\right)^n$, we bound $\mathcal{J}$:

$$\begin{aligned}
\mathcal{J} &\leqslant D\sqrt{\log(2)} + \sqrt{n}\int_0^D \sqrt{\log\left(1+\frac{D}{\epsilon}\right)}d\epsilon \\
&\overset{t=1+\frac{D}{\epsilon}}{=} D\sqrt{\log(2)} + \sqrt{n}D\int_2^\infty \frac{\sqrt{\log t}}{(t-1)^2}dt \\
&\leqslant D + 2\sqrt{n}D\int_2^\infty \frac{\sqrt{\log t}}{t^2}dt \\
&\overset{u=\log t}{=} D + 2\sqrt{n}D\int_{\log 2}^\infty u^{-1/2}e^{-u}du \\
&\overset{\Gamma(1/2)=\sqrt{\pi}}{\leqslant} (1+2\sqrt{n\pi})D \leqslant 3\sqrt{n\pi}D.
\end{aligned}$$

A union bound shows that

$$\mathbb{P}\Bigg(|\boldsymbol{\theta}(0)| \geqslant \gamma\sqrt{\frac{2\log(4/\nu)}{m}} \cup \sup_{x,y\in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)| \geqslant 8\left(\mathcal{J} + D\sqrt{\log(4/\nu)}\right)\Bigg) \leqslant \nu$$

By De Morgan's law, with probability at least $1-\nu$

$$|\boldsymbol{\theta}(0)| < \gamma\sqrt{\frac{2\log(4/\nu)}{m}} \text{ and}$$
$$\sup_{x,y\in B(R)} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)| < 8\left(\mathcal{J} + D\sqrt{\log(4/\nu)}\right).$$

both hold. The result now follows from (13). ■

### E. Proof of Theorem 1

The proof follows by setting $\boldsymbol{c}_i = \frac{g(\boldsymbol{\alpha}_i,\boldsymbol{t}_i)}{mP(\boldsymbol{\alpha}_i,\boldsymbol{t}_i)}$ and collecting all of the bounds on the terms. ■

## IV. Application

Here, we apply Theorem 1 to an approximate Model Reference Adaptive Control (MRAC) method from [1].[1]

### A. Setup and Existing Results

Chapter 12 of [1] describes an approximate MRAC scheme for plants of the form

$$\dot{x}_t = Ax_t + B(u_t + f(x_t)), \tag{16}$$

where $f : \mathbb{R}^n \to \mathbb{R}^\ell$ is an unknown nonlinearity. (The setup in [1] is more general. The discussion here is for illustration.)

We sketch the methodology, and describe how Theorem 1 can be used to give bounds on controller performance, while deferring the details of the controller and its analysis to [1].

In (16), $A$ is unknown, while $B$ is a known.

It is *assumed* that there is a nonlinear vector function $\Psi : \mathbb{R}^n \to \mathbb{R}^N$ and an unknown $\ell \times N$ parameter matrix, $\Theta$ such that $\Theta\Psi(x)$ gives a good approximation to $f$ on a bounded region. The details are described in Theorem 2, below.

It is assumed that there exist feedback and feedforward gain matrices, $K_x$ and $K_r$, satisfying the *matching conditions*

$$A + BK_x = A_r \quad BK_r = B_r$$

[1] For simulations, see https://github.com/tylerlekang/CDC2024.

to a controllable, linear reference model

$$\dot{x}_t^r = A_r\, x_t^r + B_r\, r_t, \tag{17}$$

where $A_r$ is an $n \times n$ Hurwitz matrix, $A_r$ and $B_r$ are both known, and $r_t \in \mathbb{R}^\ell$ is a bounded reference input.

The adaptive law takes the form

$$u_t \ = \hat{K}_{x,t} x_t - \widehat{\Theta}_t \Psi(x_t) + (1 - \mu(x_t))\, \hat{K}_{r,t} r_t + \mu(x_t) \tilde{u}(x_t), \tag{18}$$

where the dynamics of $\hat{K}_{x,t}$, $\widehat{\Theta}_t$, and $\hat{K}_{r,t}$ are designed via Lyapunov methods, $\tilde{u}$ is a control law that keeps the state bounded, and $\mu$ is a weighting function.

Let $P$ and $Q$ be positive definite matrices such that:

$$A_r^\top P + P A_r = -Q.$$

The controller has the following guarantees:

*Theorem 2 ([1]): Let $x_t^r$ be a fixed reference trajectory generated by a bounded reference input $r_t$. Assume that there is an* <u>*unknown*</u> *parameter matrix $\Theta$, a known bounding function $\epsilon_{\max}$, and positive numbers $R$ and $\epsilon_0$ such that:*

*(i)* $\|f(x) - \Theta\Psi(x)\| \leqslant \epsilon_{\max}(x)$ *for all* $x \in \mathbb{R}^n$

*(ii)* $R \geqslant 4\|P\|\|Q^{-1}\|\epsilon_0 + \sup_{t \geqslant 0} \|x_t^r\|$

*(iii)* $\sup_{x \in B(R)} \|f(x) - \Theta\Psi(x)\| \leqslant \epsilon_0$

*Then there is an adaptive law of the form in (18) such that for any initial condition, $x_0$:*

- *There is a time $T_1$ such that $x_t \in B(R)$ for all $t \geqslant T_1$.*
- *There is a time $T_2$ such that the tracking error satisfies $(x_t - x_t^r) \in B\left(4\|P\|\|Q^{-1}\|\epsilon_0\right)$ for all $t \geqslant T_2$.*

In particular, Theorem 2 states that as long as a good approximation of the form $\Theta\Psi(x)$ exists over a sufficiently large bounded region, the controller will drive the state to a bounded region and make the tracking error arbitrarily small. We do not actually need to know the parameter matrix, $\Theta$.

