

Stable learning using spiking neural networks equipped with affine encoders and decoders

A. Martina Neuman

Dominik Dold

Philipp Christian Petersen

*University of Vienna, Faculty of Mathematics
Vienna, Austria*

ANH.MARTINA.NEUMAN@UNIVIE.AC.AT

DOMINIK.DOLD@UNIVIE.AC.AT

PHILIPP.PETERSEN@UNIVIE.AC.AT

Editor: Christophe Giraud

Abstract

We study the learning problem associated with spiking neural networks. Specifically, we focus on spiking neural networks composed of *simple spiking neurons* having only positive synaptic weights, equipped with an affine encoder and decoder; we refer to these as *affine spiking neural networks*. These neural networks are shown to depend continuously on their parameters, which facilitates classical covering number-based generalization statements and supports stable gradient-based training. We demonstrate that the positivity of the weights enables a wide range of expressivity results, including rate-optimal approximation of smooth functions and dimension-independent approximation of Barron regular functions. In particular, we show in theory and simulations that affine spiking neural networks are capable of approximating shallow ReLU neural networks. Furthermore, we apply these affine spiking neural networks to standard machine learning benchmarks and reach competitive results (with respect to deep feedforward networks). Finally, we observe that from a generalization perspective, contrary to feedforward neural networks or previous results for general spiking neural networks, the depth has little to no adverse effect on theoretical guarantees for the generalization capabilities.

1. Introduction

Deep learning (LeCun et al., 2015; Bengio et al., 2017) is a technology that has revolutionized many areas of modern life. At its core, the term describes the gradient-based training of deep neural networks. Since its breakthrough in image classification in 2012 (Krizhevsky et al., 2012), deep learning is essentially the only viable technology for this application. Moreover, it is the basis of multiple recent breakthroughs in science (Jumper et al., 2021) and even mathematical research (Davies et al., 2021). Recently, deep learning has received wide public attention through the advent of generative AI in the form of large language models such as ChatGPT (OpenAI, 2023). However, it is well-documented that deep learning in modern applications often demands significant computational resources, with hardware requirements scaling at an unsustainable rate (Thompson et al., 2021). In constrained settings, this limits the practicality of deploying deep learning methods. In addition, these extensive computations come with an immense environmental cost (Luccioni et al., 2024a,b). Consequently, to address the growing need for more powerful computational tools, controlling computational costs becomes crucial. Neuromorphic computing

(Schuman et al., 2017) offers one promising solution to this problem. This computational paradigm leverages *spiking neural networks* (SNNs) (Maass, 1997b), which are more closely aligned with biological neural networks, and hold the potential to be significantly more energy-efficient than traditional deep neural networks (Frenkel et al., 2023; Lunghi et al., 2024; Yin et al., 2021; Göltz et al., 2021).

A neuron in a spiking neural network is referred to as a spiking neuron. Different from their artificial counterparts commonly used in deep learning, these spiking neurons operate in an inherently temporal manner (Gerstner et al., 2014). The key state variable of spiking neurons is their membrane potential, which resembles the potential difference across the membrane of a biological neuron. The outputs of a spiking neuron consist of all-or-nothing events, namely stereotypical electrical pulses called *spikes*. The connection between two neurons, known as a synapse, converts an incoming spike into either an *excitatory* or *inhibitory* change in the membrane potential of the receiving neuron. When excitatory changes in potential are rapid and significant, typically modeled as exceeding a predefined *threshold*, the receiving neuron in turn emits a spike (Hodgkin and Huxley, 1952). Beyond the basic configuration, a wide range of spiking neuron models have been developed, each designed to meet the demands of specific applications (Izhikevich, 2004; Gerstner and Naud, 2009). For instance, models such as Hodgkin-Huxley (Hodgkin and Huxley, 1952) and FitzHugh-Nagumo (FitzHugh, 1961; Nagumo et al., 1962) provide high biological realism but at the cost of significant computational complexity. On the other hand, in the neuromorphic and neuro-inspired AI communities, where a major focus lies in leveraging spiking neural networks for machine learning applications, threshold-based models such as the *leaky integrate-and-fire* model (LIF) (Lapicque, 1907) and its variations have become the spiking neuron models of choice (Frenkel et al., 2023; Eshraghian et al., 2023).

The energy efficiency of spiking neurons, and spiking neural networks by extension, is rooted in the characteristic all-or-nothing spiking mechanism, which promotes high sparsity in synaptic interactions. Information can then be encoded either in precise *spike times* (Thorpe et al., 1996), spike time sequences (Xie et al., 2024; Dold et al., 2022), or temporally averaged quantities, such as *spike rates* over sequences of spikes (Gerstner et al., 2014). Despite their considerable application potential, developing training methods for SNNs that effectively capture temporal sparsity while achieving competitive performance on machine learning tasks remains a challenging and ongoing area of research (Zenke et al., 2021; Davidson and Furber, 2021; Lunghi et al., 2024). However, recent advancements, such as surrogate gradients (Neftci et al., 2019), have made some progress in narrowing the performance gap with artificial neural networks (ANNs). A key difficulty encountered in training SNNs is the fact that, for many neuron models such as the current-based LIF model, the output of an SNN does not continuously depend on its parameters. Thus, resulting discontinuous changes of spike times, such as sudden disappearance (or re-emergence) of spikes, lead to instabilities during training with gradient-based methods (Eshraghian et al., 2023). A spike disappearance occurs, for example, when increased inhibition drives a neuron’s membrane potential below the threshold (Figure 1A). This is particularly problematic in deep SNNs, where neurons in deeper layers may become inactive, resulting in dead neurons and vanishing gradients during training (Rossbroich et al., 2022; Eshraghian et al., 2023), which is usually countered by adding a term to the loss function that aims at reviving dead neurons

when too many die out (Mostafa, 2017; Göltz et al., 2021). Moreover, these discontinuities prevent the derivation of standard covering number-based learning bounds, complicating a comparison of SNNs with ANNs.

In this work, we examine the simple *spike-response model* (SRM) (Maass and Schmitt, 1997) with single-spike encoding, which has recently obtained increased interest in theoretical studies (Singh et al., 2023; Stanojevic et al., 2023) and deep learning applications (Stanojevic et al., 2024). In the general family of SRMs, which encompasses the family of LIF neurons, each incoming spike triggers a *response*, and the membrane potential is represented as the sum of these spike-responses (Gerstner et al., 2014). In the used simple SRM, these responses are chosen to be linear, where each slope reflects an excitatory or inhibitory effect and is determined by a *synaptic weight*. We discover that simple SRM SNNs exhibit discontinuities with respect to their neural network parameters, including synaptic weights, with inhibition (modeled by synaptic negative weights) being a primary cause (Figure 1B). Such discontinuities crucially hinder the derivation of standard covering number-based learning bounds, complicating a theoretical comparison of SNNs and ANNs. Additionally, they contribute to the challenges associated with neuronal inactivity and training instability, as gradient-based optimization methods cannot accommodate the abrupt changes in the neural network’s output. To address this limitation, we propose a modified type of the simple SRM SNN, which we refer to as *affine SNN*. It is characterized by two key properties:

1. only excitatory responses, or more specifically, only positive synaptic weights, are permitted;
2. in line with (Singh et al., 2023), we add an affine encoding and a decoding layers¹.

The restriction to excitatory synapses carries significant analytical and practical consequences. It is designed to sustain neuronal spike activity, and thereby guarantee well-defined and tractable spike times. An outstanding property of affine SNNs is that their outputs are *Lipschitz continuous* with respect to both neural network inputs and *parameters*, which we show in Theorem 18. This is a property absent in more general SNNs that incorporate inhibitory responses. We demonstrate that, despite the restriction on their weights, affine SNNs are a novel computational paradigm possessing beneficial properties of feedforward neural networks (Cybenko, 1989; He et al., 2020; Barron, 1993). We collect the results in Section 5, where we find the following:

1. Affine SNNs are *universal approximators*, i.e., every continuous function on a compact domain can be arbitrarily well approximated by an affine SNN (Theorem 26).
2. Affine SNNs can replicate approximation results achieved by *linear finite elements* (Theorem 29). In particular, they approximate, at optimal approximation rates, Sobolev-regular smooth functions (Theorem 31).

1. In our context, the adjective “affine” in the term “affine SNN” refers to the affine encoding and decoding layers. Since our model uses only positive synaptic weights, we adopt the term for brevity. However, we may refer to, say, “affine SNNs with real-valued synaptic weights” in our experiments when describing the same model extended to allow negative synaptic weights.

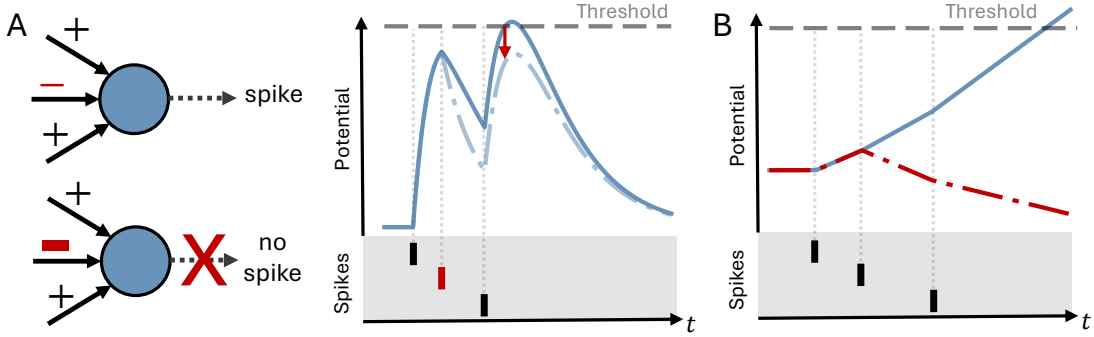


Figure 1: **A** Example of the response of a current-based LIF neuron to three input spikes. When increasing the inhibitory input (red), the potential gets pulled below the threshold (dash-dotted line) and the neuron stops spiking. **B** Potential of a simple spike-response neuron given three input spikes. In contrast to the case with only excitatory inputs, i.e., positive weights (blue line), the neuron becomes silent if the slope of the response is negative (red, dash-dotted line).

