# Metric-Entropy Limits on Nonlinear Dynamical System Learning

Yang Pan, Clemens Hutter, and Helmut Bölcskei

*Chair for Mathematical Information Science, ETH Zurich*

*Sternwartstrasse 7, 8092 Zurich, Switzerland*

**Abstract**

This paper is concerned with the fundamental limits of nonlinear dynamical system learning from input-output traces. Specifically, we show that recurrent neural networks (RNNs) are capable of learning nonlinear systems that satisfy a Lipschitz property and forget past inputs fast enough in a metric-entropy optimal manner. As the sets of sequence-to-sequence maps realized by the dynamical systems we consider are significantly more massive than function classes generally considered in deep neural network approximation theory, a refined metric-entropy characterization is needed, namely in terms of order, type, and generalized dimension. We compute these quantities for the classes of exponentially-decaying and polynomially-decaying Lipschitz fading-memory systems and show that RNNs can achieve them.

*This paper is dedicated to Professor Andrew R. Barron on the occasion of his 65$^{th}$ birthday.*

## I. Introduction

It is well known that neural networks can approximate almost any function arbitrarily well [1]–[5]. The recently developed Kolmogorov-Donoho rate-distortion theory for deep neural network approximation [6], [7] goes a step further by quantifying how effective such approximations are in terms of the description complexity of the networks relative to that of the functions they are to approximate. Specifically, [7] considers classes of functions mapping $\mathbb{R}^d$ to $\mathbb{R}$ and aims at approximating every function in a given class to within a prescribed error $\epsilon$ using a (deep) rectified linear unit (ReLU) network. Moreover, the length of the bitstring specifying the approximating network is characterized. Now, [7] establishes that for a wide variety of function classes, the length of this bitstring exhibits the same scaling behavior, in $\epsilon$, as the metric entropy of the function class under consideration (see Table I below). This means that neural networks are universally Kolmogorov-Donoho optimal for all these function classes.

In the present paper, we extend the philosophy of [7] to the approximation of nonlinear sequence-to-sequence mappings through recurrent neural networks (RNNs). Specifically, we consider Lipschitz

H. Bölcskei gratefully acknowledges support by the Lagrange Mathematics and Computing Research Center, Paris, France.

fading-memory (LFM) systems. In essence, this notion describes systems that gradually forget long-past inputs, with the speed of memory decay quantified in terms of a certain Lipschitz property. Such systems find diverse applications, inter alia, in finance [8] and material science [9]–[11]. We first develop tools for quantifying the metric entropy of classes of LFM systems with a given memory decay rate. A general construction of RNNs approximating LFM systems is shown to yield Kolmogorov-Donoho-optimality for LFM systems of exponentially or polynomially decaying memory.

*Related work.* Learning of linear dynamical systems has been studied extensively in the literature [12]–[16]. Notably, [16] provides explicit RNN constructions for a wide class of linear dynamical systems, including time-varying systems. Going beyond linear systems, learning of nonlinear finite memory systems within the finite-state-machine framework has been studied in [17]. More concretely, [17] explores the learning of finite-state finite-memory machines using RNNs. This program is extended to approximately finite-memory systems in [18] and fading-memory systems in [19]–[21]. In particular, [21] formalizes the concept of fading-memory systems in control theory, demonstrating that continuous-time fading-memory operators can be approximated using Volterra series. Subsequently, [22] established that discrete-time fading-memory systems can be identified using neural networks. Moreover, [23] demonstrated that echo state networks, a specialized architecture within the RNN family, serve as universal approximators for discrete-time fading-memory systems. None of the studies just reviewed formally addresses the issue of quantifying the RNN description complexity relative to that of the class of (nonlinear) systems they are to learn.

*Organization of the paper.* The remainder of Section I summarizes notation. In Section II, we introduce our setup and provide a definition of metric-entropy optimality in a very general context encompassing the approximation of functions as well as dynamical systems. Section III develops tools for characterizing the metric entropy of LFM systems. In Section IV, we employ these tools to derive precise scaling results for the metric entropy of exponentially Lipschitz fading-memory (ELFM) and polynomially Lipschitz fading-memory (PLFM) systems. Section V presents a construction for the approximation of general LFM systems by RNNs. Finally, in Section VI, we combine the results developed in the previous sections to prove that RNNs can learn ELFM and PLFM systems in a metric-entropy-optimal manner.

*Notation.* For $N \in \mathbb{N}$, $[\![N]\!]$ stands for the set $\{0, 1, \dots, N\}$, while $[\![N]\!]^{\pm}$ denotes the set $\{-N \dots, -1, 0, 1, \dots, N\}$. The cardinality of a finite set $U$ is designated by $|U|$. Sequences $x[t] \in \mathbb{R}$ are indexed by $t \in \mathbb{Z}$ or $t \in \mathbb{N}$ and we use $\mathbb{R}^{\mathbb{Z}}$ and $\mathbb{R}^{\mathbb{N}}$ to respectively denote the set of such sequences. We refer to the set of all finite-length bitstrings by $\{0, 1\}^*$. The transpose of the matrix $A$ is $A^T$. For matrices $A_1, \dots, A_N$, $\mathrm{diag}(A_1, A_2, \dots, A_N)$ refers to the block-diagonal matrix with the $A_i$ on the main diagonal. The $N \times N$ identity matrix is $\mathbb{I}_N$ and $0_N$ stands for the $N$-dimensional column vector with all entries equal to 0. For the vector $x \in \mathbb{R}^d$, we let $\|x\|_\infty := \max_{i=1,2,\dots,d} |x_i|$. $\log(\cdot)$ refers to the logarithm to base 2, $\log^{(n)} = \log \circ \cdots \circ \log$ is the $n$-fold iterated logarithm, and $\log^{\tau}(\cdot) = (\log(\cdot))^{\tau}$, for $\tau \in \mathbb{R}$. The composition of functions $f_1, f_2$ is denoted by $f_2 \circ f_1$ (or $f_1 \circ f_2$). For $\epsilon > 0$, let $f(\epsilon)$ and $g(\epsilon)$ be strictly positive for all small enough values of $\epsilon$. We use $f(\epsilon) = o(g(\epsilon))$ to indicate that $\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} = 0$ and we express $\limsup_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} < \infty$ by $f(\epsilon) = O(g(\epsilon))$. Moreover, we write $f(\epsilon) = \Theta(g(\epsilon))$ when both $f(\epsilon) = O(g(\epsilon))$ and $g(\epsilon) = O(f(\epsilon))$. Constants are always understood to be in $\mathbb{R}$ unless explicitly stated otherwise. Finally, we say that a constant is universal if it does not depend on any of the ambient quantities.

## II. Problem setup and metric-entropy optimality

### A. ReLU network approximation

We start by defining ReLU networks.

**Definition II.1** (ReLU network [7]). *Let $L \in \mathbb{N}$ and $N_0, N_1, \ldots, N_L \in \mathbb{N}$. A ReLU (feedforward) neural network $\Phi$ is a map $\Phi\ : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3 \end{cases} \tag{1}$$

*where, for $\ell \in \{1, 2, \ldots, L\}$, $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $W_\ell(x) := A_\ell x + b_\ell$, $x \in \mathbb{R}^{N_{\ell-1}}$, are affine transformations with (weight) matrices $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $b_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho : \mathbb{R} \to \mathbb{R}$, $\rho := \max\{x, 0\}$ acts component-wise, i.e., $\rho(x_1, \ldots, x_N) = (\rho(x_1), \ldots, \rho(x_N))$. We denote the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$ by $\mathcal{N}_{d,d'}$. Moreover, we define the following quantities related to the notion of size of the ReLU network $\Phi$:*

- *depth $\mathcal{L}(\Phi) := L$,*
- *the connectivity $\mathcal{M}(\Phi)$ of the network $\Phi$ is the total number of non-zero entries in the matrices $A_\ell,\ \ell \in \{1, 2, ..., L\}$, and the vectors $b_\ell,\ \ell \in \{1, 2, ..., L\}$,*
- *width $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$,*
- *the weight set $\mathcal{K}(\Phi)$ denotes the set of non-zero entries in the matrices $A_\ell,\ \ell \in \{1, 2, ..., L\}$, and the vectors $b_\ell,\ \ell \in \{1, 2, ..., L\}$,*
- *weight magnitude $\mathcal{B} := \max_{\ell=1,\ldots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.*

We next formalize the concept of network weight quantization.

**Definition II.2** (Quantization [7]). *Let $m \in \mathbb{N}$ and $\epsilon \in (0, 1/2)$. The network $\Phi$ is said to have $(m, \epsilon)$-quantized weights if $\mathcal{K}(\Phi) \subset 2^{-m\lceil \log(\epsilon^{-1}) \rceil}\mathbb{Z} \cap [-\epsilon^{-m}, \epsilon^{-m}]$. Moreover, for $a \in \mathbb{R}$, we define the $(m, \epsilon)$-quantization map rounding real-valued numbers to integer multiples of $2^{-m\lceil \log(\epsilon^{-1}) \rceil}$ as*

$$\mathcal{Q}_{m,\epsilon}(a) := \left\lceil a/2^{-m\lceil \log(\epsilon^{-1}) \rceil} \right\rceil \cdot 2^{-m\lceil \log(\epsilon^{-1}) \rceil}. \tag{2}$$

Every quantized ReLU network can be represented by a bitstring specifying the topology of the network along with its quantized non-zero weights, i.e., the entries of $A_\ell,\ \ell \in \{1, 2, ..., L\}$, and $b_\ell,\ \ell \in \{1, 2, ..., L\}$. In Appendix A, we specify how this bitstring is organized. Taking this bitstring back to the quantized ReLU network is done through a mapping, which we denote by $\mathcal{D}_\mathcal{N}$ and refer to as the *canonical neural network decoder*.

**Remark II.3.** *For every ReLU network $\Phi$ with $(m, \epsilon)$-quantized weights, there is a bitstring $\mathbf{b}$ of length no more than $C_0 m \log(\epsilon^{-1})\mathcal{M}(\Phi) \log(\mathcal{M}(\Phi))$ such that $\mathcal{D}_\mathcal{N}(\mathbf{b}) = \Phi$, with $C_0 > 0$ a universal constant. This follows by upper-bounding* (116) *in Appendix A.*

As we consider the approximation of sequence-to-sequence maps ($\mathbb{R}^\mathbb{Z} \to \mathbb{R}^\mathbb{Z}$), feedforward networks as defined above are not applicable since they effect maps between finite-dimensional spaces, concretely from $\mathbb{R}^{N_0}$ to $\mathbb{R}^{N_L}$. However, and perhaps surprisingly, simply applying feedforward networks iteratively in a judicious manner turns out to be sufficient for approximating interesting classes of nonlinear sequence-to-sequence maps in a Kolmogorov-Donoho-optimal manner. Concretely, this gives rise to the concept of recurrent neural networks.

**Definition II.4** (Recurrent neural networks [16])**.** *For $m \in \mathbb{N}$, let $\Phi \in \mathcal{N}_{m+1,m+1}$ be a ReLU network of depth $\mathcal{L}(\Phi) \geq 2$. The recurrent neural network (RNN) associated with $\Phi$ is the operator $\mathcal{R}_\Phi : \mathbb{R}^\mathbb{N} \to \mathbb{R}^\mathbb{N}$ mapping input sequences $(x[t])_{t\geq 0}$ in $\mathbb{R}$ to output sequences $(y[t])_{t\geq 0}$ in $\mathbb{R}$ according to*

$$\begin{pmatrix} y[t] \\ h[t] \end{pmatrix} = \Phi\left(\begin{pmatrix} x[t] \\ h[t-1] \end{pmatrix}\right), \; \forall t \geq 0, \tag{3}$$

*where $h[t] \in \mathbb{R}^m$ is the hidden state vector sequence with initial state $h[-1] = 0_m$. We denote the set of all RNNs by $\mathcal{N}^R$.*

**Remark II.5.** *When unfolded in time, an RNN simply amounts to repeated application of $\Phi$.*

From Definition II.4 it is apparent that an RNN $\mathcal{R}_\Phi$ is fully specified by its associated feedforward network $\Phi$.

**Definition II.6.** *Formally, we define $\mathbf{R} : \bigcup_{m=1}^{\infty} \mathcal{N}_{m+1,m+1} \to \mathcal{N}^R$ as the map that takes a ReLU network $\Phi$ to its associated RNN $\mathcal{R}_\Phi$ according to Definition II.4.*

Together with the canonical neural network decoder $\mathcal{D}_\mathcal{N}$, we thus obtain the following procedure for decoding a bitstring to an RNN.

**Definition II.7** (Canonical RNN decoder)**.** *We define the canonical RNN decoder as $\mathcal{D}_\mathcal{R} := \mathbf{R} \circ \mathcal{D}_\mathcal{N}$, where $\mathbf{R}$ is as in Definition II.6 and $\mathcal{D}_\mathcal{N}$ is the canonical neural network decoder.*

The main point of this paper is to show that the canonical RNN decoder is capable of approximating a wide variety of non-linear sequence-to-sequence maps ($\mathbb{R}^\mathbb{Z} \to \mathbb{R}^\mathbb{Z}$) in a metric-entropy-optimal manner. This means that there are no other decoders that use fundamentally fewer bits. In addition, the results in [7] show that the canonical neural network decoder optimally approximates a wide variety of function classes mapping $\mathbb{R}^d$ to $\mathbb{R}$. Taken together, we will hence be able to conclude that ReLU networks are able to optimally approximate function classes as well as sequence-to-sequence maps.

### *B. Metric-Entropy Optimality*

In this section, we rigorously define the notion of metric-entropy optimal approximation. Consider a metric space $(\mathcal{X}, \rho)$ and a compact subset $\mathcal{C} \subset \mathcal{X}$. Together, $\mathcal{C}$ and $\rho$ determine an approximation task. Specifically, we wish to approximate elements $f \in \mathcal{C}$ to within a prescribed error $\epsilon > 0$ in the metric $\rho$ by elements $\widetilde{f} \in \mathcal{X}$ which can be encoded by finite-length bitstrings $\mathbf{b} \in \{0,1\}^\ell$. To go from bitstrings to elements of $\mathcal{X}$, we define decoder mappings as follows.

**Definition II.8.** *A decoder $\mathcal{D} : \{0,1\}^* \to \mathcal{X}$ is a mapping from bitstrings of arbitrary length to elements of $\mathcal{X}$.*

We shall frequently want to quantify how well a given decoder $\mathcal{D}$ performs.

**Definition II.9.** *Given a metric space $(\mathcal{X}, \rho)$, a compact set $\mathcal{C} \subset \mathcal{X}$, and a decoder $\mathcal{D} : \{0,1\}^* \to \mathcal{X}$, we say that $(\mathcal{C}, \rho)$ is representable by $\mathcal{D}$, if for every $\epsilon > 0$ and every $f \in \mathcal{C}$, there exist $\ell \in \mathbb{N}$ and a bitstring $\mathbf{b} \in \{0,1\}^\ell$ such that*

$$\rho(\mathcal{D}(\mathbf{b}), f) \leq \epsilon.$$

*Furthermore, we set*

$$L(\epsilon; \mathcal{D}, \mathcal{C}, \rho) := \min\left\{\ell' \in \mathbb{N} \mid \forall f \in \mathcal{C},\ \exists \ell \leq \ell',\ \exists \mathbf{b} \in \{0,1\}^\ell \textit{ s.t. } \ \rho(\mathcal{D}(\mathbf{b}), f) \leq \epsilon\right\}.$$

**Remark II.10.** *This setting allows us to fix a decoder $\mathcal{D}$ (e.g., the canonical neural network decoder) and then study how well $\mathcal{D}$ performs on different $(\mathcal{C}, \rho)$. That is, $\mathcal{D}$ does not depend on $\mathcal{C}, \rho, f$, or $\epsilon$.*

The quantity $L(\epsilon; \mathcal{D}, \mathcal{C}, \rho)$ measures how bit-efficient the decoder $\mathcal{D}$ is in representing $\mathcal{C}$ with respect to $\rho$. It is now natural to ask what the minimum required number of bits, independently of $\mathcal{D}$, is for representing $\mathcal{C}$ with respect to $\rho$. The concept of metric entropy [24], [25] gives an answer to this question.

**Definition II.11.** *Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C} \subset \mathcal{X}$ compact. The set $\{x_1, x_2, \ldots, x_N\} \subset \mathcal{C}$* (respectively $\{x_1, x_2, \ldots, x_N\} \subset \mathcal{X}$) *is an $\epsilon$-covering* (respectively $\epsilon$-net) *for $(\mathcal{C}, \rho)$ if, for each $x \in \mathcal{C}$, there exists an $i \in \{1, 2, \ldots, N\}$ so that $\rho(x, x_i) \leq \epsilon$. The $\epsilon$-covering number $N(\epsilon; \mathcal{C}, \rho)$* (respectively the exterior $\epsilon$-covering number $N^{\text{ext}}(\epsilon; \mathcal{C}, \rho)$) *is the cardinality of a smallest $\epsilon$-covering* (respectively smallest $\epsilon$-net) *for $(\mathcal{C}, \rho)$.*

In general, it is hard to obtain precise expressions for covering numbers. One therefore typically resorts to characterizations of their asymptotic behavior as $\epsilon \to 0$. In [7], where sets of functions are considered, this is done through the concept of optimal exponents. Here, however, we are concerned with sets of systems, which are much more massive and hence require a refined framework for quantifying the asymptotic behavior of their covering numbers. Thus, inspired by [26, Section II.C], we use the following notions.

**Definition II.12** (Order, type, and generalized dimension)**.** *Consider a metric space $(\mathcal{X}, \rho)$ and a compact set $\mathcal{C} \subset \mathcal{X}$. Then $(\mathcal{C}, \rho)$ is said to be of order $\kappa \in \mathbb{N}$ and type $\lambda \in \mathbb{N}$ if the quantity*

$$\mathfrak{d} := \limsup_{\epsilon \to 0} \frac{\log^{(\kappa+1)} N^{ext}(\epsilon; \mathcal{C}, \rho)}{\log^{\lambda}(\epsilon^{-1})} \tag{4}$$

*is finite and non-zero. In this case, we call $\mathfrak{d}$ the generalized dimension.*

Order, type, and generalized dimension provide measures for the "description complexity" of $(\mathcal{C}, \rho)$ with the order $\kappa$ the coarsest one. For a given order, the type $\lambda$ constitutes a finer measure, and for fixed order and type, the generalized dimension $\mathfrak{d}$ is the finest measure [26].

Whenever the optimal exponent according to [7, Definition IV.1] is well-defined (i.e., strictly positive and finite), the underlying set has order and type equal to one and generalized dimension equal to the inverse of the optimal exponent (Lemma B.1). Based on this insight, we obtain Table I, which lists the generalized dimension for the sets considered in [7, Table 1].

| | Metric | $\mathcal{C}$ | | $\kappa$ | $\lambda$ | $\mathfrak{d}$ |
|---|---|---|---|---|---|---|
| $\{\mathbb{R} \to \mathbb{R}\}$ | $L^2([0,1])$ | $L^2$-Sobolev | $\mathcal{U}(W_2^m([0,1]))$ | 1 | 1 | $1/m$ |
| $\{\mathbb{R} \to \mathbb{R}\}$ | $L^2([0,1])$ | Hölder | $\mathcal{U}(C^\alpha([0,1]))$ | 1 | 1 | $1/\alpha$ |
| $\{\mathbb{R} \to \mathbb{R}\}$ | $L^2([0,1])$ | Bump Algebra | $\mathcal{U}(B_{1,1}^1([0,1]))$ | 1 | 1 | 1 |
| $\{\mathbb{R} \to \mathbb{R}\}$ | $L^2([0,1])$ | Bounded Variation | $\mathcal{U}(BV([0,1]))$ | 1 | 1 | 1 |
| $\{\mathbb{R}^d \to \mathbb{R}\}$ | $L^2(\Omega)$ | $L^p$-Sobolev | $\mathcal{U}(W_p^m(\Omega))$ | 1 | 1 | $\frac{d}{m}$ |
| $\{\mathbb{R}^d \to \mathbb{R}\}$ | $L^2(\Omega)$ | Besov | $\mathcal{U}(B_{p,q}^m(\Omega))$ | 1 | 1 | $\frac{d}{m}$ |
| $\{\mathbb{R}^d \to \mathbb{R}\}$ | $L^2(\Omega)$ | Modulation | $\mathcal{U}(M_{p,p}^s(\mathbb{R}^d))$ | 1 | 1 | $(\frac{1}{p} - \frac{1}{2} + \frac{2s}{d})$ |
| $\{\mathbb{R}^d \to \mathbb{R}\}$ | $L^2(\Omega)$ | Cartoon functions | $\mathcal{E}^\beta([-\frac{1}{2}, \frac{1}{2}]^d)$ | 1 | 1 | $\frac{2}{\beta(d-1)}$ |

TABLE I: Generalized dimension for the sets considered in [7]. Here, $\mathcal{U}(X) = \{f \in X : \|f\|_X \leq 1\}$ denotes the unit ball in the space $X$ and $\Omega \subseteq \mathbb{R}^d$ is a Lipschitz domain.

Returning to the previous discussion, we are now able to characterize the minimum number of bits required by any decoder to $\epsilon$-represent $(\mathcal{C}, \rho)$.

**Lemma II.13.** *Consider the metric space $(\mathcal{X}, \rho)$, the compact set $\mathcal{C} \subset \mathcal{X}$ of order $\kappa$, type $\lambda$, and generalized dimension $\mathfrak{d}$, and assume that $(\mathcal{C}, \rho)$ is representable by a decoder $\mathcal{D}$. Then, it holds that*

$$\limsup_{\epsilon \to 0} \frac{\log^{(\kappa)} L(\epsilon; \mathcal{D}, \mathcal{C}, \rho)}{\log^{\lambda}(\epsilon^{-1})} \geq \mathfrak{d}. \tag{5}$$

*Proof.* See Appendix C-A. □

It is natural to say that a decoder $\mathcal{D}$ is optimal if it satisfies (5) with equality.

**Definition II.14.** *Consider the metric space $(\mathcal{X}, \rho)$ and the compact set $\mathcal{C} \subset \mathcal{X}$ of order $\kappa$ and type $\lambda$ with generalized dimension $\mathfrak{d}$. We say that $(\mathcal{C}, \rho)$ is optimally representable by the decoder $\mathcal{D}$, if $(\mathcal{C}, \rho)$ is representable by $\mathcal{D}$ and*

$$\limsup_{\epsilon \to 0} \frac{\log^{(\kappa)} L(\epsilon; \mathcal{D}, \mathcal{C}, \rho)}{\log^{\tau}(\epsilon^{-1})} = \mathfrak{d}. \tag{6}$$

We now recall a remarkable universal optimality property of ReLU networks, namely all the function classes listed in Table I are optimally representable, in the sense of Definition II.14, by the canonical neural network decoder. This is a simple reformulation of the results in [7]; we provide the details of this reformulation in Appendix B. In the present paper, we establish that RNNs (Definition II.4), with inner ReLU networks, extend this universality to the approximation of nonlinear dynamical systems.

### *C. Lipschitz Fading-Memory Systems*

We proceed to characterize the class of dynamical systems we are interested in and start by defining their domain.