A gap in current adaptive control analysis with such linear parametrizations is *proving* that the $\Theta$ matrix exists.

In [1] it is suggested that the entries of $\Psi$ take the form $\Psi_i(x) = \phi(\alpha_i^\top x - b_i)$, where $\phi$ is a neural network activation function, $\alpha_i$ is a weight vector, and $b_i$ is a bias. Classical approximation theorems guarantee that weights and biases, $(\alpha_i, b_i)$ *exist* such that $\inf_{\Theta \in \mathbb{R}^{\ell \times N}} \sup_{x \in B(R)} \|f(x) - \Theta\Psi(x)\| \leqslant \epsilon_0$, but do not describe how to find them. Lemma 3 below implies that they can be generated randomly.

While we focus on a method from [1] for concreteness, similar gaps in the analysis are common [1], [2], [5]–[11]. These gaps are specifically articulated in [12], [13].

### B. Guaranteed Approximations for the Nonlinearity

The result below gives a randomized construction of a nonlinear vector function, $\mathbf{\Psi}$, such that for any $\epsilon_0 > 0$ and any $R > 0$, $\mathbf{\Psi}$ satisfies the conditions of Theorem 2 with high probability, as long as $f$ is sufficiently smooth.

*Lemma 3: Assume that every entry of $f : \mathbb{R}^n \to \mathbb{R}^\ell$ satisfies Assumption (1). Define $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^{m+n+1}$ by:*

$$\mathbf{\Psi}(x)^\top = \begin{bmatrix} 1 & x^\top & \sigma(\boldsymbol{\alpha}_1^\top x - \boldsymbol{t}_1) & \cdots & \sigma(\boldsymbol{\alpha}_m^\top x - \boldsymbol{t}_m) \end{bmatrix}$$

*where $(\boldsymbol{\alpha}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{\alpha}_m, \boldsymbol{t}_m)$ are independent identically distributed samples from $P$. For any $\nu \in (0,1)$, if*

$$m \geqslant \frac{\ell}{\epsilon_0^2}\left(\kappa_0 + \kappa_1\sqrt{\log(4\ell/\nu)}\right)^2,$$

*then with probability at least $1 - \nu$, there is a matrix $\mathbf{\Theta}$ such that $\sup_{x \in B(R)} \|f(x) - \mathbf{\Theta}\mathbf{\Psi}(x)\| \leqslant \epsilon_0$. Furthermore, the bounding function can be taken as:*

$$\epsilon_{\max}(x) = 2\sqrt{\ell}\rho A_{n-1} + \sqrt{\ell}\rho(1 + \|x\|\sqrt{m+1}) \cdot \left((1 + 4\pi + 2\pi R)A_{n-1} + \frac{8\pi^2\rho}{\sqrt{m}P_{\min}}\right).$$

*Proof:* Let $\boldsymbol{\theta}_i(x)$ be the $i$th entry of $\mathbf{\Theta}\mathbf{\Psi}(x) - f(x)$, where the entries of $\mathbf{\Theta}$ are constructed from the importance sampling approximation for each $f_i(x)$. For each $i$, with probability at most $\nu/\ell$ the error has

$$\sup_{x \in B(R)} |\boldsymbol{\theta}_i(x)| \geqslant \frac{1}{m}\left(\kappa_0 + \kappa_1\sqrt{4\ell/\nu}\right),$$

So, a union bounding / De Morgan argument shows that with probability at least $1 - \nu$, all entries satisfy

$$\sup_{x \in B(R)} |\boldsymbol{\theta}_i(x)| \leqslant \frac{1}{\sqrt{m}}\left(\kappa_0 + \kappa_1\sqrt{4\ell/\nu}\right),$$

which implies that

$$\|\mathbf{\Theta}\mathbf{\Psi}(x) - f(x)\| \leqslant \frac{\sqrt{\ell}}{\sqrt{m}}\left(\kappa_0 + \kappa_1\sqrt{4\ell/\nu}\right).$$

The sufficient condition for $\|\mathbf{\Theta}\mathbf{\Psi}(x) - f(x)\| \leqslant \epsilon_0$ now follows by re-arrangement.