3. Affine SNNs can produce dimension-independent approximation rates for Barron-regular functions (Theorem 33).

Our proposed class of affine SNNs occupies an intermediate position between shallow and deep ReLU networks in the approximation-generalization trade-off; such characterization is also supported by our empirical results. Notably, we find that, in order to keep the generalization gap under control, we need a number of training samples that depends only linearly on the number of parameters (Theorem 21). In particular, unlike feedforward neural networks or previous VC dimension-based analyses of SNNs (Schmitt, 1999; Maass and Schmitt, 1997), our generalization bounds scale at most *logarithmically with the depth of the network graph*, highlighting a striking advantage in efficiency—in terms of sample complexity—and scalability (Theorem 21). Overall, we observe the following key property of affine SNNs: The capacity cost for learning, i.e., the complexity of the hypothesis set, is bounded with at most logarithmic dependence on the depth of the underlying affine SNNs. Hence, affine SNNs can solve problems that shallow feedforward neural networks cannot solve at practically no higher capacity cost.

We complement these theoretical results with experimental simulations in Section 7, enabled by the continuous nature of affine SNNs which renders them *particularly well-suited for classical gradient-based training*. In particular, we demonstrate that affine SNNs feature superior generalization properties in a simple regression task than shallow and deep ReLU neural networks (Figure 5A), and that they generalize closer to shallow than deep ReLU neural networks in a classification task (Figure 5B). We further show that affine SNNs reach competitive performance levels on standard machine learning benchmarks with respect to deep feedforward networks such as MNIST ($96.75^{+0.02}_{-0.08}$ % median test accuracy, with upper and lower index being the distance to the third and first quartile) and Fashion MNIST ($87.81^{+0.04}_{-0.60}$ % median test accuracy).

1.1 Related work

Feedforward neural networks. In the last few years, a decent understanding of the learning theory of deep neural networks has been established. For comprehensive overviews, we refer to (Berner et al., 2021; Anthony and Bartlett, 1999; Petersen and Zech, 2024). Learning theory is typically split into two aspects: first, the expressivity of architecture, i.e., how well a certain type of neural networks can represent a set of functions of interest, and second, generalization bounds, which describe the mismatch between the performance of a trained model on the training set and unseen data points.

The approximation theory of feedforward neural networks is comparatively very well understood. First of all, universality properties have been shown for various architectures (Cybenko, 1989; Hornik et al., 1990; Kidger and Lyons, 2020). Moreover, for specific function classes, approximation rates can be derived. For example, focusing only on feedforward neural networks with the ReLU activation function, it was established that neural networks could reproduce approximation by linear and higher order finite elements (He et al., 2020; Opschoor et al., 2020), achieve optimal approximation of smooth functions (Yarotsky, 2017; Shen et al., 2022), and approximate high-dimensional functions without curse of dimensionality (Barron, 1993; Parhi and Nowak, 2022; Caragea et al., 2023; Lerma-Pineda et al., 2024).

Regarding learning guarantees, classical statistical learning theory facilitates generalization through, for example, VC dimension, covering number or pseudo-dimension bounds (Shalev-Shwartz and Ben-David, 2014; Berner et al., 2021; Anthony and Bartlett, 1999). Applied specifically to ReLU neural networks, such generalization bounds are, for example, derived in (Berner et al., 2020; Schmidt-Hieber, 2020).

It has to be mentioned that in the context of modern machine learning applications using overparameterized architectures, classical statistical learning theory-based arguments are potentially not the best possible, and different tools are required (Zhang et al., 2021; Belkin et al., 2019).

Spiking neural networks. The research on SNNs is vast and has been developed over multiple decades. For a more comprehensive overview, we refer to the survey articles (Maass, 1997b; Paugam-Moisy, 2006; Grüning and Bohte, 2014; Eshraghian et al., 2023). Generally, it has been shown that SNNs can represent specific functions, such as coincidence detectors, with significantly fewer parameters compared to feedforward neural networks (Abeles, 1982), making them an attractive alternative in resource-constrained environments. Moreover, they have been shown to enable inference with very low times-to-solution using the time-to-first-spike paradigm (Göltz et al., 2021; Stanojevic et al., 2024) as well as highly energy-efficient deep learning solutions (Göltz et al., 2021; Yin et al., 2021; Lunghi et al., 2024).

In the family of spike response models with temporally encoded inputs, the expressivity of SNNs has been studied extensively, e.g., (Maass, 1996, 1997b, 2015; Singh et al., 2023; Stanojevic et al., 2023). This includes the universal approximation property (Mostafa, 2017; Kheradpisheh and Masquelier, 2020; Comsa et al., 2020; Göltz et al., 2021; Klos and Memmesheimer, 2025) as well as other, more quantitative approximation rates (Singh et al., 2023). Our results distinguish themselves from previous work due to our requirement of positive weights. Even though at first glance, this offers significantly less flexibility regarding an SNN’s parametrization, it opens up a new angle for studying SNNs, with remarkable

Figure 2: A network graph G with two input nodes in burgundy and one output node in violet

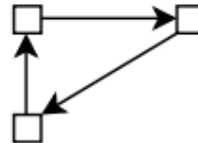
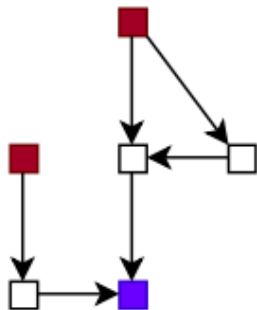


Figure 3: An example of a forbidden directed cycle

results. Concerning learning rates, it was shown in (Maass and Schmitt, 1997, 1999; Schmitt, 1999) that classical statistical learning theory bounds can be derived. These are in terms of the VC dimension or *pseudo-dimension*. In contrast, our results use covering number estimates and yield stronger generalization guarantees since the upper bounds only depend logarithmically on the depth of the SNNs.

2. Notions of spiking neural networks

In this section, we introduce the central concepts of this paper. Specifically, we will describe the SNN architectures and the concept of *spike times*. SNNs are based on the so-called network graphs introduced below.

Definition 1 *A (finite) directed, unweighted graph $G = (V, E)$ satisfying the following properties:*

1. G has no directed cycles,
2. G has no isolated nodes,

is called a directed acyclic graph or a network graph. We denote the set of all nodes with no incoming edges as V_{in} , referred to as the input nodes, and the set of nodes with no outgoing edges as V_{out} , referred to as the output nodes. We refer to the length of the longest directed path in G as the (graph) depth of G .

Figures 2, 3 provide an example and a non-example of a network graph, respectively.

Based on a network graph, various definitions of SNNs are possible. In Subsection 2.1, we specify our model of SNNs, and in Subsection 2.2, we draw comparisons between our setups and others documented in the literature. Following this, in Subsection 2.3, we introduce

affine SNNs. Finally, in Subsection 2.4, we present a useful operation applicable to affine SNNs, facilitating the construction of more complex neural networks.

In this paper, we denote the cardinality of a set S by $\#S$. Drawing inspiration from graph theory, neural networks, and biology, we will also use the terms “node”, “vertex” and “neuron” interchangeably.

2.1 Spiking neural network model

We now present the definition of a general SNN as an architecture, followed by a description of its dynamics. Afterward, we introduce a special type of *positive SNNs*.

Definition 2 Let $G = (V, E)$ be a network graph with a subset $V_{\text{in}} \subset V$ of input neurons, a subset $V_{\text{out}} \subset V$ of output neurons, and a set $E \subset V \times V$ of synapses. Each synapse $(u, v) \in E$ is a directed edge, associated with the following attributes

1. a response function $\varepsilon_{(u,v)}: \mathbb{R} \rightarrow \mathbb{R}$,
2. a synaptic delay $\theta_{(u,v)} \geq 0$,
3. a synaptic weight $\mathbf{w}_{(u,v)} \geq 0$.

Moreover, for every $v \in V \setminus V_{\text{in}}$, there exists $(u, v) \in E$ such that $\mathbf{w}_{(u,v)} > 0$.

Lastly, let $\mathbf{W} := (\mathbf{w}_{(u,v)})_{(u,v) \in E}$, $D := (\theta_{(u,v)})_{(u,v) \in E}$, and $\mathcal{E} := (\varepsilon_{(u,v)})_{(u,v) \in E}$ be the tuple of synaptic weights, synaptic delays, and response functions, respectively. Then a spiking neural network (SNN) is a tuple $\Phi = (G, \mathbf{W}, D, \mathcal{E})$.

In the sequel, we will focus on SNNs for which the response function is the same for all edges. Concretely, inspired by (Singh et al., 2023), we adopt a unified response function modeled after the ReLU activation function.

Definition 3 Let G be a network graph, and let (u, v) be one of its synapses. We define the response function $\varepsilon_{(u,v)}: \mathbb{R} \rightarrow \mathbb{R}$ associated with (u, v) as follows

$$\varepsilon_{(u,v)}(t) := \varrho(t) \tag{1}$$

where $\varrho(t) = \max\{t, 0\}$ denotes the ReLU activation function.

Information is transmitted through an SNN via spike times, triggered when the membrane potential, or simply the *potential*, reaches a critical threshold. We state the mathematical model below.

Definition 4 Let $\Phi = (G, \mathbf{W}, D, \mathcal{E})$ be an SNN, and let $G = (V, E)$ be a network graph. Let $\varepsilon_{(u,v)} \in \mathcal{E}$ be defined as in (1). Let $t_u \in \mathbb{R}$ for $u \in V_{\text{in}}$. Then, for $v \in V \setminus V_{\text{in}}$, we define the potential at v as $P_v: \mathbb{R} \rightarrow \mathbb{R}$, where for $t \in \mathbb{R}$

$$P_v(t) := \sum_{(u,v) \in E} \mathbf{w}_{(u,v)} \varepsilon_{(u,v)}(t - t_u - \theta_{(u,v)}) = \sum_{(u,v) \in E} \mathbf{w}_{(u,v)} \varrho(t - t_u - \theta_{(u,v)}). \tag{2}$$

Here in (2), $t_u = \min\{t \in \mathbb{R}: P_u(t) = 1\}$ is the spike time at u if $u \in V \setminus V_{\text{in}}$. Subsequently, the spike time t_v at v is given by $t_v = \min\{t \in \mathbb{R}: P_v(t) = 1\}$.