**Definition II.15.** *For fixed $D > 0$, we denote the set of admissible input signals by $\mathcal{S} := [-D, D]^{\mathbb{Z}}$, that is, for every $x[\cdot] \in \mathcal{S}$, it holds that $|x[t]| \leq D$, $\forall t \in \mathbb{Z}$.*

The quantity $D > 0$ is taken to be fixed throughout the paper and the dependence of $\mathcal{S}$ on $D$ is not explicitly indicated.

First, the systems $G : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}}$ we consider are causal.

**Definition II.16** (Causality)**.** *A system $G : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}}$ is causal, if for each $T \in \mathbb{Z}$, for every pair $x, x' \in \mathcal{S}$ with $x[t] = x'[t], \forall t \leq T$, it holds that $(Gx)[T] = (Gx')[T]$.*

Second, we demand time-invariance.

**Definition II.17** (Time-invariance)**.** *A system $G : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}}$ is time-invariant, if for every $\tau \in \mathbb{Z}$, it holds that*

$$\mathbf{T}_\tau G x = G \mathbf{T}_\tau x, \qquad \forall x \in \mathcal{S},$$

*with the shift operator $\mathbf{T}_\tau : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ defined as $(\mathbf{T}_\tau x)[t] := x[t - \tau]$.*

Next, we follow Volterra, who suggested that [19, p. 188] "a first extremely natural postulate is to suppose that the influence of the [input] a long time before the given moment gradually fades out." This property was termed "fading memory" in [21], and here we introduce a more quantitative version thereof, namely the concept of "Lipschitz fading memory" describing the speed at which system memory fades. This definition is inspired by examples in [8], [27]–[29], which will be discussed in more detail later.

**Definition II.18** (Lipschitz fading-memory)**.** *We say that $(w[t])_{t \geq 0}$ is a weight sequence if it is non-increasing and satisfies $w[t] \in (0, 1], \forall t \geq 0$, and $\lim_{t \to \infty} w[t] = 0$. A system $G : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}}$ has Lipschitz fading-memory with respect to the weight sequence $w$ if*

$$|(Gx)[t] - (Gy)[t]| \leq \sup_{\tau \geq 0} |w[\tau](x[t - \tau] - y[t - \tau])|, \quad \forall t \in \mathbb{Z}, \ \forall x, y \in \mathcal{S}.$$

The class of Lipschitz fading-memory (LFM) systems considered in the remainder of the paper can now formally be defined as follows.

**Definition II.19** (Lipschitz fading-memory systems)**.** *Given a weight sequence $w[\cdot]$, we define*

$$\begin{aligned} \mathcal{G}(w) := \{G : \mathcal{S} \to \mathbb{R}^{\mathbb{Z}} \, | G \text{ is causal, time-invariant, has Lipschitz fading-memory w.r.t. } w, \\ \text{and satisfies } (G0)[t] = 0, \ \forall t \in \mathbb{Z}\}. \end{aligned} \tag{7}$$

As we will want to approximate LFM systems $G \in \mathcal{G}(w)$ by RNNs, we need a metric that quantifies approximation quality. This metric should take into account that the RNNs we consider start running at time $t = 0$ and will, moreover, be of worst-case nature.

**Definition II.20.** *Let $\mathcal{S}^+ := \{s \in \mathcal{S} \, | \, s[t] = 0, \forall t < 0\}$. For $G, G' \in \{\mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}\}$ we define the metric*

$$\rho_*(G, G') = \sup_{x \in \mathcal{S}^+} \sup_{t \in \mathbb{N}} |(Gx)[t] - (G'x)[t]|.$$

We hasten to add that the restriction to one-sided input signals in Definition II.20 and to taking the supremum over $t \in \mathbb{N}$ in the output signals does not impact the hardness of the approximation task as shown by the next result.

**Lemma II.21.** *Let $(w[t])_{t\geq 0}$ be a weight sequence. For $G, G' \in \mathcal{G}(w)$, we have*

$$\rho_*(G, G') = \sup_{x\in\mathcal{S}} \sup_{t\in\mathbb{Z}} |(Gx)[t] - (G'x)[t]|.$$

*Proof.* See Appendix C-B. □

We are now ready to formally state the main goal of this paper, which is to prove that $(\mathcal{G}(w), \rho_*)$ is optimally representable by the canonical RNN decoder in Definition II.7. In fact, we will be seeking a quantitative version of this statement comparing the description complexity of the class $\mathcal{G}(w)$ to that of the RNNs approximating it.

## III. Approximation rates for LFM Systems

In this section, we study the ($\epsilon$-)scaling behavior of $N^{\text{ext}}(\epsilon; \mathcal{G}(w), \rho_*)$ for general weight sequences $w$. This will be effected by deriving an upper bound on $N^{\text{ext}}(\epsilon; \mathcal{G}(w), \rho_*)$ through the construction of a covering and a lower bound by identifying an explicit packing. We first define the concept of packings.

**Definition III.1.** *Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C} \subset \mathcal{X}$ compact. An $\epsilon$-packing for $(\mathcal{C}, \rho)$ is a set $\{x_1, x_2, \ldots, x_N\} \subset \mathcal{C}$ such that $\rho(x_i, x_j) > \epsilon$, for all distinct $i, j$. The $\epsilon$-packing number $M(\epsilon; \mathcal{C}, \rho)$ is the cardinality of a largest $\epsilon$-packing for $(\mathcal{C}, \rho)$.*

We shall frequently make use of the following two results relating the packing, covering, and exterior covering numbers.

**Lemma III.2** ([24], Theorem IV)**.** *Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C} \subset \mathcal{X}$ compact. For all $\epsilon > 0$, we have*

$$M(2\epsilon; \mathcal{C}, \rho) \leq N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \leq N(\epsilon; \mathcal{C}, \rho) \leq M(\epsilon; \mathcal{C}, \rho). \tag{8}$$

**Lemma III.3** ([24], p. 93)**.** *Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be metric spaces and consider the compact subsets $\mathcal{C}_{\mathcal{X}} \subset \mathcal{X}$ and $\mathcal{C}_{\mathcal{Y}} \subset \mathcal{Y}$. Assume that there exists an isometric isomorphism $f : \mathcal{C}_{\mathcal{X}} \to \mathcal{C}_{\mathcal{Y}}$, i.e., $f$ is bijective and for every pair $a, b \in \mathcal{C}_{\mathcal{X}}$, one has $\rho_{\mathcal{Y}}(f(a), f(b)) = \rho_{\mathcal{X}}(a, b)$. Then,*

$$N(\epsilon; \mathcal{C}_{\mathcal{X}}, \rho_{\mathcal{X}}) = N(\epsilon; \mathcal{C}_{\mathcal{Y}}, \rho_{\mathcal{Y}}) \quad \text{and} \quad M(\epsilon; \mathcal{C}_{\mathcal{X}}, \rho_{\mathcal{X}}) = M(\epsilon; \mathcal{C}_{\mathcal{Y}}, \rho_{\mathcal{Y}}). \tag{9}$$

Lemma III.3 will allow us to work with a simplified metric space $(\mathcal{G}_0(w), \rho_0)$ instead of the original one $(\mathcal{G}(w), \rho_*)$. Concretely, we exploit the properties of LFM systems to effect this reduction as follows. First, as LFM systems are causal, their output at time $t$ depends on the history of inputs up to and including time $t$ only. Second, time-invariance implies that the map taking the history of the input signal to the current output at time $t$ does not change with $t$ and we can therefore restrict ourselves to $t = 0$ w.l.o.g.. Thus, the mapping realized by an LFM system is completely characterized by the response to signals in the set

$$\mathcal{S}^- := \{s \in \mathcal{S} \mid \forall \ell > 0 : s[\ell] = 0\}. \tag{10}$$

We now define the simplified metric space $(\mathcal{G}_0(w), \rho_0)$ according to

$$\mathcal{G}_0(w) := \{g : \mathcal{S}^- \to \mathbb{R} \mid |g(x) - g(x')| \leq \|x - x'\|_w, \ \forall x, x' \in \mathcal{S}^-, g(0) = 0\}, \tag{11}$$

where

$$\|x - x'\|_w := \sup_{t \geq 0} |w[t](x[-t] - x'[-t])| \tag{12}$$

and

$$\rho_0(g, g') = \sup_{x \in \mathcal{S}^-} |g(x) - g'(x)|. \tag{13}$$

Next, we define the projection operator $\mathcal{P} : \mathcal{S} \to \mathcal{S}^-$ as

$$(\mathcal{P}x)[t] = x[t] \cdot \mathbb{1}_{\{t \leq 0\}}$$

and formalize the isometric isomorphism between functionals $g \in \mathcal{G}_0(w)$ and systems $G \in \mathcal{G}(w)$ as follows.

**Lemma III.4.** *Let $w[\cdot]$ be a weight sequence. The map*

$$\mathcal{I} : \mathcal{G}_0(w) \to \mathcal{G}(w) \tag{14}$$

$$g \to G := (x \to \{g(\mathcal{P}\mathbf{T}_{-t}x)\}_{t \in \mathbb{Z}}) \tag{15}$$

*is an isometric isomorphism between $(\mathcal{G}_0(w), \rho_0)$ and $(\mathcal{G}(w), \rho_*)$. Furthermore, $N(\epsilon; \mathcal{G}_0(w), \rho_0) = N(\epsilon; \mathcal{G}(w), \rho_*)$ and $M(\epsilon; \mathcal{G}_0(w), \rho_0) = M(\epsilon; \mathcal{G}(w), \rho_*)$, for all $\epsilon > 0$.*

*Proof.* See Appendix C-C. □

In the remainder of this section, we first lower-bound $M(\epsilon; \mathcal{G}_0(w), \rho_0)$, then upper-bound $N(\epsilon; \mathcal{G}_0(w), \rho_0)$ and finally use Lemmata III.3 and III.4 to translate these bounds into bounds on $N^{\text{ext}}(\epsilon; \mathcal{G}_0(w), \rho_0)$. The lower bound is established as follows.

**Lemma III.5.** *Let $w[\cdot]$ be a weight sequence. The $\epsilon$-packing number of $(\mathcal{G}_0(w), \rho_0)$ satisfies*

$$\log M(\epsilon; \mathcal{G}_0(w), \rho_0) \geq \left( \prod_{\ell=0}^{T} \left\lceil \frac{2Dw[\ell]}{\epsilon} \right\rceil \right) - 1,$$

*where $T := \max\{T' \in \mathbb{N} \mid w[T'] > \frac{\epsilon}{2D}\}$.*

*Proof.* The proof is taken from [24] and is detailed, for completeness, in Appendix C-D. □

To upper-bound $N(\epsilon; \mathcal{G}_0(w), \rho_0)$, we construct an $\epsilon$-net for $(\mathcal{G}_0(w), \rho_0)$. This construction is again inspired by [24] but we need to modify it to ensure that the elements of the $\epsilon$-net can efficiently be realized by ReLU networks. To be specific, we employ piece-wise linear mappings to approximate LFM systems instead of piece-wise constant mappings as considered in [24], which requires significant adjustments to the proof in [24, Section 7.2]

We start by introducing the “spike” function $\phi : \mathbb{R}^d \to \mathbb{R}$ considered in [30], and defined as

$$\phi(z) = \max\{1 + \min\{z_1, \ldots, z_d, 0\} - \max\{z_1, \ldots, z_d, 0\}, 0\}. \tag{16}$$

An illustration of spike functions for $d = 1$ and $d = 2$ is provided in Figure 1. The idea of this "spike" function can be traced back to [31], where the convex set

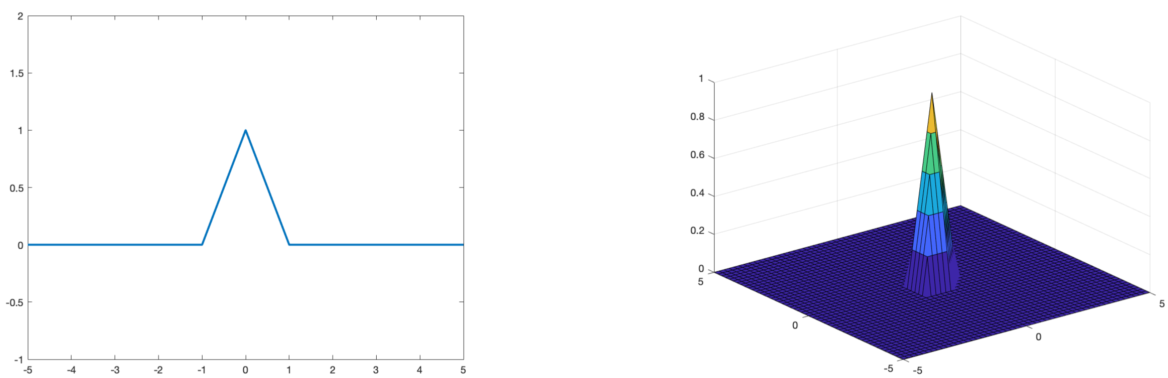
$$\{z \in \mathbb{R}^d : \max\{z_1, \ldots, z_d, 0\} - \min\{z_1, \ldots, z_d, 0\} \leq 1\} \tag{17}$$

is considered and shown to be the union of $(d+1)!$ simplices in the unit cube surrounding $0$ given by

$$\{z \in [-1, 1]^d : \; z_{\sigma(0)} \leq z_{\sigma(1)} \leq \cdots \leq z_{\sigma(d)}\}, \tag{18}$$

where $\sigma$ is a permutation of the integers $0, 1, \ldots, d$ and $z_0 := 0$. In [31], this result is employed to approximate continuous functions mapping $\mathbb{R}^d$ to $\mathbb{R}$ by functions that are piece-wise linear on the simplices in (18).

We remark that the spike function (16) is a composition of affine functions and min/max functions, which, as shown in Section V, renders it uniquely suitable for realization through ReLU networks. To be specific, in Lemmata C.2 and V.1, we provide concrete realizations of spike functions using ReLU networks. This construction is novel and distinct from the methodology considered in [30].




(a) $d = 1$ (b) $d = 2$

Fig. 1: The "spike" functions in dimensions 1 and 2.

We proceed to show how a *partition of unity* (p.o.u.) can be realized as a weighted linear combination of shifted spike functions, a property that will be of key importance in the RNN constructions described in Section V.

To this end, we consider the lattice

$$\mathfrak{M} = [\![M_1]\!]^{\pm} \times \cdots \times [\![M_d]\!]^{\pm} \subset \mathbb{R}^d, \text{ where } M_\ell \in \mathbb{N}, \text{ for } \ell \in \{1, 2, \ldots, d\}, \tag{19}$$

with the associated collection of shifted "spike" functions on $\mathbb{R}^d$

$$\Xi := \{\phi(\cdot - n)\}_{n \in \mathfrak{M}}. \tag{20}$$

The construction of the p.o.u. is as follows.

**Lemma III.6.** *Consider the spike function*

$$\phi(z) = \max\{1 + \min\{z_1, \dots, z_d, 0\} - \max\{z_1, \dots, z_d, 0\}, 0\}, \tag{21}$$

*the lattice*

$$\mathfrak{M} = [\![M_1]\!]^{\pm} \times \cdots \times [\![M_d]\!]^{\pm} \subset \mathbb{R}^d, \text{ where } M_\ell \in \mathbb{N}, \text{ for } \ell \in \{1, 2, \dots, d\}, \tag{22}$$

*and the set*

$$\Xi := \{\phi(\cdot - n)\}_{n \in \mathfrak{M}}. \tag{23}$$

*Then, $\Xi$ forms a p.o.u. on $\prod_{\ell=1}^d [-M_\ell, M_\ell]$, i.e.,*

*(i)* $0 \le \phi(z - n) \le 1$*, for* $z \in \mathbb{R}^d$ *and* $n \in \mathfrak{M}$*;*

*(ii)* $\phi(\cdot - n)$ *is compactly supported, specifically* $\operatorname{supp}(\phi(\cdot - n)) \subset n + [-1, 1]^d$*;*

*(iii) it holds that*

$$\sum_{n \in \mathfrak{M}} \phi(z - n) = 1, \quad \text{for } z \in \prod_{\ell=1}^d [-M_\ell, M_\ell].$$

*Proof.* To prove (i), we note that $\phi(z) \ge 0$ by definition and

$$1 + \min\{z_1, \dots, z_d, 0\} - \max\{z_1, \dots, z_d, 0\} \le 1.$$

To establish (ii), it suffices to prove that

$$\phi(z) = 0, \quad \text{for } z \in \mathbb{R}^d \setminus [-1, 1]^d.$$

To this end, we pick $z \in \mathbb{R}^d \setminus [-1, 1]^d$ arbitrarily and fix an arbitrary $\ell \in \{1, 2, \dots, d\}$ such that $|z_\ell| > 1$. Assume that $z_\ell > 1$. (The case $z_\ell < -1$ follows similarly.) Then,

$$1 + \min\{z_1, \dots, z_d, 0\} - \max\{z_1, \dots, z_d, 0\} \le 1 + 0 - z_\ell < 0,$$

which by (21) implies $\phi(z) = 0$.

We proceed to prove (iii). Since $\operatorname{supp}(\phi(\cdot - n)) \subset n + [-1, 1]^d$, we have

$$\sum_{n \in \mathfrak{M}} \phi(z - n) = \sum_{n \in \prod_{\ell=1}^d \{\lfloor z_\ell \rfloor, \lceil z_\ell \rceil\}} \phi(z - n), \quad \text{for } z \in \prod_{\ell=1}^d [-M_\ell, M_\ell].$$

Defining $\bar{z} \in \mathbb{R}^d$ with $\bar{z}_\ell = \lfloor z_\ell \rfloor$, for $\ell = 1, 2, \dots, d$, and noting that $\phi(z - n) = \phi((z - \bar{z}) - (n - \bar{z}))$, it suffices to show that

$$\sum_{n \in \{0,1\}^d} \phi(z - n) = 1, \quad \text{for } z \in [0, 1]^d \text{ and } n \in \{0, 1\}^d.$$

As $\min\{x_1, \dots, x_d\}$ and $\max\{x_1, \dots, x_d\}$ are permutation-invariant, so is $\phi$ by (21). We can therefore assume, w.l.o.g., that $z_1 \ge z_2 \ge \cdots \ge z_d$.

Now, set $e_k$ to be the $k$-th unit vector in $\mathbb{R}^d$ and let

$$A := \left\{0, e_1, e_1 + e_2, \dots, \sum_{i=1}^k e_i, \dots, \sum_{i=1}^d e_i\right\} \subset \{0, 1\}^d.$$

We claim that

$$\phi(z-n) = 0, \quad \text{for } z \in [0,1]^d \text{ and } n \in \{0,1\}^d \setminus A. \tag{24}$$

This can be verified as follows. First, thanks to

$$n \in \{0,1\}^d \setminus A \;\Leftrightarrow\; \exists\; i,j \in \{1,2,\dots,d\}, i<j, \text{ s.t. } n_i = 0, n_j = 1,$$

we get for $z \in [0,1]^d$ and $n \in \{0,1\}^d \setminus A$,

$$\begin{aligned}
&\min\{z_1-n_1, z_2-n_2, \dots, z_d-n_d, 0\} \le z_j - n_j = z_j - 1, \\
&\max\{z_1-n_1, z_2-n_2, \dots, z_d-n_d, 0\} \ge z_i - n_i = z_i, \\
&\Rightarrow 1 + \{z_1-n_1, z_2-n_2, \dots, z_d-n_d, 0\} - \max\{z_1-n_1, z_2-n_2, \dots, z_d-n_d, 0\} \le 0,
\end{aligned}$$

and hence $\phi(z-n) = 0$. It thus suffices to show that

$$\sum_{n \in A} \phi(z-n) = 1, \quad \text{for } z \in [0,1]^d. \tag{25}$$

Now, a direct calculation yields, for $z \in [0,1]^d$,

$$\phi(z) = 1 - z_1, \tag{26}$$

$$\phi\left(z - \sum_{i=1}^{k} e_i\right) = z_k - z_{k+1}, \quad \text{for } k = 1,\dots,d-1, \tag{27}$$

$$\phi\left(z - \sum_{i=1}^{d} e_i\right) = z_d. \tag{28}$$

Summing (26)-(28) proves (25). □

Next, we establish a traversal property of the lattice $\mathfrak{M}$.

**Definition III.7** (Regular path)**.** *For every $d \in \mathbb{N}$, $M_\ell \in \mathbb{N}$, $\ell \in \{1,\dots,d\}$, and corresponding lattice $\mathfrak{M} = [\![M_1]\!]^\pm \times \cdots \times [\![M_d]\!]^\pm$, we call a path $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{M}|}$ regular for $\mathfrak{M}$ if*

*(i) the path visits each grid point in $\mathfrak{M}$ exactly once,*

*(ii) $n_{i+1}$ and $n_i$ for $i = 1,\dots,|\mathfrak{M}|-1$, differ in exactly one position, specifically by $+1$ or $-1$.*

**Lemma III.8.** *For every $d \in \mathbb{N}$, $M_\ell \in \mathbb{N}$, $\ell \in \{1,\dots,d\}$, and corresponding lattice $\mathfrak{M} = [\![M_1]\!]^\pm \times \cdots \times [\![M_d]\!]^\pm$, there exists a regular path $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{M}|}$ for $|\mathfrak{M}|$.*

*Proof.* We write $\mathfrak{M}_d$ for the $d$-dimensional lattice $[\![M_1]\!]^\pm \times \cdots \times [\![M_d]\!]^\pm$ to emphasize the dependence on the dimension $d$ and prove the statement by induction over $d$. The base case $d = 1$ follows by simply considering the path $-M_1 \leftrightarrow -M_1+1 \leftrightarrow \cdots \leftrightarrow 0 \leftrightarrow \cdots \leftrightarrow M_1 - 1 \leftrightarrow M_1$ for lattice $\mathfrak{M}_1$. Assume now that the statement holds for $d = k$, i.e., there exists a path $n_1 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{M}_k|}$ that is regular for lattice $\mathfrak{M}_k$. For $d = k+1$, consider $\mathfrak{M}_{k+1} = [\![M_1]\!]^\pm \times \cdots \times [\![M_{k+1}]\!]^\pm$. Then, the path $(n_1, -M_{k+1}) \leftrightarrow \cdots \leftrightarrow (n_{|\mathfrak{M}_k|}, -M_{k+1}) \leftrightarrow (n_{|\mathfrak{M}_k|}, -M_{k+1}+1) \leftrightarrow \cdots \leftrightarrow (n_1, -M_{k+1}+1) \leftrightarrow (n_1, -M_{k+1}+2) \leftrightarrow \cdots \leftrightarrow (n_{|\mathfrak{M}_k|}, -M_{k+1}+2) \leftrightarrow \dots$ is regular for lattice $\mathfrak{M}_{k+1}$. □

We are now ready to describe our construction of the $\epsilon$-net for $(\mathcal{G}_0(w), \rho_0)$. In fact, we shall specify not only the elements of the $\epsilon$-net, but also the mapping taking a given functional $g \in \mathcal{G}_0(w)$ to an $\epsilon$-close element of the net. Counting the number of ball centers needed to ensure that every $g \in \mathcal{G}_0(w)$ is in the $\epsilon$-vicinity of some ball center then yields the cardinality of the $\epsilon$-net. The mapping proceeds by constructing a functional $\tilde{g}$ which is approximately faithful with respect to $g$ on amplitude-discretized and time-truncated input signals. Discretization reflects that we are interested only in an $\epsilon$-precise characterization of the action of the functional $g$ and truncation is rendered possible by the fading-memory property. The formal result is detailed as follows.