The bound on $\epsilon_{\max}(x)$ uses the triangle inequality:

$$\|f(x) - \mathbf{\Theta}\Psi(x)\| \leqslant \|f(x)\| + \|\mathbf{\Theta}\|\|\Psi(x)\|.$$

Then we bound $\|f_i\|_\infty \leqslant \|\hat{f}_i\|_1 \leqslant 2\rho A_{n-1}$ , $\|\mathbf{\Psi}(x)\| \leqslant (1 + \|x\|\sqrt{m+1})$, and use the bounds on the coefficients from Theorem 1 to bound $\|\mathbf{\Theta}\|$. ■

## V. Conclusion and Future Directions

In this paper, we gave a simple bound on the error in approximation smooth functions with random ReLU networks. We showed how the results can be applied to an adaptive control problem. The key intermediate result was a novel integral representation theorem for ReLU activation functions. Remaining theoretical challenges include quantifying the constants precisely and relaxing the smoothness requirements. Natural extensions include the examination of other activation functions and applications to different control problems. Other directions would be extensions to deep networks and networks with trained hidden layers.

## References

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

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

[3] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, Second. MIT press, 2018.

[5] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. Dixon, *Reinforcement learning for optimal feedback control*. Springer, 2018.

[6] M. L. Greene, Z. I. Bell, S. Nivison, and W. E. Dixon, "Deep neural network-based approximate optimal tracking for unknown nonlinear systems," *IEEE Transactions on Automatic Control*, 2023.

[7] W. A. Makumi, Z. I. Bell, and W. E. Dixon, "Approximate optimal indirect regulation of an unknown agent with a lyapunov-based deep neural network," *IEEE Control Systems Letters*, 2023.

[8] M. H. Cohen and C. Belta, "Safe exploration in model-based reinforcement learning using control barrier functions," *Automatica*, vol. 147, p. 110 684, 2023.

[9] N.-M. Kokolakis, K. G. Vamvoudakis, and W. Haddad, "Reachability analysis-based safety-critical control using online fixed-time reinforcement learning," in *Learning for Dynamics and Control Conference*, PMLR, 2023, pp. 1257–1270.

[10] M. Sung, S. H. Karumanchi, A. Gahlawat, and N. Hovakimyan, "Robust model based reinforcement learning using $\mathcal{L}_1$ adaptive control," in *The Twelfth International Conference on Learning Representations*, 2023.

[11] B. Lian, W. Xue, F. L. Lewis, H. Modares, and B. Kiumarsi, "Inverse reinforcement learning for optimal control systems," in *Integral and Inverse Reinforcement Learning for Optimal Control Systems and Games*, Springer, 2024, pp. 151–181.

[12] D. Soudbakhsh, A. M. Annaswamy, Y. Wang, S. L. Brunton, J. Gaudio, H. Hussain, D. Vrabie, J. Drgona, and D. Filev, "Data-driven control: Theory and applications," in *2023 American Control Conference (ACC)*, IEEE, 2023, pp. 1922–1939.

[13] A. M. Annaswamy, "Adaptive control and intersections with reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 65–93, 2023.

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

[15] S. Sonoda and N. Murata, "Neural network with unbounded activation functions is universal approximator," *Applied and Computational Harmonic Analysis*, vol. 43, no. 2, pp. 233–268, 2017.

[16] A. Petrosyan, A. Dereventsov, and C. G. Webster, "Neural network integral representations with the relu activation function," in *Mathematical and Scientific Machine Learning*, PMLR, 2020, pp. 128–143.

[17] B. Irie and S. Miyake, "Capabilities of three-layered perceptrons.," in *ICNN*, 1988, pp. 641–648.

[18] P. C. Kainen, V. Kůrková, and A. Vogt, "Integral combinations of heavisides," *Mathematische Nachrichten*, vol. 283, no. 6, pp. 854–878, 2010.

[19] R. M. Neal, "Priors for infinite networks (Technical Report CRG-TR-94-1)," 1994.

[20] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as gaussian processes," *arXiv preprint arXiv:1711.00165*, 2017.

[21] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[22] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *International Conference on Machine Learning*, 2019, pp. 1675–1685.

[23] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*, 2019, pp. 242–252.

[24] D. Hsu, C. Sanford, R. A. Servedio, and E.-V. Vlatakis-Gkaragkounis, "On the approximation power of two-layer networks of random relus," in *Conference on Learning Theory*, PMLR, 2021, pp. 2423–2461.

[25] G. Yehudai and O. Shamir, "On the power and limitations of random features for understanding neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 6598–6608, 2019.

[26] M. Li, S. Sonoda, F. Cao, Y. G. Wang, and J. Liang, "How powerful are shallow neural networks with bandlimited random weights?" In *International Conference on Machine Learning*, PMLR, 2023, pp. 19 960–19 981.

[27] 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.

[28] A. Lamperski, "Neural network independence properties with applications to adaptive control," in *IEEE Conference on Decision and Control (CDC)*, 2022.

[29] L. Blumenson, "A derivation of n-dimensional spherical coordinates," *The American Mathematical Monthly*, vol. 67, no. 1, pp. 63–66, 1960.

[30] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 999–1013, 1993.

[31] A. Lamperski, "Nonasymptotic pointwise and worst-case bounds for classical spectrum estimators," *IEEE Transactions on Signal Processing*, vol. 71, pp. 4273–4287, 2023.

[32] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.