A few remarks are in order. First, the provided definition may appear circular, as the spike time at a noninput neuron is determined by its potential, which, in turn, hinges on the spike times of other presynaptic neurons. Second, it does not inherently guarantee that $\min\{t \in \mathbb{R} : P_u(t) = 1\}$ is nonempty.

The following lemma demonstrates the well-definedness of P_v and t_v for all $v \in V \setminus V_{\text{in}}$, when the responses $\varepsilon_{(u,v)}$ are of the form (1). A proof is given in Appendix A.1.

Lemma 5 *Let $\Phi = (G, W, D, \mathcal{E})$ be an SNN where $G = (V, E)$ is a network graph. Let $\varepsilon_{(u,v)} \in \mathcal{E}$ be defined as in (1). Let $t_u \in \mathbb{R}$ for $u \in V_{\text{in}}$. Then P_v and t_v are well-defined for all $v \in V \setminus V_{\text{in}}$.*

We conclude this subsection by formalizing the selected SNNs for this work, referred to as positive SNNs.

Definition 6 *Let $\Phi = (G, W, D, \mathcal{E})$ be an SNN with $G = (V, E)$ being a network graph. Then Φ is a positive SNN if for all $(u, v) \in E$, $w_{(u,v)} > 0$, and $\varepsilon_{(u,v)}$ is given by (1). Moreover, if Φ is a positive SNN, we streamline its tuple notation as $\Phi = (G, W, D)$.*

Since Lemma 5 asserts in particular that t_v exists for $v \in V_{\text{out}}$ once t_u is assigned for all $u \in V_{\text{in}}$ in a positive SNN Φ , we can define a map taking all input spike times to the output spike times. This function is called the *realization of Φ* .

Definition 7 *Let $\Phi = (G, W, D)$ be a positive SNN. Let $d_{\text{in}}, d_{\text{out}}$ denote the cardinality of $V_{\text{in}}, V_{\text{out}}$, respectively. Then the realization of Φ , $R(\Phi) : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^{d_{\text{out}}}$, is a function whose inputs are $(t_v)_{v \in V_{\text{in}}} \in \mathbb{R}^{d_{\text{in}}}$ and whose outputs are $(t_v)_{v \in V_{\text{out}}} \in \mathbb{R}^{d_{\text{out}}}$, where t_v denotes the spike time at neuron v .*

It is to be understood from the definition above and throughout this paper that we assume a consistent enumeration of the input and output neurons.

2.2 Model discussion

In Definition 6 of positive SNNs $\Phi = (G, W, D)$, the synaptic weights are taken to be positive, and the response functions assume the form (1). This leads to an exclusively monotone increase of the potential that always crosses a positive threshold from below at a unique, and analytically calculable, time. This property is central to our analysis, which allows us to prove Lipschitz continuity (Section 3) with respect to the neural network parameters and derive generalization bounds based on covering numbers (Section 5). In line with the points raised in the introduction, we offer two examples in Appendix B demonstrating that the inclusion of negative synaptic weights will lead to the lack of continuity with respect to both network parameters as well as input spike times. In addition, the choice of a linear response function aligns with a wider class of spike-response models (Maass, 1997a) that feature linear synaptic response between neurons u and v given by

$$\tilde{\varepsilon}_{(u,v)}(t) := \begin{cases} 0 & \text{if } t \notin [0, \delta] \\ \varrho(t) & \text{if } t \in [0, \delta] \end{cases},$$

for $t \in \mathbb{R}$ and $\delta \in (0, \infty]$. On the one hand, when $\delta < \infty$, this reduces to a generalized LIF model (Gerstner, 1995). On the other hand, setting $\delta = \infty$, we recover (1), which has also been used in (Singh et al., 2023; Stanojevic et al., 2023, 2024). Due to the linear rise, similar to other types of integrate-and-fire neurons (Mostafa, 2017; Stanojevic et al., 2023, 2024), information about previous input spikes is never forgotten until the neuron spikes itself. Furthermore, information is encoded purely in the starting time of the linear rise.

2.3 Spiking neural networks with affine encoders and decoders

To contextualize and motivate the introduction of affine SNNs, we inspect the implications of Definitions 3 and 7. Let $\Phi = (G, W, D)$ be a positive SNN associated with the network graph $G = (V, E)$. Consider two tuples of input spike times $(t_u)_{u \in V_{\text{in}}}$, $(\tilde{t}_u)_{u \in V_{\text{in}}}$, such that $\tilde{t}_u \geq t_u$, for all $u \in V_{\text{in}}$. For $v \in V_{\text{out}}$, let t_v, \tilde{t}_v denote the respective corresponding spike times at v . Then it follows directly from (2) that $t_v \geq \min\{t_u : u \in V_{\text{in}}\}$ and that $\tilde{t}_v \geq t_v$. As a consequence, Definition 7 implies monotonicity of the function $R(\Phi)$. Naturally, this constitutes a strong limitation on the functions expressible by positive SNNs. As a remedy, we use affine encoders and decoders to amend the neural network construction. Note that the use of general encoders and decoders for SNNs has already been introduced in (Singh et al., 2023).

Definition 8 Let $d_{\text{in}}, d_{\text{out}}, d_0, d_1 \in \mathbb{N}$. A spiking neural network equipped with an affine encoder and decoder, or an affine spiking neural network (affine SNN), is a triple $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$. Here, $\Phi = (G, W, D)$ is a positive SNN, with $\#V_{\text{in}} = d_{\text{in}}$, $\#V_{\text{out}} = d_{\text{out}}$, and $A_{\text{in}}: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{\text{in}}}$, $A_{\text{out}}: \mathbb{R}^{d_{\text{out}}} \rightarrow \mathbb{R}^{d_1}$ are two affine maps, called an encoder and a decoder, respectively, such that for $x \in \mathbb{R}^{d_0}$ and $z \in \mathbb{R}^{d_{\text{out}}}$

$$A_{\text{in}}(x) = W_{\text{in}}x + b_{\text{in}} \quad \text{and} \quad A_{\text{out}}(z) = W_{\text{out}}z + b_{\text{out}},$$

where $W_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times d_0}$, $W_{\text{out}} \in \mathbb{R}^{d_1 \times d_{\text{out}}}$, $b_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$, $b_{\text{out}} \in \mathbb{R}^{d_1}$.

The realization of an affine SNN $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$ is given by $R(\Psi): \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_1}$, where $R(\Psi) = A_{\text{out}} \circ R(\Phi) \circ A_{\text{in}}$.

Next, we address the quantification of the size of these neural networks. With our adoption of directed acyclic graphs as network structures, there is no *canonical* concept of layers². Consequently, we evaluate the neural network’s size by examining attributes such as the number of synaptic weights and delays alongside the conventional sizing metrics associated with the affine encoder and decoder maps.

Definition 9 Let $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$ be an affine SNN. The size of Ψ , denoted $\text{Size}(\Psi)$, is defined by the total number of nonzero scalar entries in the tuple $(W, D, W_{\text{in}}, W_{\text{out}}, b_{\text{in}}, b_{\text{out}})$, i.e.,

$$\text{Size}(\Psi) := \|(W, D, W_{\text{in}}, W_{\text{out}}, b_{\text{in}}, b_{\text{out}})\|_{\ell^0}.$$

In our study, it will sometimes be important to guarantee that an affine SNN does not have arbitrarily large outputs. This requirement is standard in the analysis of learning properties of feedforward neural networks (Berner et al., 2020, Setting 2.5), (Schmidt-Hieber,

2. However, see a graph layering algorithm presented in the proof of Lemma 38 in Appendix A.3.

2020, Equation (4)). To enforce this, we introduce a *clipped realization* of an affine SNN below.

Definition 10 *Let Ψ be an affine SNN. Let $I \subset \mathbb{R}$ be a compact interval. The I -clipped realization of Ψ is given by $R_I(\Psi): \mathbb{R}^{d_0} \rightarrow I^{d_1}$, where $R_I(\Psi) = \text{clip}_I \circ R(\Psi)$. Here, for $x \in \mathbb{R}^{d_1}$ and $i = 1, \dots, d_1$*

$$(\text{clip}_I(x))_i = \arg \min\{|x_i - z| : z \in I\}$$

where x_i denotes the i -th coordinate of x .

For ease of reference, we present Table 1, which summarizes the symbols associated with SNNs and affine SNNs that will be consistently used throughout the remainder of this paper.

Symbols	Default meaning
Ψ	Affine SNN
Φ	SNN
G	Network graph
W	Synaptic weight tuple
D	Synaptic delay tuple
V_{in}	Input nodes of G
V_{out}	Output nodes of G
d_{in}	Cardinality of V_{in} and input dimension of $R(\Phi)$
d_{out}	Cardinality of V_{out} and output dimension of $R(\Phi)$
A_{in}	Affine encoder of Ψ
A_{out}	Affine decoder of Ψ
W_{in}	Matrix associated with A_{in}
W_{out}	Matrix associated with A_{out}
b_{in}	Shift associated with A_{in}
b_{out}	Shift associated with A_{out}
d_0	Input dimension of $R(\Psi)$
d_1	Output dimension of $R(\Psi)$
$\text{Size}(\Psi)$	Size of Ψ

Table 1: Commonly used symbols and their meanings

2.4 Addition of affine spiking neural networks

We introduce addition for affine SNNs, a commutative operation on pairs of SNNs with matching input and output dimensions. In later sections, we will employ affine SNN addition to reproduce approximation results based on the superposition of simple functions.

Let $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$ be an affine SNN, where $\Phi = (G, W, D)$ is a positive SNN. In what follows, we refer to $A_{\text{in}}(\Psi)$ and $A_{\text{out}}(\Psi)$ as the encoder and decoder of Ψ , respectively. Additionally, we use $G(\Phi)$, $W(\Phi)$, and $D(\Phi)$ to denote the network graph, the synaptic weight and synaptic delay matrices associated with Φ , respectively.