**Lemma III.9.** *For every $\epsilon > 0$ and $s \geq 1$, set*

$$T := \max\left\{ \ell \in \mathbb{N} \;\middle|\; w[\ell] > \frac{\epsilon}{D}\frac{s}{s+1} \right\},$$
$$\delta_\ell := \frac{s}{s+1}\frac{\epsilon}{w[\ell]}, \qquad \forall \ell \in [\![T]\!],$$
$$N_\ell := \left\lceil \frac{D}{\delta_\ell} \right\rceil, \quad \forall \ell \in [\![T]\!],$$

*and define the mapping*

$$k : S_- \to \mathbb{R}^{T+1}$$
$$k_\ell(x) := \frac{x[-\ell]}{\delta_\ell}, \qquad \forall \ell \in [\![T]\!].$$

*Furthermore, consider the lattice*

$$\mathfrak{N} := [\![N_0]\!]^{\pm} \times \cdots \times [\![N_T]\!]^{\pm} \subset \mathbb{R}^{T+1}.$$

*Then, there exists a set $\mathcal{U}_0 \subset \mathbb{R}^{|\mathfrak{N}|}$ with $|\mathcal{U}_0| = \left(2\left\lceil \frac{s+1}{2} \right\rceil + 1\right)^{|\mathfrak{N}|-1}$ such that*

$$\mathcal{U} := \left\{ \tilde{g} : S_- \to \mathbb{R} \;\middle|\; \tilde{g}(x) = \sum_{n \in \mathfrak{N}} \hat{g}_n \phi(k(x) - n),\ \{\hat{g}_n\}_{n \in \mathfrak{N}} \in \mathcal{U}_0 \right\}$$

*is an $\epsilon$-net for $(\mathcal{G}_0(w), \rho_0)$. Furthermore, it holds that*

$$|\mathcal{U}| = \left(2\left\lceil \frac{s+1}{2} \right\rceil + 1\right)^{|\mathfrak{N}|-1} = \left(2\left\lceil \frac{s+1}{2} \right\rceil + 1\right)^{\left[\prod_{\ell=0}^{T}\left(2\left\lceil \frac{Dw[\ell]}{\epsilon}\frac{s+1}{s} \right\rceil + 1\right)\right]-1}.$$

*Proof.* The construction proceeds in three steps.

*Step 1:* We pick amplitude-discretized and time-truncated signals from $\mathcal{S}^-$ on the lattice $\mathfrak{N}$ and approximately interpolate the functional $g \in \mathcal{G}_0(w)$ on these signals, using the elements of the p.o.u. $\Xi = \{\phi(\cdot - n)\}_{n \in \mathfrak{N}}$ as interpolation basis functions, to get a new functional $\hat{g}$.

Fix $g \in \mathcal{G}_0(w)$ arbitrarily. For each $n \in \mathfrak{N}$, define a signal $\widehat{x}_n \in \mathcal{S}^-$ according to

$$\widehat{x}_n[-\ell] := \begin{cases} \delta_\ell n_\ell, & \text{if } \ell \in [\![T]\!] \\ 0, & \text{if } \ell \in \mathbb{Z} \setminus [\![T]\!] \end{cases} \tag{29}$$

and let $\hat{g} : S_- \to \mathbb{R}$ be given by

$$\hat{g}(x) = \sum_{n \in \mathfrak{N}} g(\widehat{x}_n)\phi(k(x) - n). \tag{30}$$

We first prove that

$$|\hat{g}(x) - g(x)| \leq \frac{s\epsilon}{s+1}, \quad \text{for } x \in \mathcal{S}^-. \tag{31}$$

By Lemma III.6, $\{\phi(\cdot - n)\}_{n \in \mathfrak{N}}$ constitutes a p.o.u. on $\prod_{\ell=0}^{T}[-N_\ell, N_\ell]$, and hence

$$\phi(k(x) - n) = 0, \quad \text{for all } x \in \mathcal{S}^- \text{ s.t. } \|k(x) - n\|_\infty > 1. \tag{32}$$

Furthermore, for $x \in S_-$ and $n \in \mathfrak{N}$ such that $\|k(x) - n\|_\infty \leq 1$, we have

$$\left| \frac{x[-\ell]}{\delta_\ell} - n_\ell \right| \leq 1, \quad \text{for } \ell \in [\![T]\!],$$

which gives

$$\begin{aligned} w[\ell]\, |x[-\ell] - \widehat{x}_n[-\ell]| &\leq w[\ell]\delta_\ell = \frac{s\epsilon}{s+1}, \qquad \text{for } \ell \in [\![T]\!], \\ w[\ell]\, |x[-\ell] - \widehat{x}_n[-\ell]| &= w[\ell]\, |x[-\ell]| \\ &\leq \frac{\epsilon}{D}\frac{s}{s+1}D = \frac{s\epsilon}{s+1}, \quad \text{for } \ell > T, \end{aligned}$$

and therefore

$$\|x - \widehat{x}_n\|_w \leq \frac{s\epsilon}{s+1}. \tag{33}$$

Hence, we can bound the approximation error $|g(x) - \hat{g}(x)|$, for $x \in \mathcal{S}^-$, according to

$$|g(x) - \hat{g}(x)| \overset{(30),\text{ p.o.u.}}{=} \left| \sum_{n \in \mathfrak{N}} (g(x) - g(\widehat{x}_n))\phi(k(x) - n) \right| \tag{34}$$

$$\overset{(32)}{=} \left| \sum_{\|k(x)-n\|_\infty \leq 1} (g(x) - g(\widehat{x}_n))\phi(k(x) - n) \right| \tag{35}$$

$$\overset{(\phi \geq 0)}{\leq} \sum_{\|k(x)-n\|_\infty \leq 1} |(g(x) - g(\widehat{x}_n))|\, \phi(k(x) - n) \tag{36}$$

$$\overset{\text{Lipschitz}}{\leq} \sum_{\|k(x)-n\|_\infty \leq 1} \|x - \widehat{x}_n\|_w\, \phi(k(x) - n) \tag{37}$$

$$\overset{(33)}{\leq} \frac{s\epsilon}{s+1} \sum_{\|k(x)-n\|_\infty \leq 1} \phi(k(x) - n) \tag{38}$$

$$\overset{(32)}{=} \frac{s\epsilon}{s+1} \sum_{n \in \mathfrak{N}} \phi(k(x) - n) \tag{39}$$

$$\overset{\text{p.o.u.}}{=} \frac{s\epsilon}{s+1}. \tag{40}$$

*Step 2:* We modify the interpolation weights $g(\widehat{x}_n)$ in $\hat{g}$ to weights $\hat{g}_n$, for each $n \in \mathfrak{N}$, along a fixed traversal of $\mathfrak{N}$, to get a new functional $\tilde{g}$ that approximates $g$ to within an error of at most $\epsilon$.

Specifically, we construct, by induction, a functional $\tilde{g}$ such that

$$|\tilde{g}(x) - \hat{g}(x)| \leq \Delta := \frac{\epsilon}{s+1}. \tag{41}$$

To this end, we let

$$\tilde{g}(x) := \sum_{n \in \mathfrak{N}} \hat{g}_n \phi(k(x) - n) \tag{42}$$

and then set out to find, for every $n \in \mathfrak{N}$, a value $\hat{g}_n$ so that

$$|g(\widehat{x}_n) - \hat{g}_n| \leq \Delta, \tag{43}$$

which, in turn, yields

$$\begin{aligned}
|\hat{g}(x) - \tilde{g}(x)| &= \left| \sum_{n \in \mathfrak{N}} (g(\widehat{x}_n) - \hat{g}_n)\phi(k(x) - n) \right| \\
&\leq \sum_{n \in \mathfrak{N}} |g(\widehat{x}_n) - \hat{g}_n| \, \phi(k(x) - n) \\
&\leq \Delta \sum_{n \in \mathfrak{N}} \phi(k(x) - n) \\
&\overset{\text{p.o.u.}}{=} \Delta.
\end{aligned}$$

Therefore,

$$|\tilde{g}(x) - g(x)| \leq |\tilde{g}(x) - \hat{g}(x)| + |\hat{g}(x) - g(x)| \leq \Delta + \frac{s\epsilon}{s+1} = \epsilon. \tag{44}$$

It remains to specify how the values $\hat{g}_n$, $n \in \mathfrak{N}$, can be obtained such that (43) holds. This will be done by performing the mapping to $\hat{g}_n$ along a certain traversal of $\mathfrak{N}$. To this end, we note that by Lemma III.8, there exists a path $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{N}|}$ that is regular for $\mathfrak{N}$. We proceed with the proof by considering two cases.

- *Case 1: The path* $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{N}|}$ *starts or ends at* $0 \in \mathfrak{N}$.
  In particular, we assume, w.l.o.g., that $n_1 = 0$. Next, we find the values $\hat{g}_{n_k}$ satisfying (43) inductively over the index $k = 1, 2 \ldots, |\mathfrak{N}|$. The base case $k = 1$ is immediate as we can simply set $\hat{g}_{n_1} = 0$ and thereby obtain

  $$|g(\widehat{x}_{n_1}) - \hat{g}_{n_1}| \overset{(11)}{=} |g(0) - 0| = 0 \leq \Delta.$$

  Now, assume that (43) holds for some $n_k$, $k \in \{1, 2, \ldots, |\mathfrak{N}|\}$, on the path $n_1 \to \cdots \to n_{|\mathfrak{N}|}$. Then, by Lemma III.8, $n_{k+1}$ differs in exactly one position from $n_k$, namely by $+1$ or $-1$, which, based on (12) and (29), yields $\left\|\widehat{x}_{n_{k+1}} - \widehat{x}_{n_k}\right\|_w \leq \frac{\epsilon s}{s+1}$. Upon noting that

  $$\begin{aligned}
  \left|g\left(\widehat{x}_{n_{k+1}}\right) - \hat{g}_{n_k}\right| &= \left| \underbrace{g\left(\widehat{x}_{n_{k+1}}\right) - g\left(\widehat{x}_{n_k}\right)}_{\text{Lipschitz}} + \underbrace{g\left(\widehat{x}_{n_k}\right) - \hat{g}_{n_k}}_{(43)} \right| \\
  &\leq \left\|\widehat{x}_{n_{k+1}} - \widehat{x}_{n_k}\right\|_w + \Delta \\
  &\leq \frac{\epsilon s}{s+1} + \Delta = (s+1)\Delta,
  \end{aligned}$$

we can conclude that there exists an

$$m \in \left\{-2\left\lceil\frac{s+1}{2}\right\rceil, -2\left\lceil\frac{s+1}{2}\right\rceil+2, \dots, 2\left\lceil\frac{s+1}{2}\right\rceil\right\},$$
$$\text{s.t. } \hat{g}_{n_{k+1}} := \hat{g}_{n_k} + m\Delta \text{ and } \left|\hat{g}_{n_{k+1}} - g(\widehat{x}_{n_{k+1}})\right| \leq \Delta. \tag{45}$$

This completes the induction.

- *Case 2: The path* $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{N}|}$ *does not start or end at* $0 \in \mathfrak{N}$.
  In particular, we assume $n_i = 0 \in \mathfrak{N}$, for some $i \in \{2, 3, \dots, |\mathfrak{N}| - 1\}$. To follow the spirit of *Case 1*, we start the search of $\{\hat{g}_n\}_{n\in\mathfrak{N}}$ from $n_i = 0$. We then split the path $n_1 \leftrightarrow n_2 \leftrightarrow \cdots \leftrightarrow n_{|\mathfrak{N}|}$ into $\text{path}_{\leftarrow} := n_i \to n_{i-1} \to \cdots \to n_1$ and $\text{path}_{\to} := n_i \to n_{i+1} \to \cdots \to n_{|\mathfrak{N}|}$. The idea is to prove (43) by performing induction across $\text{path}_{\leftarrow}$ and $\text{path}_{\to}$ separately. This can be done following the same procedure as in *Case 1*.

*Step 3:* Repeat *Steps 1* and *2* for all $g \in \mathcal{G}_0(w)$ and collect all the resulting $\{\hat{g}_n\}_{n\in\mathfrak{N}}$ in a set $\mathcal{U}_0$. Specifically, for *Case 1*, we need to store one value for $\hat{g}_{n_1}$ and, according to (45), $\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)$ increments or decrements from $\hat{g}_{n_i}$ to $\hat{g}_{n_{i+1}}$, for $i = 0, 1, \dots, |\mathfrak{N}| - 1$. As the path $n_1 \to \cdots \to n_{|\mathfrak{N}|}$ is of length $|\mathfrak{N}| - 1$, we get a total of

$$\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{|\mathfrak{N}|-1}$$

values. For *Case 2*, noting that $\text{path}_{\leftarrow}$ and $\text{path}_{\to}$ are of length $i - 1$ and $|\mathfrak{N}| - i$, respectively, and applying the same argument as above, there is again a total of

$$\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{i-1}\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{|\mathfrak{N}|-i} = \left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{|\mathfrak{N}|-1}$$

increments or decrements. The set $\mathcal{U}_0$ containing all the values $\{\hat{g}_n\}_{n\in\mathfrak{N}}$ required for (43) to hold for all $g \in \mathcal{G}_0(w)$ and all $n \in \mathfrak{N}$ is hence of cardinality

$$|\mathcal{U}_0| = \left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{|\mathfrak{N}|-1} = \left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{\left[\prod_{\ell=0}^{T}\left(2\left\lceil\frac{Dw[\ell]}{\epsilon}\frac{s+1}{s}\right\rceil+1\right)\right]-1}. \tag{46}$$

Finally, setting

$$\mathcal{U} := \left\{\tilde{g} : S_- \to \mathbb{R} \;\middle|\; \tilde{g}(x) = \sum_{n\in\mathfrak{N}} \hat{g}_n \phi(k(x) - n),\ \{\hat{g}_n\}_{n\in\mathfrak{N}} \in \mathcal{U}_0\right\}$$

concludes the proof. □

Based on the $\epsilon$-net constructed in Lemma III.9, we can now upper-bound the metric entropy of $(\mathcal{G}_0(w), \rho_0)$ as follows.

**Corollary III.10.** *For every* $s \geq 1$, *the covering number of* $(\mathcal{G}_0(w), \rho_0)$ *satisfies*

$$\log N(\epsilon; \mathcal{G}_0(w), \rho_0) \leq \log\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)\prod_{\ell=0}^{T}\left(2\left\lceil\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s}\right\rceil+1\right),$$

*where* $T := \max\left\{\ell \in \mathbb{N} \mid w[\ell] > \frac{\epsilon}{2D}\frac{s}{s+1}\right\}$.

*Proof.* For $\epsilon > 0$, Lemma III.9 delivers an $\frac{\epsilon}{2}$-net for $(\mathcal{G}_0(w), \rho_0)$ with

$$\left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{\left[\prod_{\ell=0}^{T}\left(2\left\lceil\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s}\right\rceil+1\right)\right]-1}$$

elements. By Definition II.11, it hence follows that

$$N^{\text{ext}}(\epsilon/2; \mathcal{G}_0(w), \rho_0) \leq \left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{\left[\prod_{\ell=0}^{T}\left(2\left\lceil\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s}\right\rceil+1\right)\right]}.$$

From Lemma III.2, we then obtain an upper bound on the covering number according to

$$N(\epsilon; \mathcal{G}_0(w), \rho_0) \leq N^{\text{ext}}(\epsilon/2; \mathcal{G}_0(w), \rho_0) \leq \left(2\left\lceil\frac{s+1}{2}\right\rceil+1\right)^{\left[\prod_{\ell=0}^{T}\left(2\left\lceil\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s}\right\rceil+1\right)\right]},$$

which finishes the proof. □

We conclude the developments in this section by deriving a lower bound and an upper bound on the exterior covering number of $(\mathcal{G}(w), \rho_*)$.

**Theorem III.11.** *The exterior covering number of $(\mathcal{G}(w), \rho_*)$ satisfies*

$$\left(\prod_{\ell=0}^{T'}\left\lceil\frac{Dw[\ell]}{\epsilon}\right\rceil\right) - 1 \leq \log N^{ext}(\epsilon; \mathcal{G}(w), \rho_*) \leq \log(3)\prod_{\ell=0}^{T''}\left(2\left\lceil\frac{4Dw[\ell]}{\epsilon}\right\rceil+1\right), \tag{47}$$

*where $T' := \max\left\{\ell \in \mathbb{N} \mid w[\ell] > \frac{\epsilon}{D}\right\}$ and $T'' := \max\left\{\ell \in \mathbb{N} \mid w[\ell] > \frac{\epsilon}{4D}\right\}$.*

*Proof.* The result follows by noting that

$$\begin{aligned}
\left(\prod_{\ell=0}^{T'}\left\lceil\frac{Dw[\ell]}{\epsilon}\right\rceil\right) - 1 &\overset{\text{Lemma III.5}}{\leq} \log M(2\epsilon; \mathcal{G}_0(w), \rho_0) \\
&\overset{\text{Lemma III.4}}{=} \log M(2\epsilon; \mathcal{G}(w), \rho_*) \\
&\overset{\text{Lemma III.2}}{\leq} \log N^{\text{ext}}(\epsilon; \mathcal{G}(w), \rho_*) \leq \log N(\epsilon; \mathcal{G}(w), \rho_*) \\
&\overset{\text{Lemma III.4}}{=} \log N(\epsilon; \mathcal{G}_0(w), \rho_0) \\
&\overset{\substack{\text{Corollary III.10} \\ \text{for } s=1}}{\leq} \log(3)\prod_{\ell=0}^{T''}\left(2\left\lceil\frac{4Dw[\ell]}{\epsilon}\right\rceil+1\right).
\end{aligned}$$

□

## IV. Approximation rates for ELFM Systems and PLFM Systems

We now discuss two specific classes of LFM systems, namely exponentially Lipschitz fading-memory (ELFM) and polynomially Lipschitz fading-memory (PLFM) systems. Specifically, we characterize the description complexity of these two classes by computing their type, order, and generalized dimension. The corresponding results will then serve as a reference for the RNN approximations in Section V. Specifically, we will establish, in Section VI, that RNNs can learn ELFM and PLFM systems in a metric-entropy-optimal manner.

### A. Approximation rates for ELFM systems

The concept of exponentially Lipschitz fading memory systems is inspired, inter alia, by applications in finance, such as those discussed in [8], where asset pricing decisions are influenced by past observations. Instead of relying solely on finite-length observations, the model integrates infinite past observations by endowing them with exponentially decaying memory. Similar settings are also considered in random walk models [28], [29]. These examples, when appropriately adapted to the setup in the present paper, fit into the setting of our LFM systems (Definition II.19) with exponentially decaying weight sequences. We formally define exponentially decaying memory and the corresponding ELFM systems as follows.

**Definition IV.1.** *For $a \in (0,1]$ and $b > 0$, let*

$$w_{a,b}^{(e)}[t] := ae^{-bt}, \qquad \text{for all } t \geq 0.$$

*An LFM system with weight sequence $\{w_{a,b}^{(e)}[t]\}_{t\geq 0}$ is said to be exponentially Lipschitz fading-memory (ELFM). We write $\mathcal{G}(w_{a,b}^{(e)})$ for the class of all ELFM systems with weight sequence $w_{a,b}^{(e)}[t]$.*

The remainder of this section is devoted to computing the order, type, and generalized dimension of $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$. To this end, we first establish an auxiliary result.

**Lemma IV.2.** *Let $a \in (0,1]$, $b, c, d > 0$ and consider the weight sequence $w_{a,b}^{(e)}$ as per Definition IV.1. Set*

$$T := \max\left\{ t \in \mathbb{N} \;\middle|\; w_{a,b}^{(e)}[t] > \frac{\epsilon}{d} \right\}.$$

*Then,*

$$\log\left( \prod_{\ell=0}^{T} \frac{c w_{a,b}^{(e)}[\ell]}{\epsilon} \right) = \frac{1}{2b\log(e)} \log^2(\epsilon^{-1}) + o\left(\log^2(\epsilon^{-1})\right). \tag{48}$$

*Proof.* See Appendix C-E. □

We are now ready to state the main result of this section quantifying the massiveness of the class of ELFM systems.

**Lemma IV.3.** *Let $a \in (0,1]$ and $b > 0$. The class of ELFM systems $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ is of order $1$ and type $2$, with generalized dimension*

$$\mathfrak{d} = \frac{1}{2b\log(e)}.$$

*Proof.* Consider $\epsilon \in (0, \epsilon_0)$ with $\epsilon_0 = \frac{D w_{a,b}^{(e)}[0]}{2} = \frac{aD}{2}$. By Theorem III.11, the exterior covering number of $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ satisfies

$$\left( \prod_{\ell=0}^{T'} \left\lceil \frac{D w_{a,b}^{(e)}[\ell]}{\epsilon} \right\rceil \right) - 1 \leq \log N^{\text{ext}}(\epsilon; \mathcal{G}(w_{a,b}^{(e)}), \rho_*) \leq \log(3) \prod_{\ell=0}^{T''} \left( 2 \left\lceil \frac{4D w_{a,b}^{(e)}[\ell]}{\epsilon} \right\rceil + 1 \right), \tag{49}$$

where

$$T' := \max\left\{\ell \in \mathbb{N} \;\middle|\; w_{a,b}^{(e)}[\ell] > \frac{\epsilon}{D}\right\} \text{ and } T'' := \max\left\{\ell \in \mathbb{N} \;\middle|\; w_{a,b}^{(e)}[\ell] > \frac{\epsilon}{4D}\right\}.$$

We can further lower-bound the left-most term in (49) according to

$$\begin{aligned}\left(\prod_{\ell=0}^{T'}\left\lceil\frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right\rceil\right)-1 &\geq \left(\prod_{\ell=0}^{T'}\frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right)-1 \\ &\geq \frac{1}{2}\prod_{\ell=0}^{T'}\frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon}, \qquad (50)\end{aligned}$$

where (50) follows from

$$\frac{1}{2}\prod_{\ell=0}^{T'}\frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon} \geq 1,$$

which, in turn, is a consequence of

$$\begin{aligned}\frac{Dw_{a,b}^{(e)}[0]}{\epsilon} &\geq \frac{Dw_{a,b}^{(e)}[0]}{\epsilon_0} = 2, \\ \frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon} &\geq \frac{Dw_{a,b}^{(e)}[T']}{\epsilon} > 1, \quad \text{for} \quad \ell \in [\![T']\!] \setminus \{0\}.\end{aligned}$$

Similarly, we can further upper-bound the right-most term in (49) according to

$$\begin{aligned}\log(3)\left(\prod_{\ell=0}^{T''}\left(2\left\lceil\frac{4Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right\rceil+1\right)\right) &\leq \log(3)\left(\prod_{\ell=0}^{T''}\left(\frac{8Dw_{a,b}^{(e)}[\ell]}{\epsilon}+3\right)\right) \\ &\leq \log(3)\left(\prod_{\ell=0}^{T''}\frac{20Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right), \qquad (51)\end{aligned}$$

where (51) follows from

$$\frac{4Dw_{a,b}^{(e)}[\ell]}{\epsilon} \geq \frac{4Dw_{a,b}^{(e)}[T'']}{\epsilon} > 1, \quad \text{for} \quad \ell \in [\![T'']\!].$$

Combining (49)–(51), taking logarithms one more time, and dividing the results by $\log^2(\epsilon^{-1})$, yields

$$\frac{\log\left(\prod_{\ell=0}^{T'}\frac{Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right)-1}{\log^2(\epsilon^{-1})} \leq \frac{\log^{(2)} N^{\text{ext}}(\epsilon;\mathcal{G}(w_{a,b}^{(e)}),\rho_*)}{\log^2(\epsilon^{-1})} \leq \frac{\log\left(\prod_{\ell=0}^{T''}\frac{20Dw_{a,b}^{(e)}[\ell]}{\epsilon}\right)+\log^{(2)}(3)}{\log^2(\epsilon^{-1})}. \qquad (52)$$

Taking the limit $\epsilon \to 0$ and applying Lemma IV.2 to the lower and the upper bound in (52), we obtain

$$\lim_{\epsilon\to 0}\frac{\log^{(2)} N^{\text{ext}}(\epsilon;\mathcal{G}(w_{a,b}^{(e)}),\rho_*)}{\log^2(\epsilon^{-1})} = \frac{1}{2b\log(e)},$$

which implies that $(\mathcal{G}(w_{a,b}^{(e)}),\rho_*)$ is of order 1 and type 2, with generalized dimension

$$\mathfrak{d} = \frac{1}{2b\log(e)}.$$

This concludes the proof. □

The generalized dimension being inverse proportional to $b$ reflects that faster memory decay rates, i.e., larger $b$, result in system classes that are less massive. Additionally, we note that the description complexity of ELFM systems is primarily determined by the order being equal to 1 and the type equal to 2. Compared to the function classes in Table I, which are all of order 1 and type 1, this shows that the class of ELFM systems is significantly more massive than unit balls in function spaces.