Definition 11 *Let*

$$\Psi = (A_{\text{in}}, \Phi, A_{\text{out}}) \quad \text{and} \quad \Psi' = (A'_{\text{in}}, \Phi', A'_{\text{out}}),$$

be two affine SNNs, associated with positive SNNs,

$$\Phi = (G, W, D) \quad \text{and} \quad \Phi' = (G', W', D'),$$

respectively. Let for $x \in \mathbb{R}^{d_0}$, $y \in \mathbb{R}^{d_{\text{out}}}$, and $z \in \mathbb{R}^{d'_{\text{out}}}$,

$$\begin{aligned} A_{\text{in}}(x) &= W_{\text{in}}x + b_{\text{in}} & \text{and} & & A_{\text{out}}(y) &= W_{\text{out}}y + b_{\text{out}}, \\ A'_{\text{in}}(x) &= W'_{\text{in}}x + b'_{\text{in}} & \text{and} & & A'_{\text{out}}(z) &= W'_{\text{out}}z + b'_{\text{out}}, \end{aligned}$$

where $W_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times d_0}$, $W_{\text{out}} \in \mathbb{R}^{d_1 \times d_{\text{out}}}$, $b_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$, $b_{\text{out}} \in \mathbb{R}^{d_1}$, and $W'_{\text{in}} \in \mathbb{R}^{d'_{\text{in}} \times d_0}$, $W'_{\text{out}} \in \mathbb{R}^{d'_1 \times d'_{\text{out}}}$, $b'_{\text{in}} \in \mathbb{R}^{d'_{\text{in}}}$, $b'_{\text{out}} \in \mathbb{R}^{d'_1}$. Then the addition of Ψ , Ψ' , denoted $\Psi \oplus \Psi'$, is an affine SNN, associated with a positive SNN $\Phi \oplus \Phi'$ such that

$$\begin{aligned} G(\Phi \oplus \Phi') &:= G \cup G' \\ W(\Phi \oplus \Phi') &:= W \cup W' \\ D(\Phi \oplus \Phi') &:= D \cup D', \end{aligned}$$

where input (output) nodes of G are listed before input (output) nodes of G' . Moreover,

$$\begin{aligned} A_{\text{in}}(\Psi \oplus \Psi') &: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{\text{in}} + d'_{\text{in}}} \\ A_{\text{out}}(\Psi \oplus \Psi') &: \mathbb{R}^{d_{\text{out}} + d'_{\text{out}}} \rightarrow \mathbb{R}^{d_1}, \end{aligned}$$

are affine maps, such that, for $x \in \mathbb{R}^{d_0}$,

$$A_{\text{in}}(\Psi \oplus \Psi')(x) := \begin{pmatrix} W_{\text{in}} \\ W'_{\text{in}} \end{pmatrix} x + \begin{pmatrix} b_{\text{in}} \\ b'_{\text{in}} \end{pmatrix},$$

and for $x \in \mathbb{R}^{d_{\text{out}} + d'_{\text{out}}}$,

$$A_{\text{out}}(\Psi \oplus \Psi')(x) := (W_{\text{out}} \quad W'_{\text{out}}) x + (b_{\text{out}} + b'_{\text{out}}).$$

The following result on the size of the addition of affine SNNs follows immediately from the construction.

Lemma 12 *Let Ψ, Ψ' be two affine SNNs. Then*

$$\text{Size}(\Psi \oplus \Psi') = \text{Size}(\Psi) + \text{Size}(\Psi').$$

3. Lipschitz continuity of affine spiking neural networks

Let $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$ be an affine SNN, where $\Phi = (G, W, D)$ is a positive SNN. In this section, we delineate two types of continuity the realization $R(\Psi): \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_1}$ of Ψ exhibits:

1. continuity concerning the *neural network input* $x \in \mathbb{R}^{d_0}$,
2. continuity concerning the *neural network parameters* $(W, D, W_{\text{in}}, b_{\text{in}}, W_{\text{out}}, b_{\text{out}})$.

We explore these types of continuity in the listed order. In addressing the first type of continuity with respect to the neural network input, we offer the following theorem, the proof of which is given in Appendix A.2.

Theorem 13 *Let Ψ be an affine SNN. Then, for $x, \tilde{x} \in \mathbb{R}^{d_0}$,*

$$\|R(\Psi)(x) - R(\Psi)(\tilde{x})\|_{\ell^\infty} \leq (d_0 d_{\text{out}})^{\frac{1}{2}} \|W_{\text{in}}\|_F \|W_{\text{out}}\|_F \|x - \tilde{x}\|_{\ell^\infty},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

To assess the continuity of Ψ in terms of the neural network parameters, we begin with the continuity of Φ concerning (W, D) . In this context, we fix a network graph G and examine the discrepancies between two positive SNNs constructed on G , $\Phi = (G, W, D)$ and $\tilde{\Phi} = (G, \tilde{W}, \tilde{D})$. The following proposition serves as a foundation for the ensuing discussions. A proof is provided in Appendix A.3.

Proposition 14 *Let $\Phi = (G, W, D)$, $\tilde{\Phi} = (G, \tilde{W}, \tilde{D})$ be two positive SNNs. Suppose there exists $\mathbf{b} > 0$ such that for every $\mathbf{w}_{(u,v)} \in W$, $\tilde{\mathbf{w}}_{(u,v)} \in \tilde{W}$,*

$$\min \{\mathbf{w}_{(u,v)}, \tilde{\mathbf{w}}_{(u,v)}\} \geq \mathbf{b}. \quad (3)$$

Let L be the graph depth of G . Then for every $t \in \mathbb{R}^{d_{\text{in}}}$,

$$\|R(\Phi)(t) - R(\tilde{\Phi})(t)\|_{\ell^\infty} \leq L \cdot \left(1 + \frac{1}{\mathbf{b}^2}\right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty}, \quad (4)$$

where

$$\|\Phi - \tilde{\Phi}\|_{\ell^\infty} := \max \{\|W - \tilde{W}\|_{\ell^\infty}, \|D - \tilde{D}\|_{\ell^\infty}\}. \quad (5)$$

Remark 15 *The constant of the Lipschitz estimate in (4) is almost tight in general. To see this, consider the graph G with N nodes $(v_i)_{i=1}^N$ and edges only between v_i and v_{i+1} for $i = 1, \dots, N-1$. Let $\Phi, \tilde{\Phi}$ be two positive SNNs with respective weights $\mathbf{w}_{(v_i, v_{i+1})} = 2\mathbf{b}$, $\tilde{\mathbf{w}}_{(v_i, v_{i+1})} = \mathbf{b}$ and delays $\theta_{(v_i, v_{i+1})} = 0$, $\tilde{\theta}_{(v_i, v_{i+1})} = \mathbf{b}$ for $i = 1, \dots, N-1$. Then, it is immediate for every $t \in \mathbb{R}$ that*

$$R(\Phi)(t) = t + (N-1) \frac{1}{2\mathbf{b}} \quad \text{and} \quad R(\tilde{\Phi})(t) = t + (N-1) \left(\frac{1}{\mathbf{b}} + \mathbf{b}\right).$$

Subsequently,

$$|R(\Phi)(t) - R(\tilde{\Phi})(t)| = (N-1) \left(\frac{1}{2\mathbf{b}} + \mathbf{b}\right) = (N-1) \left(\frac{1}{2\mathbf{b}^2} + 1\right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty}.$$

Remark 16 *The previous Remark 15 demonstrates the importance of strictly positive synaptic weights indicated in (3). Indeed, if this condition is not met, it is generally not possible to constrain the difference in two neural network outputs by a multiple of the difference in their neural network parameters. This can be observed by choosing arbitrarily small \mathbf{b} in Remark 15.*

Remark 17 *The global Lipschitz continuity of positive SNNs should be contrasted with the local Lipschitz continuity of feedforward neural networks (Petersen et al., 2021, Proposition 4). For feedforward neural networks, the difference in the realizations of the neural networks can only be controlled by the \mathbf{L} -th power of the ℓ^∞ difference of the neural network weights, where \mathbf{L} is the graph depth.*

Continuing, we present a theorem that expands upon Proposition 14, clarifying the Lipschitz continuity of an affine SNN in relation to its neural network parameters. A proof is included in Appendix A.4.

Theorem 18 *Let $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$, $\tilde{\Psi} = (\tilde{A}_{\text{in}}, \tilde{\Phi}, \tilde{A}_{\text{out}})$ be two affine SNNs, where $\Phi = (G, \mathbf{W}, D)$, $\tilde{\Phi} = (G, \tilde{\mathbf{W}}, \tilde{D})$ are two positive SNNs. Suppose there exist $\mathbf{b}, \mathbf{B} \in (0, \infty)$ such that for every $\mathbf{w}_{(u,v)} \in \mathbf{W}$, $\tilde{\mathbf{w}}_{(u,v)} \in \tilde{\mathbf{W}}$,*

$$\min \{ \mathbf{w}_{(u,v)}, \tilde{\mathbf{w}}_{(u,v)} \} \geq \mathbf{b},$$

and for every $\theta_{(u,v)} \in D$, $\tilde{\theta}_{(u,v)} \in \tilde{D}$

$$\max \{ \theta_{(u,v)}, \tilde{\theta}_{(u,v)} \} \leq \mathbf{B}.$$

Let

$$\begin{aligned} W_{\text{out}}^* &:= \max \{ \|W_{\text{out}}\|_F, \|\tilde{W}_{\text{out}}\|_F \} \\ W_{\text{in}}^* &:= \max \{ \|W_{\text{in}}\|_F, \|\tilde{W}_{\text{in}}\|_F \} \\ b_{\text{in}}^* &:= \max \{ \|b_{\text{in}}\|_{\ell^\infty}, \|\tilde{b}_{\text{in}}\|_{\ell^\infty} \}. \end{aligned}$$

Then for every $x \in \mathbb{R}^{d_0}$,

$$\|R(\Psi)(x) - R(\tilde{\Psi})(x)\|_{\ell^\infty} \leq B_1 + B_2 \tag{6}$$

where, with \mathbf{L} denoting the graph depth of G ,

$$B_1 := d_{\text{out}}^{\frac{1}{2}} W_{\text{out}}^* \cdot \left(d_0^{\frac{1}{2}} \|W_{\text{in}} - \tilde{W}_{\text{in}}\|_F \|x\|_{\ell^\infty} + \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty} + \|b_{\text{in}} - \tilde{b}_{\text{in}}\|_{\ell^\infty} \right),$$

and

$$B_2 := d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}} - \tilde{W}_{\text{out}}\|_F \left(d_0^{\frac{1}{2}} W_{\text{in}}^* \|x\|_{\ell^\infty} + b_{\text{in}}^* + \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right) \right) + \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty}.$$

4. Generalization bounds for affine spiking neural networks

In this section, we derive generalization bounds for the problem of learning functions from finitely many samples using affine SNNs. We start by defining a learning problem.

Let Ω be a compact domain in a Euclidean space. We assume that there is an *unknown* probability distribution \mathcal{D} on $\Omega \times [0, 1]$. In a learning problem, our objective is to select a member from an appropriate hypothesis set \mathcal{H} of functions mapping Ω to $[0, 1]$ that fits \mathcal{D} best. Concretely, we want to find $g \in \mathcal{H}$ that minimizes the *risk* \mathcal{R} defined as

$$\mathcal{R}(g) := \mathbb{E}_{(x,y) \sim \mathcal{D}} |g(x) - y|^2, \quad (7)$$

over \mathcal{H} . Since we do not know \mathcal{D} , this optimization problem cannot be solved directly. Instead, we assume that we are given a sample of size $m \in \mathbb{N}$ of observations drawn i.i.d. from \mathcal{D} , i.e., $S = (x_i, y_i)_{i=1}^m \sim \mathcal{D}^m$. Based on this sample, we define the *empirical risk* $\widehat{\mathcal{R}}_S$ to be

$$\widehat{\mathcal{R}}_S(g) := \frac{1}{m} \sum_{i=1}^m |g(x_i) - y_i|^2. \quad (8)$$

A function $g_S \in \arg \min_{h \in \mathcal{H}} \widehat{\mathcal{R}}_S(h)$ is called an *empirical risk minimizer*. Such a function serves as a potential approximate minimizer for the optimization problem of the risk. Hence, g_S approximates a solution to the learning problem if it can be shown that the risk and the empirical risk do not differ too much. In this section, we focus on bounding the risk (7) by the empirical risk (8) for a hypothesis class \mathcal{H} consisting of clipped realizations of affine SNNs, defined in (11) below, up to a small additive term. The main result, Theorem 21, is provided at the end.

For a parameterized hypothesis class, a well-known strategy in statistical learning theory involves leveraging the Lipschitz property of the parameterization to assess the so-called covering numbers, which are then used to upper bound the *generalization error*

$$\sup_{g \in \mathcal{H}} |\mathcal{R}(g) - \widehat{\mathcal{R}}_S(g)|.$$

This approach has been employed in the context of feedforward neural networks (Berner et al., 2020; Grohs and Voigtlaender, 2024; Schmidt-Hieber, 2020). For the sake of concreteness, we recall the covering number as follows.

Definition 19 *Let Ω be a relatively compact subset of a metric space $(\mathcal{X}, \mathfrak{d})$. For $\varepsilon > 0$, we call*

$$\mathcal{N}(\Omega, \varepsilon, \mathcal{X}) := \min \left\{ m \in \mathbb{N} : \exists (x_i)_{i=1}^m \in \mathcal{X}^m \text{ such that } \bigcup_{i=1}^m \overline{B(x_i, \varepsilon)} \supset \Omega \right\},$$

the ε -covering number of Ω in \mathcal{X} , where $\overline{B(x, \varepsilon)} := \{z \in \mathcal{X} : \mathfrak{d}(z, x) \leq \varepsilon\}$ is the closed ball with radius ε centered at x .

Further, suppose that f is an C_{Lip} -Lipschitz continuous map for some $C_{\text{Lip}} > 0$ from a metric space \mathcal{X} to $L^\infty(\mathbb{R}^{d_0})$. Then it is straightforward to verify that

$$\mathcal{N}(f(\Omega), L\varepsilon, L^\infty(\mathbb{R}^{d_0})) \leq \mathcal{N}(\Omega, \varepsilon, \mathcal{X}). \quad (9)$$

Let $\mathbf{b} \in (0, 1]$, $\mathbf{B} \in [1, \infty)$. Let $G = (V, E)$ be a *fixed* network graph with $\#V_{\text{in}} = d_{\text{in}}$ and $\#V_{\text{out}} = d_{\text{out}}$. Consider the class of affine SNNs $\Psi = (A_{\text{in}}, \Phi, A_{\text{out}})$ whose positive SNN $\Phi = (G, W, D)$ is built on G . Using this, we define the following *parameterized* class of these affine SNNs

$$\begin{aligned} \mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B}) := \left\{ \Psi = (A_{\text{in}}, \Phi, A_{\text{out}}) : \right. & W_{\text{in}} \in [-\mathbf{B}, \mathbf{B}]^{d_{\text{in}} \times d_0}, b_{\text{in}} \in [-\mathbf{B}, \mathbf{B}]^{d_{\text{in}}}, \\ & W \in [\mathbf{b}, \mathbf{B}]^{\#E}, D \in [0, \mathbf{B}]^{\#E}, \\ & \left. W_{\text{out}} \in [-\mathbf{B}, \mathbf{B}]^{d_1 \times d_{\text{out}}}, b_{\text{out}} \in [-\mathbf{B}, \mathbf{B}]^{d_1} \right\}, \end{aligned} \quad (10)$$

where $\vec{d} := (d_0, d_1, d_{\text{in}}, d_{\text{out}})$. We equip $\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})$ with the metric

$$\mathfrak{d}(\Psi, \tilde{\Psi}) := \max\{\|W_{\text{in}} - \tilde{W}_{\text{in}}\|_{\ell^\infty}, \|b_{\text{in}} - \tilde{b}_{\text{in}}\|_{\ell^\infty}, \|\Phi - \tilde{\Phi}\|_{\ell^\infty}, \|W_{\text{out}} - \tilde{W}_{\text{out}}\|_{\ell^\infty}, \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty}\},$$

where we recall from (5) that for $\Phi = (G, W, D)$, $\tilde{\Phi} = (G, \tilde{W}, \tilde{D})$,

$$\|\Phi - \tilde{\Phi}\|_{\ell^\infty} = \max\{\|W - \tilde{W}\|_{\ell^\infty}, \|D - \tilde{D}\|_{\ell^\infty}\}.$$

Then it is readily seen that $\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})$ is isometrically isomorphic to a compact subset of $[-\mathbf{B}, \mathbf{B}]^M$ where $M := d_{\text{in}}d_0 + 2\#E + d_1d_{\text{out}}$.

Let $d_1 = 1$. Our hypothesis set of choice is as follows,

$$\mathcal{H} = R_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})). \quad (11)$$

Here, $R_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B}))$ denotes the image set of the map $R_{[0,1]}$, which is the $[0, 1]$ -clipped realization map (see Definition 10),

$$R_{[0,1]} : \mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B}) \rightarrow L^\infty(\mathbb{R}^{d_0}).$$

Next, we compute the covering number

$$\mathcal{N}(R_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \varepsilon, L^\infty(\mathbb{R}^{d_0})),$$

a critical quantity that will appear in Theorem 21. By leveraging Theorem 18 in Section 3, it is straightforward to see that $R_{[0,1]}$ is a Lipschitz continuous map on \mathcal{H} with the Lipschitz constant C_{Lip} , where

$$C_{\text{Lip}} \leq C_{\text{Lip}}^* := d_{\text{out}} \cdot \left(2d_{\text{in}}^{\frac{1}{2}} d_0 \mathbf{B} + \mathbf{B} \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) + 2\mathbf{B} + \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right) \right) + 1. \quad (12)$$

Therefore, we conclude, using (9) and the well-known fact that the ε -covering number of $[-\mathbf{B}, \mathbf{B}]^M$ is $\lceil 2\mathbf{B}/\varepsilon \rceil^M$, that

$$\log \left(\mathcal{N}(R_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \varepsilon, L^\infty(\mathbb{R}^{d_0})) \right) \leq M \log \left(\left\lceil \frac{2\mathbf{B}C_{\text{Lip}}^*}{\varepsilon} \right\rceil \right). \quad (13)$$

Remark 20 Note that the logarithm of the covering number derived in (13) depends linearly on the total number of parameters and only logarithmically on the depth of the underlying graph. This is a remarkable property since, for feedforward neural networks, the logarithm of the covering numbers depends linearly on the product of the number of layers and the number of parameters, see e.g. (Schmidt-Hieber, 2020, Remark 1), (Berner et al., 2020, Proposition 2.8).

However, it should be emphasized that this improved Lipschitz control of affine SNNs comes at the expense of representational efficiency, as significantly more parameters are needed to match the expressive capacity of deep FNNs.

We are now prepared to present the two main results of this section.

Theorem 21 Let $d_1 = 1$, $d_0, d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$, $\mathbf{b} \in (0, 1]$, and $\mathbf{B} \in [1, \infty)$. Let G be a network graph with d_{in} input neurons and d_{out} output neurons. Let \mathcal{D} be a distribution on $[0, 1]^{d_0} \times [0, 1]$, and let $S \sim \mathcal{D}^m$ be a sample. Then it holds for all $\Psi \in \mathcal{P}_{\text{SNN}}^*(G, \vec{\mathbf{d}}; \mathbf{b}, \mathbf{B})$ that with probability $1 - \delta$

$$\mathcal{R}(\mathcal{R}_{[0,1]}(\Psi)) \leq \widehat{\mathcal{R}}_S(\mathcal{R}_{[0,1]}(\Psi)) + \sqrt{\frac{2(M \log(m \lceil 16BC_{\text{Lip}}^* \rceil) + \log(2/\delta))}{m}},$$

for all

$$m \geq 2(M \log(m \lceil 16BC_{\text{Lip}}^* \rceil) + \log(2/\delta)). \quad (14)$$

A proof of Theorem 21 is given in Appendix A.5.