### *B. Approximation rates for PLFM systems*

Next, we consider polynomially decaying memory, a concept used, e.g., in the context of PDEs [32], [33]. Specifically, the references [32], [33] are concerned with Volterra integro-differential equations [27] of polynomially decaying memory kernels. These examples, when suitably adjusted to the framework in the present paper, align with our LFM systems (Definition II.19) of polynomially decaying weight sequences. We formalize the concept of polynomially decaying memory and polynomially Lipschitz fading-memory (PLFM) systems as follows.

**Definition IV.4.** *For $q \in (0, 1]$ and $p > 0$, let*

$$w^{(p)}_{p,q}[t] := \frac{q}{(1+t)^p}, \qquad \text{for all } t \geq 0.$$

*An LFM system with weight sequence $\{w^{(p)}_{p,q}[t]\}_{t\geq 0}$ is said to be polynomially Lipschitz fading-memory (PLFM). We write $\mathcal{G}(w^{(p)}_{p,q})$ for the class of all PLFM systems with weight sequence $w^{(p)}_{p,q}[t]$.*

As in the previous section, we first need an auxiliary result.

**Lemma IV.5.** *Let $q \in (0, 1]$, $p, c, d > 0$ and consider $w^{(p)}_{p,q}$ as per Definition IV.4. Set*

$$T := \max\left\{ t \in \mathbb{N} \;\middle|\; w^{(p)}_{p,q}[t] > \frac{\epsilon}{d} \right\}.$$

*Then,*

$$\log\left( \prod_{\ell=0}^{T} \frac{c w^{(p)}_{p,q}[\ell]}{\epsilon} \right) = \Theta(\epsilon^{-1/p}). \tag{53}$$

*Proof.* See Appendix C-F. □

We obtain the order, type, and generalized dimension of PLFM systems as follows.

**Lemma IV.6.** *Let $q \in (0, 1]$ and $p > 0$. The class of PLFM systems $(\mathcal{G}(w^{(p)}_{p,q}), \rho_*)$ is of order 2 and type 1, with generalized dimension*

$$\mathfrak{d} = \frac{1}{p}.$$

*Proof.* See Appendix C-G. □

Compared to the class of ELFM systems, which exhibits order 1, the class of PLFM systems is more massive as it has order 2. This is intuitively meaningful as polynomial decay is significantly slower than exponential decay, rendering PLFM systems to depend more strongly on past inputs. Additionally, the generalized dimension exhibiting inverse proportionality to $p$, reflects that faster (polynomial) decay, i.e., larger $p$, leads to smaller description complexity.

## V. Realizing LFM Systems through RNNs

We now proceed to realize the elements in the $\epsilon$-net for $(\mathcal{G}_0(w), \rho_0)$ constructed in Lemma III.9 by ReLU networks. Based on the connection between $(\mathcal{G}_0(w), \rho_0)$ and $(\mathcal{G}(w), \rho_*)$ identified in Section III, this will then allow us to construct RNNs approximating general LFM systems.

### A. Approximating $(\mathcal{G}_0(w), \rho_0)$ with ReLU networks

The building block in Lemma III.9 for approximating functionals in $\mathcal{G}_0(w)$ is the p.o.u. $\Xi = \{\phi(\cdot - n)\}_{n\in\mathfrak{M}}$, which is why we first focus on constructing ReLU networks realizing $\Xi$. As the elements of $\Xi$ are shifted versions of the spike function $\phi$ and shifts, by virtue of being affine transformations, are trivially realized by single-layer ReLU networks, it suffices to find a ReLU network realization of $\phi$. Note that this argument made use of the fact that compositions of ReLU networks are again ReLU networks (see Lemma C.3).

**Lemma V.1.** *For $d \in \mathbb{N}$, consider the spike function $\phi : \mathbb{R}^d \to \mathbb{R}$,*

$$\phi(z) = \max\{1 + \min\{z_1, \dots, z_d, 0\} - \max\{z_1, \dots, z_d, 0\}, 0\}. \tag{54}$$

*There exists a ReLU network $\Phi \in \mathcal{N}_{d,1}$ with $\mathcal{L}(\Phi) = \lceil \log(d+1) \rceil + 4$, $\mathcal{M}(\Phi) \le 60d - 28$, $\mathcal{W}(\Phi) \le 6d$, $\mathcal{K}(\Phi) = \{1, -1\}$, and $\mathcal{B}(\Phi) = 1$, such that*

$$\Phi(z) = \phi(z), \qquad \text{for all} \quad z \in \mathbb{R}^d. \tag{55}$$

*Proof.* Let $\Phi_d^{max}$ be the ReLU network realization of $\max\{z_1, z_2, \dots, z_d\}$ according to Lemma C.2. Then, using $\min\{z_1, z_2, \dots, z_d\} = -\max\{-z_1, -z_2, \dots, -z_d\}$, we obtain

$$\begin{aligned} \phi(z) &= \rho(1 - \rho(\Phi_d^{max}(-z)) - \rho(\Phi_d^{max}(z))) \\ &= (\underbrace{(W_3 \circ \rho \circ W_2 \circ \rho \circ P(\Phi_d^{max}, \Phi_d^{max}))}_{=:\Phi_2} \circ \underbrace{W_1}_{=:\Phi_1})(z), \end{aligned} \tag{56}$$

where $W_1(z) = \begin{pmatrix} \mathbb{I}_d & -\mathbb{I}_d \end{pmatrix}^T z$, $W_2(z) = \begin{pmatrix} -1 & -1 \end{pmatrix} z + 1$, $W_3(z) = z$, and $P(\Phi_d^{max}, \Phi_d^{max})$ is the parallelization of two $\Phi_d^{max}$-networks according to Lemma C.4 such that $P(\Phi_d^{max}, \Phi_d^{max})(z) = (\Phi_d^{max}(z), \Phi_d^{max}(z))^T$. Now, by Lemma C.3, there exists a ReLU network $\Phi$ realizing $\Phi_2 \circ \Phi_1$ in (56), with

$$
\begin{aligned}
\mathcal{L}(\Phi) &\overset{\text{Lemma C.3}}{=} \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_1) \\
&\overset{\text{Lemma C.2, C.4}}{=} \lceil \log(d+1) \rceil + 4, \\
\mathcal{M}(\Phi) &\overset{\text{Lemma C.3}}{\leq} 2\mathcal{M}(\Phi_1) + 2\mathcal{M}(\Phi_2) \\
&\overset{\text{Lemma C.4}}{=} 2\mathcal{M}(W_3) + 2\mathcal{M}(W_2) + 2\mathcal{M}(W_1) + 4\mathcal{M}(\Phi_d^{max}) \\
&\overset{\text{Lemma C.2}}{\leq} 60d - 28, \\
\mathcal{W}(\Phi) &\overset{\text{Lemma C.3}}{\leq} \max\left\{4d, \max\{\mathcal{W}(\Phi_2), \mathcal{W}(\Phi_1)\}\right\} \\
&\overset{\text{Lemma C.4}}{\leq} \max\left\{4d, 2\mathcal{W}(\Phi_d^{max}), \mathcal{W}(W_3), \mathcal{W}(W_2), \mathcal{W}(W_1)\right\} \\
&\overset{\text{Lemma C.2}}{\leq} 6d, \\
\mathcal{K}(\Phi) &\overset{\text{Lemma C.3}}{\subset} (\mathcal{K}(\Phi_2) \cup (-\mathcal{K}(\Phi_2) \cup (\mathcal{K}(\Phi_1) \cup (-\mathcal{K}(\Phi_1) \\
&\overset{\text{Lemma C.4}}{=} \left( \bigcup_{i=1}^{3} (\mathcal{K}(W_i) \cup (-\mathcal{K}(W_i)) \right) \bigcup (\mathcal{K}(\Phi_d^{max}) \cup (-\mathcal{K}(\Phi_d^{max})) \\
&\overset{\text{Lemma C.2}}{=} \{1, -1\}, \\
\mathcal{B}(\Phi) &= \max_{b \in \mathcal{K}(\Phi)} |b| = 1.
\end{aligned}
$$

This concludes the proof. □

We now show how the elements of the $\epsilon$-net constructed in Lemma III.9 can be realized by ReLU networks. In particular, our constructions will be seen to be encodable by bitstrings of finite length, accomplished by quantizing the network weights according to Definition II.2. This will be needed to establish metric-entropy optimality in Section VI.

**Lemma V.2.** *For every $s \geq 1$, there exists $\epsilon_0 > 0$, such that for $\epsilon \in (0, \epsilon_0)$, with*

$$
T := \max\left\{ \ell \in \mathbb{N} \;\middle|\; w[\ell] > \frac{\epsilon}{D} \frac{s}{s+1} \right\},
$$

*for every $g \in \mathcal{G}_0(w)$, there exists a $\Phi \in \mathcal{N}_{T+1,1}$ with $(2, \epsilon)$-quantized weights (Definition II.2) satisfying*

$$
\left| \Phi\left( \{x[-\ell]\}_{\ell=0}^{T} \right) - g(x) \right| \leq \epsilon, \qquad \text{for all } x \in \mathcal{S}^-.
$$

*Moreover,*

$$
\mathcal{L}(\Phi) = \lceil \log(T+2) \rceil + 6 \;\; \text{and} \;\; \mathcal{M}(\Phi) \leq 244(T+1) \prod_{\ell=0}^{T} \left( \frac{2Dw[\ell]}{\epsilon} \frac{s+1}{s} + 4 \right). \tag{57}
$$

*Proof.* Fix $g \in \mathcal{G}_0(w)$. To construct a $(2, \epsilon)$-quantized ReLU network for the approximation of $g$, we follow the spirit of the proof of Lemma III.9 and first consider

$$
\hat{g}(x) = \sum_{n \in \mathfrak{N}} g(\widehat{x}_n) \phi \left( \left\{ \frac{x[-\ell]}{\delta_\ell} - n_\ell \right\}_{\ell=0}^{T} \right), \tag{58}
$$

with

$$\begin{aligned}\delta_\ell &:= \frac{s}{s+1}\frac{\epsilon}{w[\ell]}, \qquad \forall \ell \in [\![T]\!]\,,\\ N_\ell &:= \left\lceil \frac{D}{\delta_\ell} \right\rceil, \quad \forall \ell \in [\![T]\!]\,,\\ \mathfrak{N} &:= [\![N_0]\!]^{\pm} \times \cdots \times [\![N_T]\!]^{\pm} \subset \mathbb{R}^{T+1},\\ \widehat{x}_n[-\ell] &:= \begin{cases} \delta_\ell n_\ell, & \text{if } \ell \leq T\ ,\\ 0, & \text{else.}\end{cases}\end{aligned} \tag{59}$$

It was shown in (31) that

$$|\hat{g}(x) - g(x)| \leq \frac{s\epsilon}{s+1}, \quad \text{for all } x \in \mathcal{S}^-. \tag{60}$$

We next quantize the parameters $\delta_\ell^{-1}$ in (58) according to

$$\widetilde{\delta}_\ell^{-1} := \mathcal{Q}_{2,\epsilon}(\delta_\ell^{-1}), \tag{61}$$

and adjust the grid points $\widehat{x}_n$ of the lattice $\mathfrak{N}$ as

$$\begin{aligned}\widetilde{N}_\ell &= \left\lceil \frac{D}{\widetilde{\delta}_\ell} \right\rceil, \quad \ell \in [\![T]\!]\,,\\ \widetilde{\mathfrak{N}} &= \left[\!\!\left[\widetilde{N}_0\right]\!\!\right]^{\pm} \times \cdots \times \left[\!\!\left[\widetilde{N}_T\right]\!\!\right]^{\pm},\\ \widetilde{x}_n[-\ell] &= \begin{cases} \widetilde{\delta}_\ell n_\ell & \text{if } \ell \leq T,\\ 0 & \text{else.}\end{cases}\end{aligned} \tag{62}$$

Furthermore, we quantize $g(\widetilde{x}_n)$ according to

$$\widetilde{g}_n := \mathcal{Q}_{2,\epsilon}(g(\widetilde{x}_n)) \tag{63}$$

and consider the function

$$\tilde{g}(x) = \sum_{n \in \widetilde{\mathfrak{N}}} \widetilde{g}_n \phi\left(\left\{\frac{x[-\ell]}{\widetilde{\delta}_\ell} - n_\ell\right\}_{\ell=0}^{T}\right) =: f\left(\{x[-\ell]\}_{\ell=0}^{T}\right). \tag{64}$$

For ease of notation, we define $k : S_- \to \mathbb{R}^{T+1}$ as $k_\ell(x) := \widetilde{\delta}_\ell^{-1} x[-\ell]$, for $\ell \in [\![T]\!]$. Next, we show that

$$|\tilde{g}(x) - g(x)| \leq \epsilon. \tag{65}$$

This follows from

$$
\begin{aligned}
|\tilde{g}(x) - g(x)| = & \left| \sum_{\|k(x)-n\|_\infty \le 1} (g(x) - \widetilde{g}_n)\phi(k(x) - n) \right| && (66)\\
\le & \sum_{\|k(x)-n\|_\infty \le 1} \left( |(g(x) - g(\widetilde{x}_n))| + |(g(\widetilde{x}_n) - \widetilde{g}_n)| \right) \phi(k(x) - n) && (67)\\
\le & \sum_{\|k(x)-n\|_\infty \le 1} \left( \|x - \widetilde{x}_n\|_w + \frac{\epsilon}{s+1} \right) \phi(k(x) - n) && (68)\\
\le & \; \epsilon \sum_{\|k(x)-n\|_\infty \le 1} \phi(k(x) - n) && (69)\\
\le & \; \epsilon. && (70)
\end{aligned}
$$

Here, (66) and (70) are by the p.o.u. property of $\phi$ and (68) is a consequence of the Lipschitz property of $g$ according to (11) and

$$
\begin{aligned}
|g(\widetilde{x}_n) - \widetilde{g}_n| &= |g(\widetilde{x}_n) - \mathcal{Q}_{2,\epsilon}(g(\widetilde{x}_n))| \\
&\le 2^{-2\lceil \log(\epsilon^{-1}) \rceil} \overset{(189)}{\le} \frac{\epsilon}{s+1}.
\end{aligned} \tag{71}
$$

Furthermore, (69) follows from the fact that for $\|k(x) - n\|_\infty \le 1$, we have

$$
\left| \frac{x[-\ell]}{\widetilde{\delta}_\ell} - n_\ell \right| \le 1, \quad \text{for all } \ell \in [\![T]\!],
$$

and hence

$$
\|x - \widetilde{x}_n\|_w \le \max_{\ell \in [\![T]\!]} \widetilde{\delta}_\ell w[\ell] \le \frac{s\epsilon}{s+1}, \tag{72}
$$

where the second inequality in (72) is by

$$
\widetilde{\delta}_\ell \overset{(61)}{=} \left( \mathcal{Q}_{2,\epsilon} \left( \delta_\ell^{-1} \right) \right)^{-1} \le \delta_\ell = \frac{s}{s+1} \frac{\epsilon}{w[\ell]}.
$$

Based on (64), we can rewrite (65) according to

$$
\left| f \left( \{x[-\ell]\}_{\ell=0}^T \right) - g(x) \right| \le \epsilon, \qquad \text{for all } x \in S_-. \tag{73}
$$

It hence suffices to construct a ReLU network $\Phi$ that realizes $f$, which then, thanks to (73), approximates $g$ to within an error of at most $\epsilon$. To this end, assume that $n^1, n^2, \ldots, n^{|\widetilde{\mathfrak{N}}|}$ is an arbitrary, but fixed, enumeration of the elements of $\widetilde{\mathfrak{N}}$. Set $\widetilde{W}^\Sigma(x) = \Lambda x$ and $\widehat{W}_{n^i}(x) := \widehat{B}x + \hat{b}_{n^i}$, with

$$
\begin{aligned}
\Lambda &= \begin{pmatrix} \widetilde{g}_{n^1} & \widetilde{g}_{n^2} & \ldots & \widetilde{g}_{n^{|\widetilde{\mathfrak{N}}|}} \end{pmatrix}, \\
\widehat{B} &= \operatorname{diag}\left( \widetilde{\delta}_0^{-1}, \widetilde{\delta}_1^{-1}, \ldots, \widetilde{\delta}_T^{-1} \right), \\
\hat{b}_{n^i} &= \begin{pmatrix} -n_0^i & -n_1^i & \ldots & -n_T^i \end{pmatrix}^T.
\end{aligned} \tag{74}
$$

Moreover, let $\Psi$ be a ReLU network realizing the spike function $\phi$ according to Lemma V.1 and define the following ReLU networks

$$\begin{aligned}\Phi_2 &:= \widetilde{W}^{\Sigma},\\ \Phi_{1,2}^{n^i} &:= \Psi, \qquad \text{for } i = 1, \dots, |\widetilde{\mathfrak{N}}|,\\ \Phi_{1,1}^{n^i} &:= \widehat{W}_{n^i}, \qquad \text{for } i = 1, \dots, |\widetilde{\mathfrak{N}}|.\end{aligned} \tag{75}$$

Now, we apply Lemma C.3 to compose $\Phi_{1,2}^{n^i}$ and $\Phi_{1,1}^{n^i}$ in order to realize the shifted versions of the spike function according to

$$\left(\Phi_{1,2}^{n^i} \circ \Phi_{1,1}^{n^i}\right)\left(\{x[-\ell]\}_{\ell=0}^{T}\right) = \phi\left(\left\{\frac{x[-\ell]}{\widetilde{\delta}_\ell} - n_\ell^i\right\}_{\ell=0}^{T}\right), \quad \text{for } i = 1, 2, \dots, |\widetilde{\mathfrak{N}}|.$$

Then, we apply Lemma C.4 to construct a ReLU network as the parallelization of the compositions $\Phi_{1,2}^{n^i} \circ \Phi_{1,1}^{n^i}$, for $i = 1, 2, \dots, |\widetilde{\mathfrak{N}}|$,

$$\Phi_1 := P\left(\left(\Phi_{1,2}^{n^1} \circ \Phi_{1,1}^{n^1}\right), \left(\Phi_{1,2}^{n^2} \circ \Phi_{1,1}^{n^2}\right), \dots, \left(\Phi_{1,2}^{n^{|\widetilde{\mathfrak{N}}|}} \circ \Phi_{1,1}^{|\widetilde{\mathfrak{N}}|}\right)\right). \tag{76}$$

Finally, we apply Lemma C.3 again to get a ReLU network that composes $\Phi_2$ and $\Phi_1$ according to

$$\Phi = \Phi_2 \circ \Phi_1 = \Phi_2 \circ P\left(\left(\Phi_{1,2}^{n^1} \circ \Phi_{1,1}^{n^1}\right), \left(\Phi_{1,2}^{n^2} \circ \Phi_{1,1}^{n^2}\right), \dots, \left(\Phi_{1,2}^{n^{|\widetilde{\mathfrak{N}}|}} \circ \Phi_{1,1}^{|\widetilde{\mathfrak{N}}|}\right)\right), \tag{77}$$

thereby realizing the linear combination

$$\sum_{n \in \widetilde{\mathfrak{N}}} \widetilde{g}_n \phi\left(\left\{\frac{x[-\ell]}{\widetilde{\delta}_\ell} - n_\ell\right\}_{\ell=0}^{T}\right) \overset{(64)}{=} f\left(\{x[-\ell]\}_{\ell=0}^{T}\right).$$

To conclude the proof, we verify that $\Phi$, indeed, has $(2, \epsilon)$-quantized weights, compute $\mathcal{L}(\Phi)$, and derive an upper bound on $\mathcal{M}(\Phi)$. We defer the corresponding details to Appendix C-I. □

### *B. Approximating the system space $(\mathcal{G}(w), \rho_*)$ using RNNs*

Having constructed ReLU networks that realize elements of $\mathcal{G}_0(w)$ according to Lemma V.2, we are now ready to describe the realization of systems in $\mathcal{G}(w)$ through RNNs. This will be done by employing the connection between $\mathcal{G}(w)$ and $\mathcal{G}_0(w)$, as established in Lemma III.4. Specifically, we construct RNNs that remember past inputs and produce approximations of the desired output

**Theorem V.3.** *For every $s \geq 1$, there exists $\epsilon_0 > 0$, such that for $\epsilon \in (0, \epsilon_0)$, with*

$$T := \max\left\{\ell \in \mathbb{N} \;\middle|\; w[\ell] > \frac{\epsilon}{D}\frac{s}{s+1}\right\},$$

*for every $G \in \mathcal{G}(w)$, there is an RNN $\mathcal{R}_\Psi$ associated with a ReLU network $\Psi \in \mathcal{N}_{T+1,T+1}$, satisfying*

$$\rho_*(\mathcal{R}_\Psi, G) \leq \epsilon.$$

*Moreover, $\mathcal{R}_\Psi$ has $(2, \epsilon)$-quantized weights and there exists a universal constant $C > 0$ such that*

$$\mathcal{M}(\Psi) \leq C(T+1)^2 \prod_{\ell=0}^{T}\left(\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s} + 4\right). \tag{78}$$

*Proof.* Fix $s \geq 1$ and $G \in \mathcal{G}(w)$ arbitrarily. We proceed in two steps.