A sharper learning bound can be established for the generalization error of an empirical risk near-minimizer in the noiseless setting, where the sample $S = (x_i, f_0(x_i))_{i=1}^m$ is generated deterministically from a target function f_0 .

Theorem 22 Let $d_1 = 1$, $d_0, d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$, $\mathbf{b} \in (0, 1]$, and $\mathbf{B} \in [1, \infty)$. Let G be a network graph with d_{in} input neurons and d_{out} output neurons. Let $f_0: [0, 1]^{d_0} \rightarrow [0, 1]$. Let \mathcal{D} be a distribution on $[0, 1]^{d_0}$, and let $S = (x_i, f_0(x_i))_{i=1}^m$ with $x_i \sim \mathcal{D}$ i.i.d. Then, for $\Psi_m \in \mathcal{P}_{\text{SNN}}^*(G, \vec{\mathbf{d}}; \mathbf{b}, \mathbf{B})$ satisfying

$$\widehat{\mathcal{R}}_S(\mathcal{R}_{[0,1]}(\Psi_m)) \leq \inf_{\Psi \in \mathcal{P}_{\text{SNN}}^*(G, \vec{\mathbf{d}}; \mathbf{b}, \mathbf{B})} \widehat{\mathcal{R}}_S(\mathcal{R}_{[0,1]}(\Psi)) + \frac{\sqrt{2} - 1}{m}, \quad (15)$$

it holds for all $\varepsilon \in (0, 1]$ that,

$$\begin{aligned} & \mathcal{R}(\mathcal{R}_{[0,1]}(\Psi_m)) \\ & \leq 4 \left(\inf_{\Psi \in \mathcal{P}_{\text{SNN}}^*(G, \vec{\mathbf{d}}; \mathbf{b}, \mathbf{B})} \mathcal{R}(\mathcal{R}_{[0,1]}(\Psi)) + \frac{18M \log \left(\left\lceil \frac{2BC_{\text{Lip}}^*}{\varepsilon} \right\rceil \right) + 73}{m} + 32\varepsilon \right) + \sqrt{\frac{\log(1/\delta)}{m}} \end{aligned} \quad (16)$$

with probability $1 - \delta$.

A proof of Theorem 22 is given in Appendix A.6.

In concluding this section, we briefly remark that there exists a substantial discourse on Lipschitz-continuous neural networks and their generalization properties, see e.g. (Béthune et al., 2022; Virmaux and Scaman, 2018; Gouk et al., 2021).

5. Expressivity of affine SNNs

It seems plausible that relying exclusively on excitatory responses (see Definition 3) will restrict the range of functions that can be approximated by affine SNNs. Nonetheless, we will see in this section that many approximation results of feedforward neural networks can still be reproduced by affine SNNs and even improved.

The core principle behind these expressivity results is the technical lemma below, whose proof is given in Appendix A.7.

Lemma 23 *Let $d_0 \in \mathbb{N}$ such that $d_0 \geq 2$, and let $\varepsilon > 0$. Then there exists an affine SNN $\Psi_\varepsilon^{\min} = (A_{\text{in}}, \Phi_\varepsilon^{\min}, A_{\text{out}})$, with $R(\Psi_\varepsilon^{\min}): \mathbb{R}^{d_0} \rightarrow \mathbb{R}$, such that for all $x_1, \dots, x_{d_0} \in \mathbb{R}$,*

$$|R(\Psi_\varepsilon^{\min})(x_1, \dots, x_{d_0}) - \min\{x_1, \dots, x_{d_0}\}| \leq \varepsilon. \quad (17)$$

Moreover, it holds that $\text{Size}(\Psi_\varepsilon^{\min}) = 2d_0 + 1$, the network graph has depth 1, all the weights in Ψ_ε^{\min} are bounded above in absolute value by $\max\{1, 1/\varepsilon\}$, and all the weights of Φ_ε^{\min} are bounded below by $1/\varepsilon$.

Remark 24 *Let us point out how remarkable the approximation of the minimum of Lemma 23 is in view of approximation by shallow ReLU neural networks. It can be seen from the proof of the lemma that the promised SNN has effectively one non-input neuron. It has been shown in (Safran et al., 2024, Theorem 4.3) that no ReLU neural network, regardless of its depth, can approximate the d -dimensional min operator with fewer than d neurons per hidden layer. Moreover, shallow feedforward neural networks, i.e., those with fewer than three layers cannot efficiently approximate the min operator with $\mathcal{O}(d)$ neurons.*

In the remainder of this section as well as the next, we continue to take $d_1 = 1$.

5.1 Universality

We start with the following essential lemma, which is a direct consequence of Lemma 23. Its proof is given in Appendix A.8.

Lemma 25 *Let $d_0 \in \mathbb{N}$, and let $\varepsilon > 0$. Let $a \in \mathbb{R}^{d_0}$, $b, c, d \in \mathbb{R}$. Then there exists an affine SNN $\Psi_{a,b,c,d,\varepsilon}$, with $R(\Psi_{a,b,c,d,\varepsilon}): \mathbb{R}^{d_0} \rightarrow \mathbb{R}$, such that for all $x \in \mathbb{R}^{d_0}$,*

$$|R(\Psi_{a,b,c,d,\varepsilon})(x) - c \max\{a^\top x + b, 0\} - d| \leq |c|\varepsilon. \quad (18)$$

Moreover, it holds that $\text{Size}(\Psi_{a,b,c,d,\varepsilon}) \leq d_0 + 5$, all the weights in $\Psi_{a,b,c,d,\varepsilon}$ are bounded above in absolute value by $\max\{1/\varepsilon, \|a\|_{\ell^\infty}, |b|, |c|, |d|\}$, and all the synaptic weights are bounded below by $1/\varepsilon$.

To highlight the significance of Lemma 25, consider setting $d_0 = 1$, $a = c = 1$, $b = d = 0$ in (18). This results in an affine SNN with constant size, whose realization is capable of approximating the ReLU function with an error of ε . An immediate consequence is that affine SNNs exhibit the same level of expressiveness as shallow ReLU feedforward neural networks. Specifically, it can be further observed that, for each $M \in \mathbb{N}$, the space

$$H_M := \left\{ x \mapsto \sum_{i=1}^M c_i \max\{a_i^\top x + b_i, 0\} + d_i : a_i \in \mathbb{R}^{d_0}, b_i, c_i, d_i \in \mathbb{R} \right\} \quad (19)$$

is contained in the closure of the set of realizations of affine SNNs. Since the finite sums of so-called *ridge functions* (19) are known to be universal approximators, see e.g. (Leshno et al., 1993), it implies that affine SNNs also share this property. We detail this in the theorem below, whose proof is given in Appendix A.9.

Theorem 26 *Let $d_0 \in \mathbb{N}$, and let $\varepsilon > 0$. Let $\Omega \subset \mathbb{R}^{d_0}$ be a compact domain. Then for every $f \in \mathcal{C}(\Omega)$, there exists an affine SNN Ψ^f such that $\sup_{x \in \Omega} |f(x) - \mathbf{R}(\Psi^f)(x)| \leq \varepsilon$.*

5.2 Emulation of finite element spaces

Lemma 23 confirms that affine SNNs are capable of approximating the minimum of multiple inputs to a given accuracy. This suggests the potential for effective approximation of *linear finite element spaces*, which comprise continuous piecewise affine functions. A similar capacity has already been documented in (He et al., 2020) for feedforward neural networks. We aim to translate these results to affine SNNs below.

For a compact domain $\Omega \subset \mathbb{R}^{d_0}$, a *finite element space* is based on a simplicial triangulation of Ω using simplices that we define below.

Definition 27 *Let $d_0 \in \mathbb{N}$, $n \in \mathbb{N}_0$ such that $n \leq d_0$. We call $x_0, \dots, x_n \in \mathbb{R}^{d_0}$ *affinely independent points* if and only if either $n = 0$ or $n \geq 1$, and the vectors $x_1 - x_0, \dots, x_n - x_0$ are linearly independent.*

An n -simplex is the convex hull of a set of $n + 1$ affinely independent points x_0, \dots, x_n , denoted as $\text{co}(x_0, \dots, x_n)$.

A simplicial triangulation of Ω is a partition of Ω into simplices.

Definition 28 *Let $d_0 \in \mathbb{N}$, and let $\Omega \subset \mathbb{R}^{d_0}$ be compact. Let $\mathcal{N} \subset \Omega$ be a finite set and let \mathcal{T} be a finite set of d_0 -simplices such that for each $\tau \in \mathcal{T}$, the set $N(\tau) := \mathcal{N} \cap \tau$ has cardinality $d_0 + 1$ and $\tau = \text{co}(N(\tau))$. We call \mathcal{T} a *regular triangulation* of Ω , if and only if*

1. $\bigcup_{\tau \in \mathcal{T}} \tau = \Omega$,
2. for all $\tau, \tau' \in \mathcal{T}$ it holds that $\tau \cap \tau' = \text{co}(N(\tau) \cap N(\tau'))$.

We call $\eta \in \mathcal{N}$ a *node*. Finally, we call

$$h_{\min}(\mathcal{T}) := \min_{\tau \in \mathcal{T}} \min_{\eta_1 \neq \eta_2 \in N(\tau)} |\eta_1 - \eta_2|,$$

$$h_{\max}(\mathcal{T}) := \max_{\tau \in \mathcal{T}} \max_{\eta_1, \eta_2 \in N(\tau)} |\eta_1 - \eta_2|,$$

the *min mesh-size* and the *max mesh-size* of \mathcal{T} , respectively.

For each $\eta \in \mathcal{N}$, we let $T(\eta) := \{\tau \in \mathcal{T} : \eta \in \tau\}$ be the set of *simplices that contain* η , and let $G(\eta) := \bigcup_{\tau \in T(\eta)} \tau$. Given a triangulation \mathcal{T} , the associated *linear finite element space* is defined to be

$$V_{\mathcal{T}} := \{f \in \mathcal{C}(\Omega) : f \text{ is affine on all } \tau \in \mathcal{T}\}.$$

Let ϕ_η be the unique function in $V_{\mathcal{T}}$ such that for all $\eta' \in \mathcal{N}$,

$$\phi_\eta(\eta') = \delta_{\eta, \eta'}. \quad (20)$$

It follows that $(\phi_\eta)_{\eta \in \mathcal{N}}$ is a basis for $V_{\mathcal{T}}$. In the case that $G(\eta)$ is convex, there exists a simplified formula for ϕ_η . Specifically, it is shown in (He et al., 2020, Lemma 3.1) that

$$\phi_\eta(x) = \max\{0, \min_{\tau \in T(\eta)} g_\tau(x)\} = \min_{\tau \in T(\eta)} g_\tau(x) - \min\{0, \min_{\tau \in T(\eta)} g_\tau(x)\}, \quad (21)$$

for all $x \in \Omega$, where g_τ is the unique globally affine function such that $g_\tau = \phi_\eta$ on τ . Thus, Lemma 23, coupled with (21), suggests the existence of moderately sized affine SNNs capable of approximating the functions ϕ_η with arbitrary precision, and subsequently, the elements in $V_{\mathcal{T}}$. This argument is formalized in Theorem 29 below, with its proof given in Appendix A.10. In what follows, a *convex regular triangulation* \mathcal{T} refers to a regular triangulation \mathcal{T} for which $G(\eta)$ is convex for each node η .

Theorem 29 *Let $d_0 \in \mathbb{N}$, and let $\Omega \subset \mathbb{R}^{d_0}$ be compact. Let \mathcal{T} be a convex regular triangulation of Ω with node set \mathcal{N} . Let $\varepsilon > 0$. Then for all $f \in V_{\mathcal{T}}$ there exists an affine SNN Ψ_ε^f such that for all $x \in \Omega$,*

$$|\mathbf{R}(\Psi_\varepsilon^f)(x) - f(x)| \leq \sum_{\eta \in \mathcal{N}} |f(\eta)| \varepsilon.$$

Moreover,

$$\text{Size}(\Psi_\varepsilon^f) \leq \sum_{\eta \in \mathcal{N}} (\#T(\eta)(d_0 + 2) + 6),$$

all the weights in Ψ_ε^f are bounded above in absolute value by

$$\max \left\{ 1, \|(f(\eta))_{\eta \in \mathcal{N}}\|_{\ell^\infty} / h_{\min}(\mathcal{T}), C d_0 \|(f(\eta))_{\eta \in \mathcal{N}}\|_{\ell^\infty} / h_{\min}(\mathcal{T}), 3/\varepsilon \right\}$$

for some $C = C(\Omega) > 0$, and all the synaptic weights are bounded below by $\min\{1, 3/\varepsilon\}$.

We conclude this subsection by highlighting another notable approximation result of affine SNNs for smooth functions. To articulate it, we need a geometric condition to hold for the compact domain Ω .

Definition 30 *We say that a compact domain $\Omega \subset \mathbb{R}^{d_0}$ is admissible if for every $N \in \mathbb{N}$, there exist a triangulation \mathcal{T}_N and universal constants $C_1, C_2, c_2 > 0$ such that $\#\mathcal{T}_N \leq C_1 N$ and that*

$$c_2 N^{-1/d_0} \leq h_{\min}(\mathcal{T}_N) \leq h_{\max}(\mathcal{T}_N) \leq C_2 N^{-1/d_0}. \quad (22)$$

The anticipated approximation result for smooth functions is as follows. We provide a proof in Appendix A.11.

Theorem 31 *Let $d_0 \in \mathbb{N}$, and let $\Omega \subset \mathbb{R}^{d_0}$ be an admissible domain. Let $s \in \{1, 2\}$. Then for every $f \in W^{s,\infty}(\Omega)$ and every $N \in \mathbb{N}$, there exist an affine SNN Ψ_N^f and a constant $C_1 = C_1(d_0)$ such that*

$$\|\mathbf{R}(\Psi_N^f) - f\|_{L^\infty(\Omega)} \leq C_1 N^{-s/d_0} \|f\|_{W^{s,\infty}(\Omega)}. \quad (23)$$

Moreover, $\text{Size}(\Psi_N^f) \leq C_2 N$, and all weights of Ψ_N^f are bounded above in absolute value by

$$\max \left\{ C_3 N^{1/d_0} \|f\|_{L^\infty(\Omega)}, 2N^{s/d_0+1} \right\}$$

and all synaptic weights are bounded below by $2N^{s/d_0+1}$, for some constants $C_2 = C_2(\Omega)$, $C_3 = C_3(\Omega)$.

5.3 Curse of dimensionality

It is well-known that sums of ReLU feedforward neural networks can overcome the curse of dimensionality when approximating specific functions of bounded variation (Barron, 1993; E and Wojtowytsch, 2022; Siegel and Xu, 2022; Caragea et al., 2023), called *Barron functions*. We recall the definition of these spaces below.

Definition 32 *Let $d_0 \in \mathbb{N}$. Let $K > 0$. The Barron class Γ_K with a constant K is defined to be the set of functions $f \in L^1_{\text{loc}}(\mathbb{R}^{d_0})$ for which there exists a measurable function \hat{f} such that, for all $x \in \mathbb{R}^{d_0}$*

$$f(x) = \int_{\mathbb{R}^{d_0}} e^{ix\xi} \hat{f}(\xi) d\xi, \quad \text{and} \quad \int_{\mathbb{R}^{d_0}} |\xi| |\hat{f}(\xi)| d\xi \leq K.$$

The central point of this subsection is to showcase that affine SNNs can approximate Barron functions with a dimension-independent rate. The formal result is as follows.

Theorem 33 *Let $d_0 \in \mathbb{N}$. There is a universal constant $\nu > 0$ such that the following holds. For every $K > 0$, every $f \in \Gamma_K$, and every $M \in \mathbb{N}$, there exists an affine SNN Ψ_M^f such that*

$$\sup_{x \in \overline{B(0,1)}} |\Psi_M^f(x) - f(x)| \leq \frac{\nu d_0^{\frac{1}{2}} K}{\sqrt{M}}. \quad (24)$$

Furthermore, there exist $C, c > 0$ such that $\text{Size}(\Psi_M^f) \leq C d_0 M$, all weights in Ψ_M^f can be bounded above by $C \cdot (M^{3/2}/\sqrt{K} + \sqrt{K})$, and all synaptic weights are bounded below by $cM^{3/2}/\sqrt{K}$.

A proof of Theorem 33 is given in Appendix A.12. It is a product of Lemmas 12, 25, and the known result that Barron functions can be uniformly approximated on compact domains without the curse of dimensionality by sums of ridge functions (Barron, 1992; Caragea et al., 2023).

6. Full error analysis

We recall from Section 4 the following parameterized hypothesis class

$$\mathcal{H} = \mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})).$$

The class encompasses clipped realizations of affine SNNs associated with a fixed network graph $G = (V, E)$. As defined in (10), (11), $\vec{d} = (d_0, d_1, d_{\text{in}}, d_{\text{out}}) = (d_0, 1, d_{\text{in}}, d_{\text{out}})$, \mathbf{B} denotes an upper bound for all the weight and delay parameters, and \mathbf{b} denotes a lower bound for the synaptic weights of the associated positive SNNs. Moreover, the class $\mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B}))$ was identified to be isometrically isomorphic to a compact domain in $[-\mathbf{B}, \mathbf{B}]^{\mathbf{M}}$, where the total Euclidean dimension $\mathbf{M} = d_{\text{in}}d_0 + 2\#E + d_1d_{\text{out}} = d_{\text{in}}d_0 + 2\#E + d_{\text{out}}$.

We first present the main learning theorem of this section, which will thereafter be applied to the approximation results of the previous section.

Theorem 34 *Let $d_0 \in \mathbb{N}$. Let $\kappa_{\mathbf{B}}, \kappa_{\mathbf{M}} > 0$. Let $f_0 : [0, 1]^{d_0} \rightarrow [0, 1]$, and suppose for every $\varepsilon \in (0, 1)$, there exists a hypothesis class of clipped realizations of affine SNNs,*

$$\mathcal{H}(\varepsilon) := \mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G(\varepsilon), \vec{d}(\varepsilon); \mathbf{b}(\varepsilon), \mathbf{B}(\varepsilon))), \quad (25)$$

such that

$$\inf \{ \|f_0 - g\|_{L^\infty([0,1]^{d_0})} : g \in \mathcal{H}(\varepsilon) \} \leq \varepsilon. \quad (26)$$

Here in (25), $\vec{d}(\varepsilon) := (d_0, 1, d_{\text{in}}(\varepsilon), d_{\text{out}}(\varepsilon))$, and $G(\varepsilon)$ is a network graph with $d_{\text{in}}(\varepsilon)$ input nodes and $d_{\text{out}}(\varepsilon)$ output nodes, such that the total dimension satisfies

$$\mathbf{M}(\varepsilon) \leq \varepsilon^{-\kappa_{\mathbf{M}}}. \quad (27)$$

Moreover, $\mathbf{b}(\varepsilon), \mathbf{B}(\varepsilon) \in \mathbb{R}$ are constrained by

$$\varepsilon^{-\kappa_{\mathbf{B}}} \geq \mathbf{B}(\varepsilon) \geq 1 \geq \mathbf{b}(\varepsilon) \geq \varepsilon^{\kappa_{\mathbf{B}}}, \quad (28)$$

Let \mathcal{D} be a distribution on $[0, 1]^{d_0}$. Let $S = (x_i, f_0(x_i))_{i=1}^m$ be a sample where $x_i \sim \mathcal{D}$ are i.i.d. Let $g_m \in \mathcal{H}(m^{-1/(\kappa_{\mathbf{M}}+2)})$ that is also an empirical risk minimizer based on the sample S . Then there exists a universal constant $c > 0$, such that the event

$$\mathcal{R}(g_m) \leq c \max \{1, \kappa_{\mathbf{B}}/\kappa_{\mathbf{M}}\} m^{-2/(\kappa_{\mathbf{M}}+2)} \log(m) + \sqrt{\frac{\log(1/\delta)}{m}} \quad (29)$$

holds with a probability at least $1 - \delta$.