*Step 1: We construct a ReLU network* $\Phi : \mathbb{R}^{T+1} \to \mathbb{R}$ *such that*

$$\sup_{x \in \mathcal{S}^+} \sup_{t \in \mathbb{N}} \left| \Phi\left(\{x[t-\ell]\}_{\ell=0}^{T}\right) - G(x)[t] \right| \leq \epsilon. \tag{79}$$

To this end, we first note that by Lemma III.4, one can find a $g \in \mathcal{G}_0(w)$ so that

$$g\left(\mathcal{P}\mathbf{T}_{-t}x\right) = G(x)[t], \quad \text{for all } x \in \mathcal{S} \text{ and } t \in \mathbb{Z}. \tag{80}$$

Furthermore, by Lemma V.2, there exists $\epsilon_0 > 0$, such that for every $\epsilon \in (0, \epsilon_0)$, there is a ReLU network $\Phi$ satisfying

$$\left| \Phi\left(\{z[-\ell]\}_{\ell=0}^{T}\right) - g(z) \right| \leq \epsilon, \quad \text{for all } z \in \mathcal{S}^-. \tag{81}$$

Next, fix an input $x \in \mathcal{S}^+$ and a time index $t \in \mathbb{N}$ and define

$$z' := \mathcal{P}\mathbf{T}_{-t}\{x\}. \tag{82}$$

Note that

$$\begin{aligned} z'[\ell] &= 0, \quad \text{for } \ell > 0, \text{ and hence } z' \in \mathcal{S}^-, \\ z'[-\ell] &= x[t-\ell], \qquad \text{for } \ell \geq 0. \end{aligned} \tag{83}$$

Inserting $z'$ from (82) into (81) and using (83) and (80), it follows that

$$\left| \Phi\left(\{x[t-\ell]\}_{\ell=0}^{T}\right) - G(x)[t] \right| \leq \epsilon.$$

As $x \in \mathcal{S}^-$ and $t \in \mathbb{N}$ were arbitrary, this proves (79).

*Step 2: We construct an RNN* $\mathcal{R}_\Psi$ *realizing the map* $x \to \Phi\left(\{x[t-\ell]\}_{\ell=0}^{T}\right)_{t \in \mathbb{N}}$.

Fix $x \in \mathcal{S}^+$ arbitrarily. Recall the RNN Definition II.4. The basic idea is to identify an RNN $\mathcal{R}_\Psi$ which, for every time step $t \in \mathbb{N}$,

- delivers the output
$$y[t] = \Phi\left(\{x[t-\ell]\}_{\ell=0}^{T}\right) \tag{84}$$
- and memorizes the $T$ past inputs $x[t], x[t-1], \ldots, x[t-T+1]$ in the hidden state vector $h[t]$, i.e.,
$$h_\ell[t] = x[t-\ell+1], \qquad \text{for} \quad \ell \in \{1, 2, \ldots, T\}. \tag{85}$$

To memorize the $T$ past inputs, we construct a one-layer neural network

$$\Phi_T(z) = \begin{pmatrix} \mathbb{I}_T & 0_T \end{pmatrix} z, \quad \text{for } z \in \mathbb{R}^{T+1}, \tag{86}$$

by noting that

$$\begin{aligned} \Phi_T\left(\{x[t-\ell]\}_{\ell=0}^{T}\right) &= \begin{pmatrix} \mathbb{I}_T & 0_T \end{pmatrix} (x[t], x[t-1], \ldots, x[t-T])^T \\ &= (x[t], x[t-1], \ldots, x[t-T+1])^T \in \mathbb{R}^T. \end{aligned} \tag{87}$$

Now, we apply Lemma C.5 to augment $\Phi_T$ to depth $\mathcal{L}(\Phi)$ without changing its input-output relation. This results in the ReLU network $\Phi_T^*$. Then, we apply Lemma C.4 to parallelize $\Phi$ and $\Phi_T^*$ leading to the desired ReLU network

$$\Psi = P(\Phi, \Phi_T^*). \tag{88}$$

By Definition II.4, the corresponding RNN $\mathcal{R}_\Psi$ effects the input-output mapping according to

$$\begin{pmatrix} y[t] \\ h[t] \end{pmatrix} = \Psi\left(\begin{pmatrix} x[t] \\ h[t-1] \end{pmatrix}\right), \quad \text{for all } t \geq 0.$$

With these choices, (84) and (85) can now be proved by induction over $t \geq 0$. The base case is immediate, as $x[t] = 0$, for $t < 0$ owing to $x \in \mathcal{S}^+$ and, by Definition II.4, $h[-1] = 0_T$. To establish the induction step, we assume that (84) and (85) hold for $t-1$ with $t \in \mathbb{N}$, i.e.,

$$\begin{aligned} y[t-1] &= \Phi\left(\{x[t-1-\ell]\}_{\ell=0}^T\right), \\ h_\ell[t-1] &= x[t-\ell], \qquad \text{for} \quad \ell \in \{1, 2, \ldots, T\}. \end{aligned}$$

Now, for time step $t$, we note that

$$\begin{aligned} \Psi\left(\begin{pmatrix} x[t] \\ h[t-1] \end{pmatrix}\right) &= P(\Phi, \Phi_T^*)\left(\begin{pmatrix} x[t] \\ h[t-1] \end{pmatrix}\right) \\ &= \begin{pmatrix} \Phi\left(\{x[t-\ell]\}_{\ell=0}^T\right) \\ \Phi_T^*\left(\{x[t-\ell]\}_{\ell=0}^T\right) \end{pmatrix} \\ &= \begin{pmatrix} \Phi\left(\{x[t-\ell]\}_{\ell=0}^T\right) \\ (x[t], x[t-1], \ldots, x[t-T+1])^T \end{pmatrix} \\ &= \begin{pmatrix} \Phi\left(\{x[t-\ell]\}_{\ell=0}^T\right) \\ h[t] \end{pmatrix}. \end{aligned}$$

As $x$ was arbitrary, this completes the induction and thereby Step 2.

To conclude, we combine the results in Steps 1 and 2 according to

$$\rho_*(\mathcal{R}_\Psi, G) = \sup_{x \in \mathcal{S}^+} \sup_{t \in \mathbb{N}} |y[t] - G(x)[t]| \tag{89}$$

$$= \sup_{x \in \mathcal{S}^+} \sup_{t \in \mathbb{N}} \left|\Phi\left(\{x[t-\ell]\}_{\ell=0}^T\right) - G(x)[t]\right| \tag{90}$$

$$\leq \epsilon, \tag{91}$$

where (90) follows from (84) and (91) is by (79). Furthermore, we have

$$\begin{aligned} \mathcal{K}(\Psi) &\overset{\text{(88), Lemma C.4}}{=} (\mathcal{K}(\Phi)) \cup (\mathcal{K}(\Phi_T^*)) \\ &\overset{\text{Lemma C.5}}{\subset} (\mathcal{K}(\Phi)) \cup (\mathcal{K}(\Phi_T)) \cup (-\mathcal{K}(\Phi_T)) \cup \{1, -1\} \\ &\overset{(86)}{=} (\mathcal{K}(\Phi)) \cup \{1, -1\} \\ &\overset{\text{Lemma V.2}}{\subset} 2^{-2\lceil \log(\epsilon^{-1}) \rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \end{aligned}$$

Thus, $\Psi$ has $(2,\epsilon)$-quantized weights. Moreover, we can obtain an upper bound on $\mathcal{M}(\Psi)$ according to

$$\begin{aligned}\mathcal{M}(\Psi) &\overset{\text{(88), Lemma C.4}}{=} \mathcal{M}(\Phi)+\mathcal{M}(\Phi_T^*)\\ &\overset{\text{(86), Lemma C.5}}{\leq} \mathcal{M}(\Phi)+\mathcal{M}(\Phi_T)+T\mathcal{W}(\Phi_T)+2T(\mathcal{L}(\Phi)-\mathcal{L}(\Phi_T))\\ &\overset{\text{(86), Lemma V.2}}{\leq} C(T+1)^2\prod_{\ell=0}^{T}\left(\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s}+4\right),\end{aligned}$$

with the universal constant $C>0$ chosen sufficiently large. □

## VI. Metric-Entropy-Optimal Realizations of ELFM and PLFM systems

So far we have characterized the description complexity of ELFM and PLFM systems based on order, type, and generalized dimension (Section IV) and we constructed RNNs approximating general LFM systems (Section V). We are now ready to state the main results of the paper, namely that the RNNs we constructed are optimal for ELFM and PLFM system approximation in terms of description complexity.

To this end, we first compute the number of bits needed by the canonical RNN decoder in Definition II.7 to obtain the RNN constructed in Theorem V.3, specifically its topology and quantized weights. The following result holds for general LFM systems and will later be particularized to ELFM and PLFM systems.

**Corollary VI.1.** *The class of LFM systems $(\mathcal{G}(w),\rho_*)$ is representable by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ with*

$$L(\epsilon;\mathcal{D}_{\mathcal{R}},\mathcal{G}(w),\rho_*)\leq C_1 M\log(M)\log(\epsilon^{-1}),$$

*where*

$$M:=(T+1)^2\prod_{\ell=0}^{T}\left(\frac{12Dw[\ell]}{\epsilon}\right),$$
$$T:=\max\left\{\ell\in\mathbb{N}\;\middle|\;w[\ell]>\frac{\epsilon}{2D}\right\},$$

*and $C_1>0$ is a universal constant.*

*Proof.* Applying Theorem V.3 and setting $s=1$, it follows that there exists an $\epsilon_0>0$, such that for every $\epsilon\in(0,\epsilon_0)$ and every $G\in\mathcal{G}(w)$, we can find an RNN $\mathcal{R}_\Psi$ associated with a ReLU network $\Psi\in\mathcal{N}_{T+1,T+1}$, satisfying

$$\rho_*(\mathcal{R}_\Psi,G)\leq\epsilon. \tag{92}$$

Moreover, $\mathcal{R}_\Psi$ has $(2,\epsilon)$-quantized weights and the number of non-zero weights in $\Psi$ can be upper-bounded according to

$$\mathcal{M}(\Psi)\leq C(T+1)^2\prod_{\ell=0}^{T}\left(\frac{4Dw[\ell]}{\epsilon}+4\right). \tag{93}$$

By the definition of the canonical neural network decoder, Remark II.3, and Definition II.7, there exists a bitstring $b\in\{0,1\}^L$ with

$$L\leq 2C_0\mathcal{M}(\Psi)\log(\mathcal{M}(\Psi))\log(\epsilon^{-1}), \tag{94}$$

such that

$$\mathcal{D}_{\mathcal{R}}(b) = \mathcal{R}_{\Psi}. \tag{95}$$

Combining (92), (94), (95), and Definition II.9, upon noting that $G$ was chosen arbitrarily, it follows that $(\mathcal{G}(w), \rho_*)$ is representable by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$, with the minimum number of required bits satisfying

$$\begin{aligned}
L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w), \rho_*) \leq\ & 2C_0 \mathcal{M}(\Psi) \log(\mathcal{M}(\Psi)) \log(\epsilon^{-1}) && (96)\\
\leq\ & 2C_0 C (T+1)^2 \left[ \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right] \log(\epsilon^{-1}) \\
& \log \left( C(T+1)^2 \left( \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right) \right) && (97)\\
\leq\ & 2C_0 C C' (T+1)^2 \left[ \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right] \log(\epsilon^{-1}) \\
& \log \left( (T+1)^2 \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right) && (98)\\
=\ & C_1 M \log(M) \log(\epsilon^{-1}), && (99)
\end{aligned}$$

where (96) is a consequence of (94) and Definition II.9, (97) follows from (93) and the fact that $2Dw[\ell]/\epsilon > 1$, for $\ell \in [\![T]\!]$, (98) holds upon choosing the universal constant $C'$ sufficiently large, namely s.t.

$$\log \left( C(T+1)^2 \left( \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right) \right) \leq C' \log \left( (T+1)^2 \left( \prod_{\ell=0}^{T} \left( \frac{12Dw[\ell]}{\epsilon} \right) \right) \right),$$

and (99) follows by setting $C_1 := 2C_0 C C'$. □

### *A. RNNs can optimally learn ELFM systems*

We now particularize the result in Corollary VI.1 to ELFM systems, which allows us to determine the growth rate of the minimum number of bits $L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ needed to represent $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ with respect to the prescribed error $\epsilon$. A comparison with the description complexity of $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ established in Lemma IV.3 then allows us to conclude that RNNs can optimally learn ELFM systems in a metric-entropy optimal manner.

**Theorem VI.2.** *Let $a \in (0,1]$ and $b > 0$. The class of ELFM systems (Definition IV.1) $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ is optimally representable by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ (Definition II.7).*

*Proof.* By Corollary VI.1, the minimum number of bits $L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ needed to represent $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ satisfies

$$L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w_{a,b}^{(e)}), \rho_*) \leq C_1 M \log(M) \log(\epsilon^{-1}), \tag{100}$$

where

$$M := (T+1)^2 \prod_{\ell=0}^{T} \left( \frac{12D w_{a,b}^{(e)}[\ell]}{\epsilon} \right),$$
$$T := \max \left\{ \ell \in \mathbb{N} \;\middle|\; w_{a,b}^{(e)}[\ell] > \frac{\epsilon}{2D} \right\}.$$

By Lemma IV.3, $(\mathcal{G}(w_{a,b}^{(e)}), \rho_*)$ is of order $1$ and type $2$, with generalized dimension

$$\mathfrak{d} = \frac{1}{2b \log(e)}.$$

Thus, by Lemma II.13 and Definition II.14, it suffices to prove that

$$\lim_{\epsilon \to 0} \frac{\log \left( C_1 M \log(M) \log(\epsilon^{-1}) \right)}{\log^2(\epsilon^{-1})} = \frac{1}{2b \log(e)}. \tag{101}$$

To this end, we first note that

$$\log(M) = 2 \log(T+1) + \log \left( \prod_{\ell=0}^{T} \left( \frac{12D w_{a,b}^{(e)}[\ell]}{\epsilon} \right) \right) \tag{102}$$
$$= O(\log^{(2)}(\epsilon^{-1})) + \frac{1}{2b \log(e)} \log^2(\epsilon^{-1}) + o\left( \log^2(\epsilon^{-1}) \right) \tag{103}$$
$$= \frac{1}{2b \log(e)} \log^2(\epsilon^{-1}) + o\left( \log^2(\epsilon^{-1}) \right), \tag{104}$$

where (103) follows from Lemma IV.2 and $T+1 = O(\log(\epsilon^{-1}))$ thanks to (154). Now, we rewrite the numerator in (101) according to

$$\log \left( C_1 M \log(M) \log(\epsilon^{-1}) \right) = \log(M) + \log^{(2)}(M) + o\left( \log^2(\epsilon^{-1}) \right) \tag{105}$$
$$= \frac{1}{2b \log(e)} \log^2(\epsilon^{-1}) + o\left( \log^2(\epsilon^{-1}) \right) + \log \left( \frac{1}{2b \log(e)} \log^2(\epsilon^{-1}) + o\left( \log^2(\epsilon^{-1}) \right) \right) \tag{106}$$
$$= \frac{1}{2b \log(e)} \log^2(\epsilon^{-1}) + o\left( \log^2(\epsilon^{-1}) \right). \tag{107}$$

Dividing (107) by $\log^2(\epsilon^{-1})$ and taking $\epsilon \to 0$, concludes the proof. □

*B. RNNs can optimally learn PLFM systems*

We next particularize the result in Corollary VI.1 to PLFM systems and will see that the minimum number of bits $L(\epsilon; \mathcal{D}_\mathcal{R}, \mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ required to represent $(\mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ by the canonical RNN decoder $\mathcal{D}_\mathcal{R}$ grows significantly faster (with respect to the prescribed error $\epsilon$) than for ELFM systems. This can be attributed to the fact that the memory of PLFM systems decays much more slowly and the increased complexity reflects this longer memory. Nonetheless, as shown next, RNNs can learn PLFM systems in a metric-entropy-optimal manner.

**Theorem VI.3.** *Let $q \in (0,1]$ and $p > 0$. The class of PLFM systems $(\mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ (Definition IV.4) is optimally representable by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ (Definition II.7).*

*Proof.* By Corollary VI.1, the minimum number of bits $L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ needed to represent $(\mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ by the canonical RNN decoder $\mathcal{D}_{\mathcal{R}}$ satisfies

$$L(\epsilon; \mathcal{D}_{\mathcal{R}}, \mathcal{G}(w_{p,q}^{(p)}), \rho_*) \leq C_1 M \log(M) \log(\epsilon^{-1}), \tag{108}$$

where

$$M := (T+1)^2 \prod_{\ell=0}^{T} \left( \frac{12 D w_{p,q}^{(p)}[\ell]}{\epsilon} \right),$$
$$T := \max \left\{ \ell \in \mathbb{N} \;\middle|\; w_{p,q}^{(p)}[\ell] > \frac{\epsilon}{2D} \right\}.$$

By Lemma IV.6, $(\mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ is of order $2$ and type $1$, with generalized dimension

$$\mathfrak{d} = \frac{1}{p}.$$

Thus, by Lemma II.13 and Definition II.14, it suffices to prove that

$$\lim_{\epsilon \to 0} \frac{\log^{(2)} \left( C_1 M \log(M) \log(\epsilon^{-1}) \right)}{\log(\epsilon^{-1})} = \frac{1}{p}. \tag{109}$$

To this end, we first note that

$$\log(M) = 2\log(T+1) + \log\left( \prod_{\ell=0}^{T} \left( \frac{12 D w_{p,q}^{(p)}[\ell]}{\epsilon} \right) \right) \tag{110}$$
$$= O(\log(\epsilon^{-1})) + \Theta(\epsilon^{-1/p}) \tag{111}$$
$$= \Theta(\epsilon^{-1/p}), \tag{112}$$

where (111) follows from Lemma IV.5 and (160). Now, we rewrite the numerator in (109) according to

$$\log^{(2)} \left( C_1 M \log(M) \log(\epsilon^{-1}) \right) = \log\left( \log(M) + \log^{(2)}(M) + o\left( \log(\epsilon^{-1}) \right) \right) \tag{113}$$
$$= \log\left( \Theta(\epsilon^{-1/p}) + \log\left( \Theta(\epsilon^{-1/p}) \right) + o\left( \log(\epsilon^{-1}) \right) \right) \tag{114}$$
$$= \log\left( \Theta(\epsilon^{-1/p}) \right), \tag{115}$$

where (115) follows from

$$\Theta(\epsilon^{-1/p}) + \log\left( \Theta(\epsilon^{-1/p}) \right) + o\left( \log(\epsilon^{-1}) \right) = \Theta(\epsilon^{-1/p}).$$

Finally, dividing (115) by $\log(\epsilon^{-1})$ and taking $\epsilon \to 0$, concludes the proof. □

## VII. Conclusion

Returning to Table I, we note that it can be complemented by our results for ELFM and PLFM systems as summarized in Table II below. As both classes of systems in Table II are optimally representable by the canonical RNN decoder (Definition II.7), we can conclude that, remarkably, the metric-entropy-optimal universality of ReLU networks extends from function classes to nonlinear systems, in the latter case simply by embedding the network in a recurrence. Moreover, the description complexity of the LFM systems to be learned can be matched simply by adjusting the complexity of the inner ReLU network of the RNN.

| Metric | | $\mathcal{C}$ | | $k$ | $\lambda$ | $\mathfrak{d}$ |
|---|---|---|---|---|---|---|
| $\{\mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}\}$ | $\rho_*$ (Def. II.20) | ELFM systems (Section IV-A) | $\mathcal{G}(w^{(e)}_{a,b})$ | 1 | 2 | $\frac{1}{2b\log e}$ |
| $\{\mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}\}$ | $\rho_*$ (Def. II.20) | PLFM systems (Section IV-B) | $\mathcal{G}(w^{(p)}_{p,q})$ | 2 | 1 | $\frac{1}{p}$ |

TABLE II: Scaling behavior of the covering numbers (Definition II.12) for classes of nonlinear systems.

Finally, we remark that many of the results in this paper apply to LFM systems with general weight sequences $w[\cdot]$, specifically the bounds in Section II-C on the exterior covering number of $(\mathcal{G}(w), \rho_*)$ as well as the RNN constructions in Section V.

## APPENDIX A
## REPRESENTING A NEURAL NETWORK BY A BITSTRING.

**Definition A.1.** *Let $\Phi$ be a ReLU network with $(m, \epsilon)$-quantized weights. Denote the number of non-zero weights by $M := \mathcal{M}(\Phi)$. We organize the bitstring representation of $\Phi$ in $6$ segments as follows.*

*[Segment 1] The bitstring starts with $M$ 1's followed by a single 0.*

*[Segment 2] $L(\Phi)$ is specified in binary representation. As $L(\Phi) \leq M$, it suffices to allocate $\lceil \log M \rceil$ bits.*

*[Segment 3] $N_0, \ldots, N_L \leq M$ are specified in binary representation using a total of $(M+1)\lceil \log M \rceil$ bits.*

*[Segment 4] The topology of the network, i.e., the locations of the non-zero entries in the $A_\ell$ and $b_\ell$, $\ell \in \{1, \ldots, L\}$, is encoded as follows. We denote the bitstring corresponding to the binary representation of an integer $i \in \{1, \ldots, M\}$ by $b(i) \in \{0,1\}^{\lceil \log(M) \rceil}$. For $\ell \in \{1, \ldots, L\}$, $i \in \{1, \ldots, N_\ell\}$, $j \in \{1, \ldots, N_{\ell-1}\}$, a non-zero entry $(A_\ell)_{ij}$ is indicated by $[b(\ell), b(i), b(j)]$ and a non-zero entry $(b_\ell)_i$ by $[b(\ell), b(i), b(i)]$. Thus, encoding the topology of the network requires a total of $3\lceil \log M \rceil M$ bits.*

*[Segment 5] The quantity $m\lceil \log(\epsilon^{-1}) \rceil$ is represented by a bitstring of that many 1's followed by a single 0.*

*[Segment 6] The value of each non-zero weight and bias is represented by a bitstring of length $B_\epsilon = 2m\lceil \log \epsilon^{-1} \rceil + 1$.*

*The overall length of the bitstring is now given by*

$$\underbrace{M+1}_{\textit{Segment 1}} + \underbrace{\lceil \log M \rceil}_{\textit{Segment 2}} + \underbrace{(M+1)\lceil \log M \rceil}_{\textit{Segment 3}} + \underbrace{3\lceil \log(M) \rceil M}_{\textit{Segment 4}} + \underbrace{m\lceil \log(\epsilon^{-1}) \rceil + 1}_{\textit{Segment 5}} + \underbrace{MB_\epsilon}_{\textit{Segment 6}} . \tag{116}$$

*The ReLU network $\Phi$ can be recovered by sequentially reading out $M$,$L$, the $N_\ell$, the topology, the quantity $m\lceil \log(\epsilon^{-1}) \rceil$, and the quantized weights from the overall bitstring. It is not difficult to verify that the bitstring is crafted such that this yields unique decodability.*

## APPENDIX B
## COMPARISON WITH [7]

As mentioned in Section II-B, compared to [7], we use refined notions of massiveness of sets, as the system classes we consider are significantly more massive than the function classes dealt with in [7]. We next detail how the results from [7] fit into our framework.

In [7] the scaling behavior of covering numbers is quantified in terms of the optimal exponent $\gamma^*$ [7, Definition IV.1]. The following result relates $\gamma^*$ to the generalized dimension employed here (Definition II.12).

**Lemma B.1.** *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subset L^2(\Omega)$ be compact and such that the optimal exponent $\gamma^*(\mathcal{C})$ according to [7, Definition IV.1] is finite and non-zero. Then, $\mathcal{C}$ is, with respect to the metric $\rho(f,g) := \|f-g\|_{L^2(\Omega)}$, of order $\kappa = 1$, type $\tau = 1$, and generalized dimension*

$$\mathfrak{d} = \frac{1}{\gamma^*}.$$

*Proof.* For order $\kappa = 1$ and type $\lambda = 1$, the generalized dimension is given by

$$\mathfrak{d} = \limsup_{\epsilon \to 0} \frac{\log^{(2)} N^{\text{ext}}(\epsilon; \mathcal{C}, \rho)}{\log(\epsilon^{-1})}. \tag{117}$$

We note that by [34, Remark 5.10], it holds that

$$\gamma^* = \sup \left\{ \gamma > 0 : \log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \in \mathcal{O}\left(\epsilon^{-1/\gamma}\right), \epsilon \to 0 \right\}. \tag{118}$$

We first establish that $\mathfrak{d} \leq \frac{1}{\gamma^*}$. To this end, fix $\Delta > 0$ arbitrarily and observe that $\log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \in \mathcal{O}\big(\epsilon^{-1/(\gamma^* - \Delta)}\big)$. Hence, there exist $\epsilon_0, C > 0$ such that

$$\log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \leq C \epsilon^{-1/(\gamma^* - \Delta)}, \qquad \forall \epsilon \in (0, \epsilon_0),$$

and thus

$$\mathfrak{d} = \limsup_{\epsilon \to 0} \frac{\log^{(2)} N^{\text{ext}}(\epsilon; \mathcal{C}, \rho)}{\log(\epsilon^{-1})} \leq \limsup_{\epsilon \to 0} \left( \frac{1}{\gamma^* - \Delta} + \frac{\log(C)}{\log(\epsilon^{-1})} \right) = \frac{1}{\gamma^* - \Delta}. \tag{119}$$

As $\Delta > 0$ was arbitrary, we have established that $\mathfrak{d} \leq \frac{1}{\gamma^*}$.

Next, we show that $\mathfrak{d} \geq \frac{1}{\gamma^*}$. Again, fix $\Delta > 0$ arbitrarily. By (117), there exists an $\epsilon_0 > 0$ such that for all $\epsilon \in (0, \epsilon_0)$,

$$\frac{\log^{(2)} N^{\text{ext}}(\epsilon; \mathcal{C}, \rho)}{\log(\epsilon^{-1})} \leq \mathfrak{d} + \Delta. \tag{120}$$

This implies

$$\log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \leq \epsilon^{-(\mathfrak{d} + \Delta)}, \qquad \forall \epsilon \in (0, \epsilon_0), \tag{121}$$

and thus $\log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) \in \mathcal{O}\left(\epsilon^{-1/(\mathfrak{d}+\Delta)^{-1}}\right)$. Hence, $\gamma^* \geq (\mathfrak{d} + \Delta)^{-1}$. As $\Delta$ was arbitrary, this finalizes the proof. ☐

Table I in the present paper now follows from [7, Table I] by application of Lemma B.1. Furthermore, [7] shows through the transference principle [7, Section VII] that these function classes are *optimally representable by neural networks* [7, Definition VI.5]. The following Lemma hence allows us to conclude that every row in Table I is optimally representable (according to Definition II.14) by the canonical neural network decoder.

**Lemma B.2.** *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subset L^2(\Omega)$ be compact. If the function class $\mathcal{C} \subset L^2(\Omega)$ is* optimally representable by neural networks *according to [7, Definition VI.5], then $(\mathcal{C}, \rho)$ is optimally representable by the canonical neural network decoder with respect to the metric $\rho(f, g) := \|f - g\|_{L^2(\Omega)}$.*

*Proof.* First, we note that by Lemma B.1 $\mathcal{C}$ is of order $\kappa = 1$, type $\tau = 1$, and has generalized dimension $\mathfrak{d} = \frac{1}{\gamma^*(\mathcal{C})}$. By [7, Definition VI.5], we have $\gamma^*(\mathcal{C}) = \gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C})$. Next, fix $\Delta > 0$ arbitrarily. Now, following the proof of [7, Theorem VI.4, p. 2602] with $\gamma_{\mathcal{N}}^{*,\mathit{eff}} - \Delta$ in place of $\gamma$, we can conclude the existence of a polynomial $\pi^*$, a constant $C$, and a map $\Psi : (0, \frac{1}{2}) \times \mathcal{C} \to \mathcal{N}_{d,1}$ with the following properties. For every $f \in \mathcal{C}$ and every $\epsilon \in (0, \frac{1}{2})$, the network $\widetilde{\Phi}_{\epsilon,f} = \Psi(\epsilon, f)$ has $(\lceil \pi^*(\log(\epsilon^{-1})) \rceil, \epsilon)$-quantized weights and satisfies

$$\|f - \widetilde{\Phi}_{\epsilon,f}\|_{L^2(\Omega)} \leq \epsilon \quad \text{and} \quad \mathcal{M}(\widetilde{\Phi}_{\epsilon,f}) \leq C \epsilon^{-1/(\gamma_{\mathcal{N}}^{*,\mathit{eff}} - \Delta)} =: M_\epsilon.$$

Thus, $\widetilde{\Phi}_{\epsilon,f}$ can, according to Remark II.3, be reconstructed uniquely by the canonical neural network decoder $\mathcal{D}_{\mathcal{N}}$ from a bitstring of length no more than

$$C_0 \lceil \pi^*(\log(\epsilon^{-1})) \rceil \log(\epsilon^{-1}) M_\epsilon \log(M_\epsilon).$$

Therefore, $\mathcal{C}$ is representable by the canonical neural network decoder $\mathcal{D}_{\mathcal{N}}$ with $L(\epsilon; \mathcal{D}_{\mathcal{N}}, \mathcal{C}, \rho) \leq C_0 M_\epsilon \log(M_\epsilon) \log^{q_0}(\epsilon^{-1})$, where $q_0$ is a constant depending on $\pi^*$ only, and hence

$$\log L(\epsilon; \mathcal{D}_{\mathcal{N}}, \mathcal{C}, \rho) \leq \frac{1}{\gamma_{\mathcal{N}}^{*,\mathit{eff}} - \Delta} \log(\epsilon^{-1}) + o(\log(\epsilon^{-1})). \tag{122}$$

As $\Delta > 0$ was arbitrary, we thus get

$$\limsup_{\epsilon \to 0} \frac{\log L_{\mathcal{D}_{\mathcal{N}}}(\epsilon; \mathcal{C})}{\log(\epsilon^{-1})} \leq \frac{1}{\gamma_{\mathcal{N}}^{*,\mathit{eff}}}, \tag{123}$$

which, together with Lemma II.13 and the fact that $\mathfrak{d} = \frac{1}{\gamma^*} = \frac{1}{\gamma_{\mathcal{N}}^{*,\mathit{eff}}}$ implies that $\mathcal{C}$ is optimally representable by the canonical neural network decoder according to Definition II.14. □

# Appendix C
# Proofs

## A. Proof of Lemma II.13

To simplify notation, we set $L(\epsilon) := L(\epsilon; \mathcal{D}, \mathcal{C}, \rho)$. We first establish that

$$L(\epsilon) \geq \log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho) - 1, \quad \forall \epsilon > 0. \tag{124}$$

By way of contradiction, assume that

$$2^{L(\epsilon)+1} < N^{\text{ext}}(\epsilon; \mathcal{C}, \rho), \quad \text{for some } \epsilon > 0.$$

It then follows from Definition II.9 that for this $\epsilon$, for every $f \in \mathcal{C}$, there is an integer $\ell \leq L(\epsilon)$ and a bitstring $\mathbf{b}_f \in \{0,1\}^\ell$, such that $\rho(\mathcal{D}(\mathbf{b}_f), f) \leq \epsilon$. This directly implies that the set

$$\mathcal{U} := \left\{ \mathcal{D}(\mathbf{b}) \,\middle|\, \mathbf{b} \in \bigcup_{\ell=0}^{L(\epsilon)} \{0,1\}^\ell \right\},$$

is an $\epsilon$-net for $\mathcal{C}$. Furthermore,

$$|\mathcal{U}| \leq \sum_{\ell=0}^{L(\epsilon)} 2^\ell = \frac{2^{L(\epsilon)+1} - 1}{2 - 1} \leq 2^{L(\epsilon)+1} < N^{\text{ext}}(\epsilon; \mathcal{C}, \rho).$$

Hence, $\mathcal{U}$ is an $\epsilon$-net of cardinality strictly smaller than $N^{\text{ext}}(\epsilon; \mathcal{C}, \rho)$, which stands in contradiction to Definition II.11 and so (124) must hold. This, in turn, implies that

$$\begin{aligned}
\limsup_{\epsilon \to 0} \frac{\log^{(\kappa)} L(\epsilon)}{\log^\tau(\epsilon^{-1})} &\geq \limsup_{\epsilon \to 0} \frac{\log^{(\kappa)} (\log N^{\text{ext}}(\epsilon; \mathcal{C}, \rho_*) - 1)}{\log^\tau(\epsilon^{-1})} \\
&= \limsup_{\epsilon \to 0} \frac{\log^{(\kappa+1)} N^{\text{ext}}(\epsilon; \mathcal{C}, \rho_*)}{\log^\tau(\epsilon^{-1})} = \mathfrak{d},
\end{aligned}$$

as desired. □

### B. *Proof of Lemma II.21*

Fix $G, G' \in \mathcal{G}(w)$. First, note that

$$\rho_*(G,G') = \sup_{x\in\mathcal{S}^+}\sup_{t\in\mathbb{N}} |(Gx)[t]-(G'x)[t]| \\ \leq \sup_{x\in\mathcal{S}}\sup_{t\in\mathbb{Z}} |(Gx)[t]-(G'x)[t]|. \tag{125}$$

Next, arbitrarily fix a $\Delta > 0$. By definition of the supremum, it follows that there exist $x_0 \in \mathcal{S}$ and $e \in \mathbb{Z}$, such that

$$\sup_{x\in\mathcal{S}}\sup_{t\in\mathbb{Z}} |(Gx)[t]-(G'x)[t]| - \Delta/2 \leq \left|(Gx_0)[e]-(G'x_0)[e]\right|. \tag{126}$$

Since $\lim_{t\to\infty} w[t] = 0$, there exists $T > 0$ so that $w[t] \leq \Delta/(4D)$, for all $t > T$. Next, define

$$x_1[t] = x_0[t-(T-e)] \in \mathcal{S} \quad \text{and} \quad y[t] = x_1[t]\cdot \mathbb{1}_{\{t\geq 0\}} \in \mathcal{S}^+.$$

We then get

$$\sup_{\tau\geq 0} |w[\tau](x_1[T-\tau]-y[T-\tau])| = \sup_{\tau>T} |w[\tau]x_1[T-\tau]| \leq \sup_{\tau>T} |w[\tau]D| \leq \Delta/4, \tag{127}$$

where we used that $|x_1[\cdot]| \leq D$ by Definition II.15. We furthermore obtain

$$|(Gx_1)[T]-(G'x_1)[T]| \leq |(Gx_1)[T]-(Gy)[T]| + |(G'x_1)[T]-(G'y)[T]| \\ + |(Gy)[T]-(G'y)[T]| \tag{128}$$

$$\leq \Delta/4 + \Delta/4 + |(Gy)[T]-(G'y)[T]| \tag{129}$$

$$= \Delta/2 + |(Gy)[T]-(G'y)[T]|, \tag{130}$$

where in (128) we used the triangle inequality and in (129) we invoked the Lipschitz fading-memory property of $G$ and $G'$ in combination with (127). Next, note that by time-invariance of $G$ and $G'$, it holds that

$$\begin{aligned}(Gx_1)[T] &= (G\mathbf{T}_{(T-e)}x_0)[T] = (\mathbf{T}_{(T-e)}Gx_0)[T] = (Gx_0)[e], \\ (G'x_1)[T] &= (G'\mathbf{T}_{(T-e)}x_0)[T] = (\mathbf{T}_{(T-e)}G'x_0)[T] = (G'x_0)[e].\end{aligned} \tag{131}$$

Combining all these results yields

$$\sup_{x\in\mathcal{S}}\sup_{t\in\mathbb{Z}} |(Gx)[t]-(G'x)[t]| - \Delta/2 \overset{(126)}{\leq} \left|(Gx_0)[e]-(G'x_0)[e]\right| \overset{(131)}{=} \left|(Gx_1)[T]-(G'x_1)[T]\right| \\ \overset{(130)}{\leq} \Delta/2 + \left|(Gy)[T]-(G'y)[T]\right| \\ \overset{y\in\mathcal{S}^+}{\leq} \Delta/2 + \sup_{x\in\mathcal{S}^+}\sup_{t\in\mathbb{N}} |(Gx)[t]-(G'x)[t]| \\ \overset{\text{Def. II.20}}{=} \Delta/2 + \rho_*(G,G'). \tag{132}$$

As $\Delta > 0$ was arbitrary, we thus have established that

$$\sup_{x \in \mathcal{S}} \sup_{t \in \mathbb{Z}} |(Gx)[t] - (G'x)[t]| \leq \rho_*(G, G'),$$

which, together with (125), completes the proof. □

### *C. Proof of Lemma III.4*

Recall the projection operator $\mathcal{P} : \mathcal{S} \to \mathcal{S}^-$ defined according to

$$(\mathcal{P}x)[t] = x[t] \cdot \mathbb{1}_{\{t \leq 0\}}.$$

First, we need to show that $\mathcal{I} : \mathcal{G}_0(w) \to \mathcal{G}(w)$ given by

$$g \to \mathcal{I}(g), \quad \text{with } ((\mathcal{I}(g))(x))[t] = g(\mathcal{P}\mathbf{T}_{-t}x),$$

is a well-defined map from $\mathcal{G}_0(w)$ to $\mathcal{G}(w)$, i.e., we need to verify that for every $g \in \mathcal{G}_0(w)$, indeed $\mathcal{I}(g) \in \mathcal{G}(w)$. This will be done by showing that $\mathcal{I}(g)$ satisfies the conditions of Definition II.19. We first verify that $\mathcal{I}(g)$ is causal. Note, that for every $T \in \mathbb{Z}$, for every $x, x' \in \mathcal{S}$ with $x[t] = x'[t]$, $\forall t \leq T$, it holds that $\mathcal{P}\mathbf{T}_{-T}x = \mathcal{P}\mathbf{T}_{-T}x'$, and hence

$$((\mathcal{I}(g))(x))[T] = g(\mathcal{P}\mathbf{T}_{-T}x) = g(\mathcal{P}\mathbf{T}_{-T}x') = ((\mathcal{I}(g))(x'))[T].$$

Thus, $\mathcal{I}(g)$ is, indeed, causal according to Definition II.16. Next, we verify time-invariance as follows:

$$\begin{aligned}
(\mathbf{T}_\tau(\mathcal{I}(g)(x)))[t] &= ((\mathcal{I}(g))(x))[t - \tau] \\
&= g(\mathcal{P}\mathbf{T}_{\tau - t}x) \\
&= g(\mathcal{P}\mathbf{T}_{-t}\mathbf{T}_\tau x) \\
&= ((\mathcal{I}(g))(\mathbf{T}_\tau x))[t].
\end{aligned}$$

The Lipschitz fading-memory property of $\mathcal{I}(g)$ according to Definition II.18 is established by

$$\begin{aligned}
\big|((\mathcal{I}(g))(x))[t] - ((\mathcal{I}(g))(x'))[t]\big| &= |g(\mathcal{P}\mathbf{T}_{-t}x) - g(\mathcal{P}\mathbf{T}_{-t}x')| \\
&\leq \sup_{\tau \geq 0} |w[\tau]((\mathcal{P}\mathbf{T}_{-t}x)[-\tau] - (\mathcal{P}\mathbf{T}_{-t}x')[-\tau])| \qquad (133) \\
&= \sup_{\tau \geq 0} |w[\tau](x[t - \tau] - x'[t - \tau])|, \quad \forall t \in \mathbb{Z}, \forall x, x' \in \mathcal{S},
\end{aligned}$$

where in (133) we used (11). Finally, we observe that, by (11),

$$((\mathcal{I}(g))(0))[t] = g(\mathcal{P}\mathbf{T}_{-t}0) = g(0) = 0, \ \forall t \in \mathbb{Z}.$$

In summary, we have thus shown that $\mathcal{I}(g) \in \mathcal{G}(w)$ and hence $\mathcal{I}$ is, indeed, well-defined.

Furthermore, we have

$$\rho_*(\mathcal{I}(g),\mathcal{I}(g')) = \sup_{x\in\mathcal{S}}\sup_{t\in\mathbb{Z}}\left|((\mathcal{I}(g))(x))[t]-((\mathcal{I}(g'))(x))[t]\right| \tag{134}$$

$$= \sup_{x\in\mathcal{S}}\sup_{t\in\mathbb{Z}}|g(\mathcal{P}\mathbf{T}_{-t}x)-g'(\mathcal{P}\mathbf{T}_{-t}x)|$$

$$= \sup_{y\in\mathcal{S}}|g(\mathcal{P}y)-g'(\mathcal{P}y)| \tag{135}$$

$$= \sup_{z\in\mathcal{S}^-}|g(z)-g'(z)|$$

$$= \rho_0(g,g'), \qquad \forall g\in\mathcal{G}_0(w), \forall g'\in\mathcal{G}_0(w),$$

where in (134) we invoked Lemma II.21 and in (135) we used that $\mathcal{S}$ is complete under time shifts. This establishes that $\mathcal{I}$ is isometric and consequently injective.

Next, we prove that $\mathcal{I}$ is surjective. To this end, we fix $G\in\mathcal{G}(w)$ arbitrarily, consider

$$g(s) := (Gs)[0], \quad \forall s\in S_-, \tag{136}$$

and show that $g\in\mathcal{G}_0(w)$ as well as $\mathcal{I}(g)=G$. First, we establish that $g\in\mathcal{G}_0(w)$. The Lipschitz property of $g$ can be verified according to

$$|g(x)-g(x')| = |(Gx)[0]-(Gx')[0]|$$

$$\le \sup_{\tau\ge 0}|w[\tau](x[-\tau]-x'[-\tau])| \tag{137}$$

$$= \|x-x'\|_w, \quad \forall x,x'\in\mathcal{S}^-, \tag{138}$$

where in (137) we used the fact that $G\in\mathcal{G}(w)$ has Lipschitz fading memory (Definition II.18) and (138) is by (12). Furthermore, we have $g(0)=(G0)[0]=0$ and hence $g\in\mathcal{G}_0(w)$.

It remains to show that $\mathcal{I}(g)=G$. To this end, we fix $x\in\mathcal{S}$ and $t\in\mathbb{Z}$ arbitrarily and prove that

$$((\mathcal{I}(g))(x))[t] = (Gx)[t].$$

First, note that $(\mathcal{P}\mathbf{T}_{-t}x)[t'] = (\mathbf{T}_{-t}x)[t']$, for all $t'\le 0$. By causality of $G$, we conclude that $(G\mathcal{P}\mathbf{T}_{-t}x)[0]=(G\mathbf{T}_{-t}x)[0]$, which yields

$$((\mathcal{I}(g))(x))[t] = g(\mathcal{P}\mathbf{T}_{-t}x)$$

$$= (G\mathcal{P}\mathbf{T}_{-t}x)[0] \tag{139}$$

$$= (G\mathbf{T}_{-t}x)[0] = (\mathbf{T}_{-t}Gx)[0] \tag{140}$$

$$= (Gx)[t],$$

where in (139) we used (136), and in (140) we invoked the causality and the time-invariance of $G$. As $x\in\mathcal{S}$ and $t\in\mathbb{Z}$ were arbitrary, this proves that $\mathcal{I}(g)=G$. Furthermore, since $G\in\mathcal{G}(w)$ was arbitrary, we have established that $\mathcal{I}$ is surjective. Thus, $\mathcal{I}$ is, indeed, an isometric isomorphism between $(\mathcal{G}(w),\rho_*)$ and $(\mathcal{G}_0(w),\rho_0)$. Application of Lemma III.3 then yields $N(\epsilon;\mathcal{G}_0(w),\rho_0)=N(\epsilon;\mathcal{G}(w),\rho_*)$ and $M(\epsilon;\mathcal{G}_0(w),\rho_0)=M(\epsilon;\mathcal{G}(w),\rho_*)$. □

*D. Proof of Lemma III.5*

The proof relies on the following auxiliary result.

**Lemma C.1.** *Let $w[\cdot]$ be a weight sequence. The packing number of $\mathcal{S}^-$ w.r.t. $\|\cdot\|_w$ satisfies*

$$M(\epsilon; \mathcal{S}^-, \|\cdot\|_w) \geq \prod_{\ell=0}^{T} \left\lceil \frac{2Dw[\ell]}{\epsilon} \right\rceil,$$

*with $T := max\{T' \in \mathbb{N} \mid w[T'] > \frac{\epsilon}{2D}\}$.*

*Proof.* For $t \in [\![T]\!]$, we set $N_t := \left\lceil \frac{2Dw[t]}{\epsilon} \right\rceil \in \left[\frac{2Dw[t]}{\epsilon}, \frac{2Dw[t]}{\epsilon} + 1\right)$ and $\delta_t = \frac{2D}{N_t - 1} > \frac{\epsilon}{w[t]}$. Now, we define the set

$$\mathcal{U} := \{x_{i_0,\ldots,i_T} \mid i_t \in [\![N_t - 1]\!],\ \text{for } t \in [\![T]\!]\},\ \text{where} \tag{141}$$

$$x_{i_0,\ldots,i_T} := \begin{cases} -D + i_t\delta_t, & -T \leq t \leq 0 \\ 0, & \text{else} \end{cases}, \tag{142}$$

and show that $\mathcal{U}$ constitutes an $\epsilon$-packing for $(\mathcal{S}^-, \|\cdot\|_w)$. First, we establish that $\mathcal{U} \subset \mathcal{S}^-$ by verifying that $x_{i_0,\ldots,i_T}[t] \in [-D, D]$, for all $t \in \mathbb{Z}$, and $x_{i_0,\ldots,i_T}[t] = 0$, for all $t > 0$, holds for all $x_{i_0,\ldots,i_T} \in \mathcal{U}$. Indeed, for $t \in \{-T, \ldots, 0\}$, we have

$$-D \leq x_{i_0,\ldots,i_T}[t] = -D + i_t\delta_t \leq -D + (N_t - 1)\delta_t = -D + 2D = D,$$

and for $t \notin \{-T, \ldots, 0\}$,

$$x_{i_0,\ldots,i_T}[t] = 0.$$

Next, we show that for distinct $x_{i_0,\ldots,i_T}, x_{j_0,\ldots,j_T} \in \mathcal{U}$, i.e., there is at least one $\ell \in [\![T]\!]$ such that $i_\ell \neq j_\ell$, it holds that $\|x_{i_0,\ldots,i_T} - x_{j_0,\ldots,j_T}\|_w > \epsilon$. Indeed, for any such $\ell$, we have

$$\begin{aligned} \|x_{i_0,\ldots,i_T} - x_{j_0,\ldots,j_T}\|_w &= \sup_{t \geq 0} |w[t]\,(x_{i_0,\ldots,i_T}[-t] - x_{j_0,\ldots,j_T}[-t])| \\ &\geq |w[\ell]\,(x_{i_0,\ldots,i_T}[-\ell] - x_{j_0,\ldots,j_T}[-\ell])| \\ &= |w[\ell]\,(-D + i_\ell\delta_\ell + D - j_\ell\delta_\ell)| \\ &= (i_\ell - j_\ell)\delta_\ell w[\ell] > \epsilon, \end{aligned} \tag{143}$$

where (143) follows from $\delta_\ell > \frac{\epsilon}{w[\ell]}$. This establishes that $\mathcal{U}$, indeed, constitutes an $\epsilon$-packing for $(\mathcal{S}^-, \|\cdot\|_w)$, and we therefore have

$$M(\epsilon; \mathcal{S}^-, \|\cdot\|_w) \geq |\mathcal{U}| = \prod_{\ell=0}^{T} N_t = \prod_{\ell=0}^{T} \left\lceil \frac{2Dw[\ell]}{\epsilon} \right\rceil. \qquad \square$$

*Proof of Lemma III.5.* Let $M := \prod_{\ell=0}^{T} \left\lceil \frac{2Dw[\ell]}{\epsilon} \right\rceil$ and take $\{x_1, \ldots, x_M\}$ to be an $\epsilon$-packing for $(\mathcal{S}^-, \|\cdot\|_w)$ according to Lemma C.1. Hence, with $\delta := \min_{\ell \neq j} \|x_\ell - x_j\|_w > \epsilon$, there exists an $\epsilon' \in (\epsilon, \delta)$. Next, define balls of radius $\epsilon'/2$ centered at the packing points $\{x_1, \ldots, x_M\}$ according to $\mathcal{S}_\ell := \{x[\cdot] \in \mathcal{S}^- \mid \|x - x_\ell\|_w \leq \epsilon'/2\}$, for $\ell \in \{1, \ldots, M\}$. We now show that these balls are non-overlapping. Indeed, assuming that, for $j \neq \ell$, there exists an $x \in \mathcal{S}^-$ with $\|x - x_j\|_w < \epsilon'/2$ and $\|x - x_\ell\|_w < \epsilon'/2$, leads to the contradiction

$$\epsilon' < \delta \leq \|x_\ell - x_j\|_w = \|x_\ell - x + x - x_j\|_w \leq \|x_\ell - x\|_w + \|x - x_j\|_w \leq \epsilon'/2 + \epsilon'/2.$$

As the balls $\mathcal{S}_\ell$, $\ell \in \{1, \dots, M\}$, are non-overlapping, the signal $x[\cdot] = 0$ is contained in at most one ball $\mathcal{S}_\ell$ which we take to be $\mathcal{S}_M$ w.l.o.g.. Now, we define the set $\mathcal{U}$, with elements indexed by bitstring subscripts, according to

$$\mathcal{U} := \big\{ g_{\alpha_1,\dots,\alpha_{M-1}}(\cdot) \mid \alpha_\ell \in \{0,1\},\ \text{for } \ell \in \{1,\dots,M-1\} \big\},\ \text{where}$$
$$g_{\alpha_1,\dots,\alpha_{M-1}}(x) = \begin{cases} (2\alpha_\ell - 1)(\epsilon'/2 - \|x - x_\ell\|_w), & \text{for } x \in \mathcal{S}_\ell,\ \ell \in \{1,\dots,M-1\} \\ 0, & \text{else,} \end{cases}$$

and show that $\mathcal{U}$ constitutes an $\epsilon$-packing for $(\mathcal{G}_0(w), \rho_0)$. First, we establish that $\mathcal{U} \subset \mathcal{G}_0(w)$ by verifying that every $g_{\alpha_1,\dots,\alpha_{M-1}}(\cdot) \in \mathcal{U}$ satisfies the conditions in (11). Indeed, $g_{\alpha_1,\dots,\alpha_{M-1}}(0) = 0$ because the zero signal is either in no ball or in $\mathcal{S}_M$. Next, we show that $|g_{\alpha_1,\dots,\alpha_{M-1}}(x) - g_{\alpha_1,\dots,\alpha_{M-1}}(x')| \leq \|x - x'\|_w$, $\forall x, x' \in \mathcal{S}^-$. This will be done by distinguishing cases. First, assume that $x$ and $x'$ are contained in the same ball $\mathcal{S}_\ell$, for some $\ell \in \{1,\dots,M-1\}$. Then, we have

$$\begin{aligned} |g_{\alpha_1,\dots,\alpha_{M-1}}(x) - g_{\alpha_1,\dots,\alpha_{M-1}}(x')| &= |2\alpha_\ell - 1| \cdot \big| \|x' - x_\ell\|_w - \|x - x_\ell\|_w \big| \\ &\leq \|(x' - x_\ell) - (x - x_\ell)\|_w = \|x' - x\|_w, \end{aligned} \tag{144}$$

where we used the reverse triangle inequality. Next, assume that $x \in \mathcal{S}_\ell$ and $x' \in \mathcal{S}_j$, with $\ell, j \in \{1,\dots,M-1\}$ and $\ell \neq j$. We define $z(\mu) := x + \mu(x' - x),\ \mu \in [0,1]$. Since $z(0) = x \in \mathcal{S}_\ell$ and $z(1) = x' \notin \mathcal{S}_\ell$ (because the balls are non-overlapping), there must be a $\mu_1 \in (0,1)$ such that $\|z(\mu_1) - x_\ell\|_w = \epsilon'/2$. As $z(\mu_1) \notin \mathcal{S}_j$ and $z(1) = x' \in \mathcal{S}_j$, there must similarly be a $\mu_2 \in (\mu_1, 1)$ such that $\|z(\mu_2) - x_j\|_w = \epsilon'/2$. For notational simplicity, we set $z_1 := z(\mu_1)$, $z_2 := z(\mu_2)$, and $g := g_{\alpha_1,\dots,\alpha_{M-1}}$. Next, we bound

$$\begin{aligned} |g(x) - g(x')| &\leq |g(x) - g(z_1)| + |g(z_1) - g(z_2)| + |g(z_2) - g(x')| \\ &\leq |g(x) - g(z_1)| + 0 + |g(z_2) - g(x')| && (145) \\ &\leq \|x - z_1\|_w + \|z_2 - x'\|_w && (146) \\ &= \|\mu_1(x - x')\|_w + \|(1 - \mu_2)(x - x')\|_w && (147) \\ &= (1 + \mu_1 - \mu_2)\|x - x'\|_w < \|x - x'\|_w, && (148) \end{aligned}$$

where in (145) we used that $g(z_1) = g(z_2) = 0$, in (146) we applied (144) upon noting that $x, z_1$ and $z_2, x'$ are contained in the same ball each, in (147) we inserted the definition of $z_1$ and $z_2$, and in (148) we used $\mu_2 > \mu_1$. Finally, assume that $x \in \mathcal{S}_\ell$, for some $\ell \in \{1,\dots,M-1\}$, and $x' \notin \mathcal{S}_j$, for all $j \in \{1,\dots,M-1\}$. We let $z(\mu) := x + \mu(x' - x),\ \mu \in [0,1]$. Since $z(0) = x \in \mathcal{S}_\ell$ and $z(1) = x' \notin \mathcal{S}_\ell$, there must be a $\mu_1 \in (0,1)$ such that $\|z(\mu_1) - x_\ell\|_w = \epsilon'/2$. Again, we set $z_1 = z(\mu_1)$ and bound

$$
\begin{aligned}
|g(x) - g(x')| &\leq |g(x) - g(z_1)| + |g(z_1) - g(x')| \\
&\leq |g(x) - g(z_1)| + 0 && (149) \\
&\leq \|x - z_1\|_w = \|\mu_1(x - x')\|_w && (150) \\
&= \mu_1 \|x - x'\|_w < \|x - x'\|_w, && (151)
\end{aligned}
$$

where in (149) we used that $g(z_1) = g(x') = 0$, in (150) we applied (144) upon noting that $x, z_1$ are contained in the same ball, and in (151) we used $\mu_1 < 1$. We have thus established that $|g_{\alpha_1,\ldots,\alpha_{M-1}}(x) - g_{\alpha_1,\ldots,\alpha_{M-1}}(x')| \leq \|x - x'\|_w, \ \forall x, x' \in \mathcal{S}^-$, and hence $\mathcal{U} \subset \mathcal{G}_0(w)$.

Next, we show that for distinct $g_{\alpha_1,\ldots,\alpha_{M-1}}, g_{\beta_1,\ldots,\beta_{M-1}} \in \mathcal{U}$, i.e., there is at least one $\ell \in \{1, \ldots, M-1\}$ such that $\alpha_\ell \neq \beta_\ell$, it holds that $\rho_0(g_{\alpha_1,\ldots,\alpha_{M-1}}, g_{\beta_1,\ldots,\beta_{M-1}}) > \epsilon$. Indeed, for any such $\ell$, we have

$$
\begin{aligned}
\rho_0(g_{\alpha_1,\ldots,\alpha_{M-1}}, g_{\beta_1,\ldots,\beta_{M-1}}) &= \sup_{\widetilde{x} \in \mathcal{S}^-} |g_{\alpha_1,\ldots,\alpha_{M-1}}(\widetilde{x}) - g_{\beta_1,\ldots,\beta_{M-1}}(\widetilde{x})| && (152) \\
&\geq |g_{\alpha_1,\ldots,\alpha_{M-1}}(x_\ell) - g_{\beta_1,\ldots,\beta_{M-1}}(x_\ell)| && (153) \\
&= |2(\alpha_\ell - \beta_\ell)\epsilon'/2| = \epsilon' > \epsilon,
\end{aligned}
$$

where in (152) we used (13) and in (153) we inserted the particular choice $\widetilde{x} = x_\ell$ to lower-bound the sup. This establishes that $\mathcal{U}$, indeed, constitutes an $\epsilon$-packing for $(\mathcal{G}_0(w), \rho_0)$ and we therefore have

$$
\log M(\epsilon; \mathcal{G}_0(w), \rho_0) \geq \log |\mathcal{U}| = M - 1 = \left( \prod_{\ell=0}^{T} \left\lceil \frac{2Dw[\ell]}{\epsilon} \right\rceil \right) - 1. \qquad \square
$$

## *E. Proof of Lemma IV.2*

Note that $w_{a,b}^{(e)}[t] > \epsilon/d$ gives $t < \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)}$, and thereby

$$
T = \max\left\{ t \in \mathbb{N} \,\middle|\, t < \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\} = \left\lceil \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\rceil - 1. \tag{154}
$$

Now, we note that

$$
\begin{aligned}
&\log\left( \prod_{\ell=0}^{T} \frac{c w_{a,b}^{(e)}[\ell]}{\epsilon} \right) = \log\left( \prod_{\ell=0}^{T} \frac{ace^{-b\ell}}{\epsilon} \right) \\
&= (T+1)\log(\epsilon^{-1}) - b\log(e)\frac{T(T+1)}{2} + (T+1)\log(ac) \\
&= \left\lceil \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\rceil \log(\epsilon^{-1}) - b\log(e) \frac{\left\lceil \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\rceil \left( \left\lceil \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\rceil - 1 \right)}{2} \\
&\quad + \left\lceil \frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} \right\rceil \log(ac). && (155)
\end{aligned}
$$

Next, assuming that $\epsilon$ is sufficiently small to guarantee that $\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} - 1 \geq 0$, we can upper-bound (155) according to

$$\log\left(\prod_{\ell=0}^{T} \frac{cw_{a,b}^{(e)}[\ell]}{\epsilon}\right) \tag{156}$$

$$\begin{aligned}
&\leq \left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} + 1\right)\log(\epsilon^{-1}) - b\log(e)\frac{\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)}\left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} - 1\right)}{2} + \left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} + 1\right)\log(ac) \\
&= \frac{1}{2b\log(e)}\log^2(\epsilon^{-1}) + \left(\frac{\log(ad)}{b\log(e)} + 1 - \frac{1}{2b\log(e)}\left(2\log(ad) - b\log(e)\right) + \frac{\log(ac)}{b\log(e)}\right)\log(\epsilon^{-1}) \\
&\quad - \frac{\log(ad)\left(\log(ad) - b\log(e)\right)}{2b\log(e)} + \log(ac)\left(\frac{\log(ad)}{b\log(e)} + 1\right) \\
&= \frac{1}{2b\log(e)}\log^2(\epsilon^{-1}) + o\left(\log^2(\epsilon^{-1})\right).
\end{aligned} \tag{157}$$

In the same spirit, we can lower-bound (155) as

$$\log\left(\prod_{\ell=0}^{T} \frac{cw_{a,b}^{(e)}[\ell]}{\epsilon}\right) \tag{158}$$

$$\begin{aligned}
&\geq \left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)}\right)\log(\epsilon^{-1}) - b\log(e)\frac{\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)}\left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)} + 1\right)}{2} + \left(\frac{\log\left(\frac{ad}{\epsilon}\right)}{b\log(e)}\right)\log(ac) \\
&= \frac{1}{2b\log(e)}\log^2(\epsilon^{-1}) + \left(\frac{\log(ad)}{b\log(e)} - \frac{1}{2b\log(e)}\left(2\log(ad) + b\log(e)\right) + \frac{\log(ac)}{b\log(e)}\right)\log(\epsilon^{-1}) \\
&\quad - \frac{\log(ad)\left(\log(ad) + b\log(e)\right)}{2b\log(e)} + \log(ac)\left(\frac{\log(ad)}{b\log(e)}\right) \\
&= \frac{1}{2b\log(e)}\log^2(\epsilon^{-1}) + o\left(\log^2(\epsilon^{-1})\right).
\end{aligned} \tag{159}$$

Finally, combining (157) and (159) yields

$$\log\left(\prod_{\ell=0}^{T} \frac{cw_{a,b}^{(e)}[\ell]}{\epsilon}\right) = \frac{1}{2b\log(e)}\log^2(\epsilon^{-1}) + o\left(\log^2(\epsilon^{-1})\right),$$

as desired. □

### *F. Proof of Lemma IV.5*

We start by noting that

$$T = \max\left\{t \in \mathbb{N} \,\middle|\, w_{p,q}^{(p)} > \frac{\epsilon}{d}\right\} = \left\lceil \left(\frac{dq}{\epsilon}\right)^{1/p} \right\rceil - 2. \tag{160}$$

Next,

$$
\begin{aligned}
\log\left(\prod_{\ell=0}^{T}\frac{cw_{p,q}^{(p)}[\ell]}{\epsilon}\right) &= (T+1)\log c + (T+1)\log(\epsilon^{-1}) + \sum_{\ell=0}^{T}\log\left(w_{p,q}^{(p)}[\ell]\right) \\
&= (T+1)\log(cq) + (T+1)\log(\epsilon^{-1}) - p\log((T+1)!) \\
&= (T+1)\log(cq) + (T+1)\log(\epsilon^{-1}) - p\left((T+1)\log(T+1) - (T+1)\log(e)\right) \\
&\quad + O(\log(T+1)) \qquad (161) \\
&= (T+1)\log(cqe^{p}) + (T+1)\left(\log(\epsilon^{-1}) - p\log\left(\left\lceil\left(\frac{dq}{\epsilon}\right)^{1/p}\right\rceil - 1\right)\right) \\
&\quad + O(\log(\epsilon^{-1})), \qquad (162)
\end{aligned}
$$

where (161) follows from the logarithm form of Stirling's approximation, namely

$$
\log(n!) = n\log(n) - n\log(e) + O(\log(n)). \qquad (163)
$$

Now, assuming that $\epsilon$ is sufficiently small to guarantee that $1 \le \frac{1}{2}(dq/\epsilon)^{1/p}$, applying $\log\left((dq/\epsilon)^{1/p} - 1\right) \ge \log\left((\frac{1}{2}(dq/\epsilon)^{1/p})\right) \ge 0$, we can upper-bound (162) as

$$
\begin{aligned}
&(T+1)\log(cqe^{p}) + (T+1)\left(\log(\epsilon^{-1}) - p\log\left(\left(\frac{dq}{\epsilon}\right)^{1/p} - 1\right)\right) + O(\log(\epsilon^{-1})) \\
&\le (T+1)\log(cqe^{p}) + (T+1)\left(\log(\epsilon^{-1}) - p\log\left(\frac{1}{2}\left(\frac{dq}{\epsilon}\right)^{1/p}\right)\right) + O(\log(\epsilon^{-1})) \\
&= (T+1)(\log(ce^{p}/d) + p) + O(\log(\epsilon^{-1})) \\
&= \left(\left\lceil\left(\frac{dq}{\epsilon}\right)^{1/p}\right\rceil - 1\right)(\log(ce^{p}/d) + p) + O(\log(\epsilon^{-1})) \\
&\le \left(\frac{dq}{\epsilon}\right)^{1/p}(\log(ce^{p}/d) + p) + O(\log(\epsilon^{-1})). \qquad (164)
\end{aligned}
$$

Furthermore, we can lower-bound (162) according to

$$
\begin{aligned}
&(T+1)\log(cqe^{p}) + (T+1)\left(\log(\epsilon^{-1}) - p\log\left(\left(\frac{dq}{\epsilon}\right)^{1/p}\right)\right) + O(\log(\epsilon^{-1})) \\
&= (T+1)\log(ce^{p}/d) + O(\log(\epsilon^{-1})) \\
&= \left(\left\lceil\left(\frac{dq}{\epsilon}\right)^{1/p}\right\rceil - 1\right)\log(ce^{p}/d) + O(\log(\epsilon^{-1})) \\
&\ge \left(\left(\frac{dq}{\epsilon}\right)^{1/p} - 1\right)\log(ce^{p}/d) + O(\log(\epsilon^{-1})) \\
&= \left(\frac{dq}{\epsilon}\right)^{1/p}\log(ce^{p}/d) + O(\log(\epsilon^{-1})). \qquad (165)
\end{aligned}
$$

Combining (162), (164), and (165) yields the desired result

$$\log\left(\prod_{\ell=0}^{T}\frac{cw_{p,q}^{(p)}[\ell]}{\epsilon}\right)=\Theta(\epsilon^{-1/p}).$$

□

### G. Proof of Lemma IV.6

Consider $\epsilon\in(0,\epsilon_0)$ with $\epsilon_0=\frac{Dw_{p,q}^{(p)}[0]}{2}=\frac{aD}{2}$. By Theorem III.11, the exterior covering number of $(\mathcal{G}(w_{p,q}^{(p)}),\rho)$ satisfies

$$\left(\prod_{\ell=0}^{T'}\left\lceil\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right\rceil\right)-1\leq\log N^{\text{ext}}(\epsilon;\mathcal{G}(w_{p,q}^{(p)}),\rho_*)\leq\log(3)\prod_{\ell=0}^{T''}\left(2\left\lceil\frac{4Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right\rceil+1\right),\tag{166}$$

where

$$T':=\max\left\{\ell\in\mathbb{N}\;\middle|\;w_{p,q}^{(p)}[\ell]>\frac{\epsilon}{D}\right\}\ \text{and}\ T'':=\max\left\{\ell\in\mathbb{N}\;\middle|\;w_{p,q}^{(p)}[\ell]>\frac{\epsilon}{4D}\right\}.$$

We can further lower-bound the left-most term in (166) according to

$$\begin{aligned}\left(\prod_{\ell=0}^{T'}\left\lceil\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right\rceil\right)-1&\geq\left(\prod_{\ell=0}^{T'}\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right)-1\\&\geq\frac{1}{2}\prod_{\ell=0}^{T'}\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon},\end{aligned}\tag{167}$$

where (167) follows from

$$\frac{1}{2}\prod_{\ell=0}^{T'}\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon}\geq 1,$$

which, in turn, is a consequence of

$$\begin{aligned}&\frac{Dw_{p,q}^{(p)}[0]}{\epsilon}\geq\frac{Dw_{p,q}^{(p)}[0]}{\epsilon_0}=2,\\&\frac{Dw_{p,q}^{(p)}[\ell]}{\epsilon}\geq\frac{Dw_{p,q}^{(p)}[T']}{\epsilon}>1,\quad\text{for}\quad\ell\in[\![T']\!]\setminus\{0\}.\end{aligned}$$

Similarly, we can further upper-bound the right-most term in (166) as

$$\begin{aligned}\log(3)\left(\prod_{\ell=0}^{T''}\left(2\left\lceil\frac{4Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right\rceil+1\right)\right)&\leq\log(3)\left(\prod_{\ell=0}^{T''}\left(\frac{8Dw_{p,q}^{(p)}[\ell]}{\epsilon}+3\right)\right)\\&\leq\log(3)\left(\prod_{\ell=0}^{T''}\frac{20Dw_{p,q}^{(p)}[\ell]}{\epsilon}\right),\end{aligned}\tag{168}$$

where (168) follows from

$$\frac{4Dw_{p,q}^{(p)}[\ell]}{\epsilon}\geq\frac{4Dw_{p,q}^{(p)}[T'']}{\epsilon}>1,\quad\text{for}\quad\ell\in[\![T'']\!].$$

Combining (166)–(168), taking logarithms two more times, dividing the results by $\log(\epsilon^{-1})$, and applying Lemma IV.5, we obtain

$$\frac{\log\left(\Theta(\epsilon^{-1/p})\right)}{\log(\epsilon^{-1})} \leq \frac{\log^{(3)} N^{\text{ext}}(\epsilon; \mathcal{G}(w_{p,q}^{(p)}), \rho_*)}{\log(\epsilon^{-1})} \leq \frac{\log\left(\Theta(\epsilon^{-1/p})\right)}{\log(\epsilon^{-1})}. \tag{169}$$

Taking the limit $\epsilon \to 0$, we finally get

$$\lim_{\epsilon \to 0} \frac{\log^{(3)} N^{\text{ext}}(\epsilon; \mathcal{G}(w_{p,q}^{(p)}), \rho_*)}{\log(\epsilon^{-1})} = \frac{1}{p},$$

which implies that $(\mathcal{G}(w_{p,q}^{(p)}), \rho_*)$ is of order 2 and type 1, with generalized dimension

$$\mathfrak{d} = \frac{1}{p}.$$

□

*H. Auxiliary results on ReLU networks*

**Lemma C.2.** *For $d \in \mathbb{N}$, consider the following functions,*

$$f_d^{min}(z) := \min\{z_1, z_2, \dots, z_d\}, \qquad \textit{for } z \in \mathbb{R}^d,$$
$$f_d^{max}(z) := \max\{z_1, z_2, \dots, z_d\}, \qquad \textit{for } z \in \mathbb{R}^d.$$

*Then, there exist ReLU networks $\Phi_d^{min} \in \mathcal{N}_{d,1}$ and $\Phi_d^{max} \in \mathcal{N}_{d,1}$, with $\mathcal{L}(\Phi_d^{min}) = \mathcal{L}(\Phi_d^{max}) = \lceil \log(d+1) \rceil + 1$, $\mathcal{M}(\Phi_d^{min}) = \mathcal{M}(\Phi_d^{max}) \leq 14d - 9$, $\mathcal{W}(\Phi_d^{min}) = \mathcal{W}(\Phi_d^{max}) \leq 3d$, $\mathcal{K}(\Phi_d^{min}) = \mathcal{K}(\Phi_d^{max}) = \{1, -1\}$, $\mathcal{B}(\Phi_d^{min}) = \mathcal{B}(\Phi_d^{max}) = 1$, such that*

$$\Phi_d^{min}(z) = f_d^{min}(z), \qquad \textit{for all} \quad z \in \mathbb{R}^d,$$
$$\Phi_d^{max}(z) = f_d^{max}(z), \qquad \textit{for all} \quad z \in \mathbb{R}^d.$$

*Proof.* We only need to show the result for the max function as $\min\{z_1, z_2, \dots, z_d\} = -\max\{-z_1, -z_2, \dots, -z_d\}$. First, we realize $\max\{x_1, x_2\}$ according to

$$\max\{x_1, x_2\} = x_1 + \rho(x_2 - x_1) \tag{170}$$

by the ReLU network

$$\widehat{\Phi} = \widehat{W}_2 \circ \rho \circ \widehat{W}_1, \tag{171}$$

with

$$\begin{aligned} \widehat{W}_1(x) &= \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ -1 & 1 \end{pmatrix} x =: A_1 x, \\ \widehat{W}_2(x) &= \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} x =: A_2 x. \end{aligned} \tag{172}$$

Now, for arbitrary $d \in \mathbb{N}$, choose $\ell \in \mathbb{N}$ such that $2^\ell \leq d < 2^{\ell+1}$. Then, we double the first $k = 2^{\ell+1} - d$

elements and retain the remaining $d-k$ elements as follows

$$\max\{x_1,\ldots,x_d\} = \max\{x_1,x_1,\ldots,x_k,x_k,x_{k+1},x_{k+2},\ldots,x_d\}. \tag{173}$$

The primary reason for doubling elements is that by extending the set $\{x_1,\ldots,x_d\}$ to a set whose cardinality is a power of 2, we can utilize (170) together with a divide-and-conquer approach to determine the maximum value of the set $\{x_1,\ldots,x_d\}$. This then leads to a ReLU network realization of depth scaling logarithmically in $d$. Now, set $W_{\ell+1}(x) := A_2x$, $W_j(x) := B_jx$, for $j \in \{\ell,\ell-1,\ldots,1\}$, and $W_0(x) := A_0x$, with

$$\begin{aligned}
B_o &= A_1 \operatorname{diag}(A_2, A_2), \\
B_j &= \operatorname{diag}(\underbrace{B_o,\ldots,B_o}_{2^{\ell-j}})x, \\
&= \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix}, \\
A_0 &= \operatorname{diag}(\underbrace{A_1,\ldots,A_1}_{2^\ell}) \operatorname{diag}(\underbrace{\mathbb{1}_2,\ldots,\mathbb{1}_2}_{k}, \mathbb{I}_{d-k}) \\
&= \operatorname{diag}(\underbrace{A_1\mathbb{1}_2,\ldots,A_1\mathbb{1}_2}_{k}, \underbrace{A_1,\ldots,A_1}_{2^\ell-k}), \\
A_1\mathbb{1}_2 &= (1,-1,0)^T.
\end{aligned} \tag{174}$$

Combining (170)-(174), we will now be able to establish

$$\Phi_d^{max} := W_{\ell+1} \circ \rho \circ W_\ell \circ \rho \circ \ldots \rho \circ W_1 \circ \rho \circ W_0 = f_d^{\max}. \tag{175}$$

The proof of (175) is summarized as follows:

(i) We double the first $k = 2^{\ell+1}-d$ elements and retain the remaining $d-k$ elements of $\{x_1,x_2,\ldots,x_d\}$ according to

$$\begin{aligned}
&\begin{pmatrix} x_1 & x_1 & \ldots & x_k & x_k & x_{k+1} & x_{k+2} & \ldots & x_d \end{pmatrix}^T \\
&= \operatorname{diag}(\underbrace{\mathbb{1}_2,\ldots,\mathbb{1}_2}_{k}, \mathbb{I}_{d-k}) \begin{pmatrix} x_1 & x_2 & \ldots & x_d \end{pmatrix}^T.
\end{aligned}$$

For simplicity, we denote the resulting set $\{x_1,x_1,\ldots,x_k,x_k,x_{k+1},x_{k+2},\ldots,x_d\}$ as $\{y_1,y_2,\ldots,y_{2^{\ell+1}}\}$.

(ii) Application of (171) to the $2^\ell$ pairs in $\{y_1,y_2,\ldots,y_{2^{\ell+1}}\}$, results in $\{\max\{y_1,y_2\},\max\{y_3,y_4\},\ldots,\max\{y_{2^{\ell+1}-1},y_{2^{\ell+1}}\}\}$ and hence reduces the number of elements from $2^{\ell+1}$ to $2^\ell$. This reduction is applied iteratively until we get $\max\{x_1,x_2,\ldots,x_d\}$. The compositions $A_1 \circ \operatorname{diag}(A_2,A_2)$, formed in the iterative application of (171), constitute the main diagonal elements of the matrices $B_k$, $k \in \{\ell,\ell-1,\ldots,1\}$. We will illustrate this part of the procedure using the simplest possible example, namely for $\ell = 1$ and hence $\max\{y_1,y_2,y_3,y_4\}$.

Specifically, we have

$$\begin{aligned}\max\{y_1,y_2,y_3,y_4\} &= \max\{\max\{y_1,y_2\},\max\{y_3,y_4\}\}\\ &= \max\left\{\operatorname{diag}(A_2,A_2)\,\rho\left(\operatorname{diag}(A_1,A_1)\left((y_1,y_2),(y_3,y_4)\right)^T\right)\right\}\\ &= A_2\rho\left(A_1\operatorname{diag}(A_2,A_2)\,\rho\left(\operatorname{diag}(A_1,A_1)\left((y_1,y_2),(y_3,y_4)\right)^T\right)\right)\\ &= (W_2\circ\rho\circ W_1\circ\rho\circ W_0)(y).\end{aligned}$$

Finally, we can determine the size of the resulting network (175) as follows

$$\begin{aligned}\mathcal{L}(\Phi_d^{\max}) &\overset{(175)}{=} \ell+1 = \lceil\log(d+1)\rceil+1,\\ \mathcal{M}(\Phi_d^{\max}) &\overset{(175)}{=} \sum_{j=0}^{\ell+1}\mathcal{M}(W_j) = \mathcal{M}(A_2)+\sum_{j=1}^{\ell}\mathcal{M}(B_j)+\mathcal{M}(A_0)\\ &\overset{(174),(172)}{=} 3+2k+4(2^\ell-k)+12\sum_{j=1}^{\ell}2^{\ell-j}\le 14d-9,\\ \mathcal{W}(\Phi_d^{\max}) &\overset{(175)}{=} \max_{j=0,1\dots,\ell+1}\mathcal{W}(W_j)\overset{(174),(172)}{=} 3\cdot 2^\ell\le 3d,\\ \mathcal{K}(\Phi_d^{\max}) &\overset{(175)}{\subset} \bigcup_{j=0}^{\ell+1}\mathcal{K}(W_j)\overset{(174),(172)}{=}\{1,-1\},\\ \mathcal{B}(\Phi_d^{\max}) &= \max_{b\in\mathcal{K}(\Phi_d^{\max})}|b| = 1.\end{aligned}$$

□

**Lemma C.3** (Composition of ReLU networks [7])**.** *For $i=1,\dots,n$, let $d_i\in\mathbb{N}$ and $\Phi_i\in\mathcal{N}_{d_i,d_{i+1}}$. Then, there exists a network $\Psi\in\mathcal{N}_{d_1,d_{n+1}}$ with*

$$\Psi(x) = (\Phi_n\circ\Phi_{n-1}\circ\dots\circ\Phi_1)(x),\quad \textit{for all } x\in\mathbb{R}^{d_1}, \tag{176}$$

*satisfying*

$$\begin{aligned}\mathcal{L}(\Psi) &= \sum_{i=1}^{n}\mathcal{L}(\Phi_i),\\ \mathcal{M}(\Psi) &\le 2\sum_{i=1}^{n}\mathcal{M}(\Phi_i),\\ \mathcal{W}(\Psi) &\le \max\left\{\max_{i=1,\dots,n-1}\{2d_i\},\max_{i=1,\dots,n}\{\mathcal{W}(\Phi_i)\}\right\},\\ \mathcal{K}(\Psi) &\subset \bigcup_{i=1}^{n}(\mathcal{K}(\Phi_i)\cup(-\mathcal{K}(\Phi_i))),\\ \mathcal{B}(\Psi) &= \max_{i=1,\dots,n}\mathcal{B}(\Phi_i).\end{aligned} \tag{177}$$

*Proof.* Follows along the same lines as the proof of [7, Lemma 2.3]. □

**Lemma C.4** (Parallelization of ReLU networks of the same depth [7])**.** *For $i=1,\dots,n$, let $d_i,d_i'\in\mathbb{N}$*

*and* $\Phi_i \in \mathcal{N}_{d_i,d_i'}$ *with* $\mathcal{L}(\Phi_i) = L$*. Then, there exists a ReLU network* $P(\Phi_1, \Phi_2, \ldots, \Phi_n) \in \mathcal{N}_{\sum_{i=1}^n d_i, \sum_{i=1}^n d_i'}$ *with*

$$P(\Phi_1, \Phi_2, \ldots, \Phi_n)(x) = (\Phi_1(x), \Phi_2(x), \ldots, \Phi_n(x))^T, \quad \textit{for all } x \in \mathbb{R}^{\sum_{i=1}^n d_i}, \tag{178}$$

*satisfying*

$$\begin{aligned}
\mathcal{L}(P(\Phi_1, \Phi_2, \ldots, \Phi_n)) &= L, \\
\mathcal{M}(P(\Phi_1, \Phi_2, \ldots, \Phi_n)) &= \sum_{i=1}^n \mathcal{M}(\Phi_i), \\
\mathcal{W}(P(\Phi_1, \Phi_2, \ldots, \Phi_n)) &\leq \sum_{i=1}^n \mathcal{W}(\Phi_i), \\
\mathcal{K}(P(\Phi_1, \Phi_2, \ldots, \Phi_n)) &= \bigcup_{i=1}^n \mathcal{K}(\Phi_i), \\
\mathcal{B}(P(\Phi_1, \Phi_2, \ldots, \Phi_n)) &= \max_{i=1,\ldots,n} \mathcal{B}(\Phi_i).
\end{aligned} \tag{179}$$

*Proof.* Follows along similar lines as the proof of [7, Lemma 2.5]. □

**Lemma C.5** (Augmenting network depth [7])**.** *Let* $d_1, d_2, K \in \mathbb{N}$*, and* $\Phi \in \mathcal{N}_{d_1,d_2}$ *with* $\mathcal{L}(\Phi) < K$*. Then, there exists a network* $\Psi \in \mathcal{N}_{d_1,d_2}$ *with*

$$\begin{aligned}
\mathcal{L}(\Psi) &= K, \\
\mathcal{M}(\Psi) &\leq \mathcal{M}(\Phi) + d_2\mathcal{W}(\Phi) + 2d_2(K - \mathcal{L}(\Phi)), \\
\mathcal{W}(\Psi) &= \max\{2d_2, \mathcal{W}(\Phi)\}, \\
\mathcal{K}(\Psi) &\subset (\mathcal{K}(\Phi)) \cup (-\mathcal{K}(\Phi)) \cup \{1, -1\}, \\
\mathcal{B}(\Psi) &= \max\{1, \mathcal{B}(\Phi)\},
\end{aligned} \tag{180}$$

*satisfying* $\Psi(x) = \Phi(x)$*, for all* $x \in \mathbb{R}^{d_1}$*.*

*Proof.* Follows along the same lines as the proof of [7, Lemma 2.4]. □

### *I. Remainder of the proof of Lemma V.2*

We verify that $\Phi = \Phi_2 \circ \Phi_1$ defined in (77) has $(2, \epsilon)$-quantized weights.

- The weights in $\Phi_2 = \widetilde{W}^\Sigma$ defined in (74):

  $$\widetilde{g}_n \overset{(63)}{=} \mathcal{Q}_{2,\epsilon}(g(\widetilde{x}_n)) \overset{\text{Definition II.2}}{=} \left\lceil g(\widetilde{x}_n)/2^{-2\lceil\log(\epsilon^{-1})\rceil} \right\rceil \cdot 2^{-2\lceil\log(\epsilon^{-1})\rceil} \in 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z},$$

  $$|\widetilde{g}_n| \leq |g(\widetilde{x}_n)| + 2^{-2\lceil\log(\epsilon^{-1})\rceil} \overset{(11)}{\leq} D + 2^{-2\lceil\log(\epsilon^{-1})\rceil} \overset{(189)}{\leq} D + \epsilon^2 \leq D + 1 \overset{(189)}{\leq} \epsilon^{-2},$$

  which yields

  $$\mathcal{K}(\Phi_2) \subset 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \tag{181}$$

- The weights in $\Phi_1$ defined in (76):

$$\mathcal{K}(\Phi_1) = \bigcup_{n\in\widetilde{\mathfrak{N}}} \left(\mathcal{K}(\Phi_{1,2}^n) \cup \mathcal{K}(\Phi_{1,1}^n)\right) = \mathcal{K}(\Psi) \cup \left(\bigcup_{n\in\widetilde{\mathfrak{N}}} \mathcal{K}(\widehat{W}_n)\right). \tag{182}$$

Recalling that $\Psi$ realizes the spike function and applying Lemma V.1, we obtain

$$\mathcal{K}(\Psi) = \{1, -1\} \subset 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \tag{183}$$

Based on (74), we get

$$\bigcup_{n\in\widetilde{\mathfrak{N}}} \mathcal{K}(\widehat{W}_n) = \bigcup_{\ell=0}^{T} \left(\bigcup_{n\in\widetilde{\mathfrak{N}}} \{\widetilde{\delta}_\ell^{-1}, n_\ell\}\right). \tag{184}$$

Moreover, for all $\ell = 0, \dots, T$ and $n \in \widetilde{\mathfrak{N}}$, we have

$$\widetilde{\delta}_\ell^{-1} \stackrel{(61)}{=} \mathcal{Q}_{2,\epsilon}(\delta_\ell^{-1}) \stackrel{\text{Definition II.2}}{=} \left\lceil \frac{s+1}{s}\frac{w[\ell]}{\epsilon} \Big/ 2^{-2\lceil\log(\epsilon^{-1})\rceil} \right\rceil \cdot 2^{-2\lceil\log(\epsilon^{-1})\rceil} \in 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z},$$

$$\left|\widetilde{\delta}_\ell^{-1}\right| \leq \frac{s+1}{s}\frac{w[\ell]}{\epsilon} + 2^{-2\lceil\log(\epsilon^{-1})\rceil} \stackrel{\text{Definition II.18}}{\leq} \frac{s+1}{s}\epsilon^{-1} + \epsilon^2 \stackrel{(189)}{\leq} \epsilon^{-2}.$$

Hence,

$$\widetilde{\delta}_\ell^{-1} \in 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right], \tag{185}$$

and

$$\begin{aligned} n_\ell &\in \mathbb{Z} \subset 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z}, \\ |n_\ell| &\leq \widetilde{N}_\ell \stackrel{(62)}{=} \left\lceil \frac{D}{\widetilde{\delta}_\ell} \right\rceil \leq D\left|\widetilde{\delta}_\ell^{-1}\right| + 1 \\ &\leq D\left(\frac{s+1}{s}\epsilon^{-1} + \epsilon^2\right) + 1 \\ &\stackrel{(189)}{\leq} \left(D\frac{s+1}{s} + D + 1\right)\epsilon^{-1} \stackrel{(189)}{\leq} \epsilon^{-2}, \end{aligned}$$

which yields

$$n_\ell \in 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \tag{186}$$

Based on (184), (185), and (186), we have

$$\bigcup_{n\in\widetilde{\mathfrak{N}}} \mathcal{K}(\widehat{W}_n) \subset 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \tag{187}$$

Combining (182), (183), (184), and (187) yields

$$\mathcal{K}(\Phi_1) \subset 2^{-2\lceil\log(\epsilon^{-1})\rceil}\mathbb{Z} \cap \left[-\epsilon^{-2}, \epsilon^{-2}\right]. \tag{188}$$

Using (181) and (188), Lemma C.3 shows that $\Phi = \Phi_2 \circ \Phi_1$, indeed, has $(2, \epsilon)$-quantized weights.

Finally, we compute $\mathcal{L}(\Phi)$ and derive an upper bound on $\mathcal{M}(\Phi)$ according to

$$\begin{aligned}
\mathcal{L}(\Phi) &\overset{(77),\text{Lemma}C.3}{=} \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_1) \\
&\overset{(76),(75),\text{Lemma}C.4}{=} \mathcal{L}(\Psi) + 2 \\
&\overset{\text{Lemma}V.1}{=} \lceil \log(d+1) \rceil + 6, \\
\mathcal{M}(\Phi) &\overset{\text{Lemma C.3}}{\leq} 2\left(\mathcal{M}(\Phi_2) + \mathcal{M}(\Phi_1)\right) \\
&\overset{\text{Lemma C.4}}{=} 2\left(\mathcal{M}(\Phi_2) + \sum_{i=1}^{|\widetilde{\mathfrak{N}}|}\left(2\mathcal{M}(\Psi) + 2\mathcal{M}(\widehat{W}_{n^i})\right)\right) \\
&\overset{(74)\text{-}(77),\text{ Lemma V.1}}{\leq} 2|\widetilde{\mathfrak{N}}|\left(1 + 120(T+1) - 56 + 2(T+1)\right) \\
&\leq 244(T+1)|\widetilde{\mathfrak{N}}| \\
&\overset{(62)}{\leq} 244(T+1)\prod_{\ell=0}^{T}\left(2D\widetilde{\delta}_\ell^{-1} + 3\right) \\
&\overset{(59),(61),(62)}{\leq} 244(T+1)\prod_{\ell=0}^{T}\left(2D\left(\frac{s+1}{s}\frac{w[\ell]}{\epsilon} + \epsilon^2\right) + 3\right) \\
&\overset{(189)}{\leq} 244(T+1)\prod_{\ell=0}^{T}\left(\frac{2Dw[\ell]}{\epsilon}\frac{s+1}{s} + 4\right).
\end{aligned}$$

The proof is concluded by setting

$$\epsilon_0 = \min\left\{1, \frac{1}{2}, \frac{1}{s+1}, \sqrt{\frac{1}{D+1}}, \left(D\frac{s+1}{s} + D + 1\right)^{-1}\right\}. \tag{189}$$

□

## Acknowledgment

The authors would like to thank Erwin Riegler for useful suggestions regarding the proof of Lemma C.2.

## References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[2] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989.

[3] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[4] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, p. 930–945, May 1993.

[5] ——, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, no. 1, p. 115–133, Jan. 1994.

[6] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 8–45, Jan. 2019.

[7] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei, "Deep neural network approximation theory," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2581–2623, May 2021.

[8] S. Nagel and Z. Xu, "Asset pricing with fading memory," *The Review of Financial Studies*, vol. 35, no. 5, pp. 2190–2245, 2022.

[9] E. H. Dill, "Simple materials with fading memory," *Academic Press, New York*, 1975.

[10] C. C. Wang, "The principle of fading memory," *Archive for Rational Mechanics and Analysis*, vol. 18, pp. 343–366, 1965.

[11] W. A. Day, *The Thermodynamics of Simple Materials with Fading Memory*. Springer Science & Business Media, 2013, vol. 22.

[12] A. Bakshi, A. Liu, A. Moitra, and M. Yau, "A new approach to learning linear dynamical systems," in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023, pp. 335–348.

[13] B. Boots, G. J. Gordon, and S. Siddiqi, "A constraint generation approach to learning stable linear dynamical systems," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[14] E. Hazan, K. Singh, and C. Zhang, "Learning linear dynamical systems via spectral filtering," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conference on Learning Theory*. PMLR, 2020, pp. 3487–3582.

[16] C. Hutter, R. Gül, and H. Bölcskei, "Metric entropy limits on recurrent neural network learning of linear dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 59, pp. 198–223, 2022.

[17] C. L. Giles, B. G. Horne, and T. Lin, "Learning a class of large finite state machines with a recurrent neural network," *Neural Networks*, vol. 8, no. 9, pp. 1359–1365, 1995.

[18] I. W. Sandberg, "Approximately finite memory and the circle criterion," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 41, no. 7, pp. 473–476, 1994.

[19] V. Volterra, *Theory of Functionals and of Integral and Integro-Differential Equations*, ser. Dover Books on Mathematics. New York: Dover Publications, 1959.

[20] N. Wiener, *Nonlinear problems in random theory*, 1966.

[21] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 11, pp. 1150–1161, 1985.

[22] M. B. Matthews, "Approximating nonlinear fading-memory operators using neural network models," *Circuits, Systems and Signal Processing*, vol. 12, no. 2, pp. 279–307, 1993.

[23] L. Grigoryeva and J.-P. Ortega, "Echo state networks are universal," *Neural Networks*, vol. 108, pp. 495–508, 2018.

[24] V. Tikhomirov, "$\varepsilon$-entropy and $\varepsilon$-capacity of sets in functional spaces," in *Selected works of A.N. Kolmogorov*. Springer, 1993, pp. 86–170.

[25] M. J. Wainwright, *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019, vol. 48.

[26] G. Zames, "On the metric complexity of causal linear systems: $\epsilon$-entropy and $\epsilon$-dimension for continuous time," *IEEE Transactions on Automatic Control*, vol. 24, no. 2, pp. 222–230, 1979.

[27] V. Lakshmikantham, *Theory of Integro-differential Equations*. CRC press, 1995, vol. 1.

[28] G. Alves, J. De Araújo, J. Cressoni, L. da Silva, M. da Silva, and G. Viswanathan, "Superdiffusion driven by exponentially decaying memory," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 4, p. P04026, 2014.

[29] T. R. Moura, G. Viswanathan, M. da Silva, J. Cressoni, and L. da Silva, "Transient superdiffusion in random walks with a q-exponentially decaying memory profile," *Physica A: Statistical Mechanics and its Applications*, vol. 453, pp. 259–263, 2016.

[30] S. Karnik, R. Wang, and M. Iwen, "Neural network approximation of continuous functions in high dimensions with applications to inverse problems," *Journal of Computational and Applied Mathematics*, vol. 438, p. 115557, 2024.

[31] D. Yarotsky, "Optimal approximation of continuous functions by very deep ReLU networks," in *Conference on Learning Theory*. PMLR, 2018, pp. 639–649.

[32] E. Mainini and G. Mola, "Exponential and polynomial decay for first order linear Volterra evolution equations," *Quarterly of Applied Mathematics*, vol. 67, no. 1, pp. 93–111, 2009.

[33] M. Fabrizio and S. Polidoro, "Asymptotic decay for some differential systems with fading memory," *Applicable Analysis*, vol. 81, no. 6, pp. 1245–1264, 2002.

[34] P. Grohs, *Optimally sparse data representations*. Birkhäuser, Cham, 2015.