A proof of Theorem 34 is given in Appendix A.13. It follows from the proof that the conclusion remains valid if the bounds $\varepsilon^{-\kappa_{\mathbf{B}}}, \varepsilon^{\kappa_{\mathbf{B}}}$ in (28) or $\varepsilon^{-\kappa_{\mathbf{M}}}$ in (27) are only assumed to hold up to a multiplicative constant, with the constant then affecting the final estimate.

Let us describe Theorem 34 in words. First, the condition (26) necessitates that the family of hypothesis classes $(\mathcal{H}(\varepsilon))_{\varepsilon \in (0,1)}$ possesses the ability to achieve uniform ε -approximation of f . Second, when (26) is met alongside the growth conditions on the

parameters (27), (28), the theorem asserts that learning can be accomplished with high probability. We want to emphasize that (26) holds when f belongs to a Sobolev class, as per Theorem 31, or a Barron class, as per Theorem 33. Therefore, below, we will provide the appropriate values for $d_{\text{in}}(\varepsilon)$, $d_{\text{out}}(\varepsilon)$, κ_{B} , κ_{M} for these two results and present the overall learning error estimate.

- **Sobolev functions.** Let $s \in \{1, 2\}$ and $\Omega \subset \mathbb{R}^{d_0}$ be a compact domain. Let $f \in W^{s,\infty}(\Omega)$ such that $\|f\|_{W^{s,\infty}(\Omega)} \leq 1$. Let $G(\varepsilon)$ be the network graph of the SNN of Theorem 31 with $d_{\text{in}}(\varepsilon) = d_{\text{out}}(\varepsilon) = N = \lceil \varepsilon^{-d_0/s} \rceil$. Let $\kappa_{\text{B}} = d_0/s + 1$, $\kappa_{\text{M}} = d_0/s$, and $\varepsilon = m^{-1/(\kappa_{\text{M}}+2)}$. Then Theorem 34 yields that the risk of the empirical risk minimizer over affine SNNs is asymptotically bounded by a constant multiple of

$$m^{-2/(d_0/s+2)} \log(m) + \sqrt{\frac{\log(1/\delta)}{m}},$$

with probability $1 - \delta$.

- **Barron functions.** Let $K > 0$. Let $f \in \Gamma_K$. Let $G(\varepsilon)$ the network graph of the SNNs of Theorem 33 with $d_{\text{in}}(\varepsilon) = d_{\text{out}}(\varepsilon) = \lceil \nu^2 d_0 K^2 \varepsilon^{-2} \rceil$. Let $\kappa_{\text{B}} = 3$, $\kappa_{\text{M}} = 2$, and $\varepsilon = m^{-1/(\kappa_{\text{M}}+2)}$. Then Theorem 34 yields that the risk of the empirical risk minimizer over affine SNNs is asymptotically bounded by a constant multiple of

$$m^{-1/2} \log(m) + \sqrt{\frac{\log(1/\delta)}{m}}, \quad (30)$$

with probability $1 - \delta$. It is worth noting that there is no dimension dependence in the exponent of m in (30). Therefore, the overall learning error bound can be seen to have overcome the curse of dimensionality.

7. Simulation results

We complement the mathematical results with simulations of affine SNNs in pyTorch for a series of machine learning tasks, unifying the derived theory with practical applications. For all experiments, the neural network architecture consists of an affine encoder, L layers of simple spiking neurons, and an affine decoder. Encoder and decoder were realized using standard linear layers in pyTorch, with weights and biases lower and upper bounded by $-B$ and B , respectively. In all experiments, we set $B = 10$. For the simple neurons, a custom module was implemented that solves for output spike times analytically, similar to related spiking neuron models (Mostafa, 2017; Comsa et al., 2020; Göltz et al., 2021),

$$t_v = \frac{1 + \sum_{u \in \mathcal{C}_v} \mathbf{w}_{(u,v)}(t_u + \theta_{(u,v)})}{\sum_{u \in \mathcal{C}_v} \mathbf{w}_{(u,v)}}, \quad (31)$$

where \mathcal{C}_v is a set containing the indices of all causal input spike times, i.e.,

$$\mathcal{C}_v := \{u : (u, v) \in E \text{ and } t_u + \theta_{(u,v)} \leq t_v\}.$$

For simplicity, we set all synaptic delays $\theta_{(u,v)}$ to 0 in simulations unless stated otherwise. For the spiking layers, weights $\mathbf{w}_{(u,v)}$ are lower and upper bounded by b and B . In pyTorch,

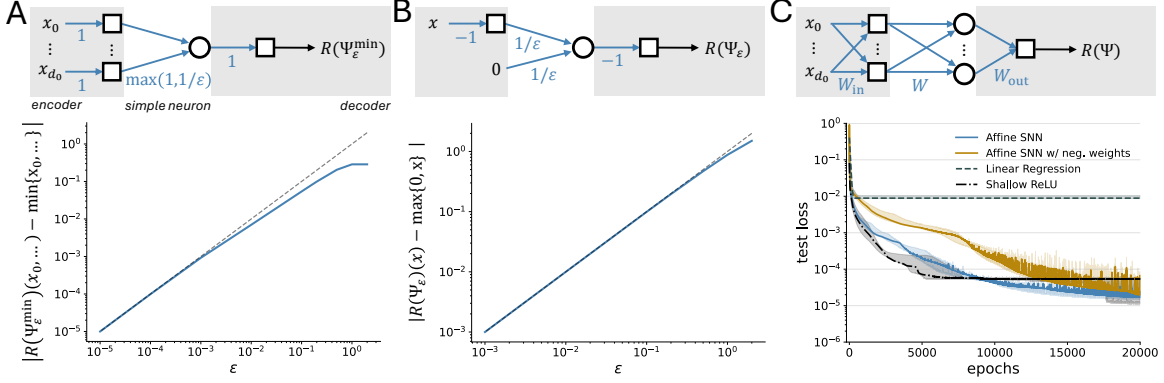


Figure 4: Affine SNN approximating **A** the min operator and **B** the max operator with an error (blue line) bounded by ε (dashed diagonal line). **C** An affine SNN is trained to reproduce a shallow ReLU neural network. As references, we also trained a linear neural network, a shallow ReLU neural network, an affine SNN in our model, and an affine SNN with both positive and negative synaptic weights. We show the median (line) and first and third quartiles (shaded area) over five different random seeds.

we realize this by applying the clamp function to the weights in each layer’s forward function. Since the encoder, output spike times, and decoder are all differentiable, the standard backward function of pyTorch was used to train affine SNNs. This allows the usage of exact gradients to train the SNN, as also done for related single-spike neural models (Mostafa, 2017; Comsa et al., 2020; Göltz et al., 2021). Alternatively, approximate methods based on SpikeProp (Bohte et al., 2000), such as surrogate gradients (Neftci et al., 2019), can be used.

First, we illustrate the results of Lemmas 23 and 25. In Figures 4A and B, we show that an affine SNN can be set up to approximate the min and max operator with an error upper bounded by ε . For the min operator, we use the identity function as encoder and decoder, and we set the weights to $w_{(u,v)} = \max(1, \frac{1}{\varepsilon})$. To evaluate the neural network, we randomly generate 1000 inputs from a uniform distribution (centered around 0), with $d_0 = 784$, $d_{\text{in}} = d_1 = d_{\text{out}} = 1$. For the max operator, we use the identity function multiplied by -1 as encoder and decoder, and we set the weights to $w_{(u,v)} = \frac{1}{\varepsilon}$. To evaluate it, we again generate 1000 random inputs from a uniform distribution (centered around 0), with $d_0 = d_{\text{in}} = d_1 = d_{\text{out}} = 1$.

We further show that an affine SNN is capable of approximating shallow ReLU neural networks. To do this, we initialized a random ReLU neural network with $d_0 = 40$ and $M = 20$ (see Lemma 25). The weights were initialized by sampling from a rescaled uniform distribution, i.e. $W_{ij} \sim \mathcal{U}(-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}})$ for $W_{ij} \in \mathbb{R}^{n \times m}$, which is the default initialization in pyTorch. From this ReLU neural network, we randomly sampled 10^4 training and 10^3 test examples. We trained an affine SNN ($d_0 = 40$, $d_{\text{in}} = 40$, $d_1 = 20$, $d_{\text{out}} = 1$, and $b = 0.01$) using a mean squared error loss, a learning rate of 10^{-3} , and L2 regularization with coefficient 10^{-5} . The whole experiment was repeated for five different random seeds.

In Figure 4C, the median test loss is shown. In addition, we show the test loss achieved by linear regression, an affine SNN with negative and positive weights, as well as a ReLU neural network with the same architecture as the neural network that was used to generate the training and test data. The affine SNN performs equally well as the ReLU neural network, with both learning significantly faster and more stably than an affine SNN with both positive and negative weights.

To demonstrate the generalization properties of affine SNNs, we trained three different network architectures on a simple regression task (fitting a quadratic function $f(x) = x/4 + x^2$): a 10-layer deep ReLU network with weights bounded by $-B$ and B and 100 neurons per hidden layer, an affine SNN ($d_{\text{in}} = 10, d_0 = 1, d_1 = 1, d_{\text{out}} = 100, L = 10, b = 0.1$) with 100 neurons per hidden layer and trainable delays bounded by $[0, B]$, and a shallow ReLU network with 30000 hidden neurons—chosen so that all three networks have approximately the same number of trainable parameters. To cover different severity levels of overfitting, we trained all networks on different numbers of training samples, from just 5 up to 10000. Testing was done on a separate set of 2000 samples, and we report the test loss obtained from the epoch with the lowest training loss, i.e., the one where the model fitted the training data best. In all cases, no weight decay or other regularization methods were used. Furthermore, we used a training rate of 10^{-3} (Adam optimizer), and a mean squared error loss function. In Figure 5A, we show the mean test loss and standard error for all three models after five runs with different random seeds. For a low number of training samples, both the affine SNN and shallow ReLU network outperform the deep ReLU network. However, for larger training datasets, the affine SNN starts outperforming both the shallow and deep ReLU networks.

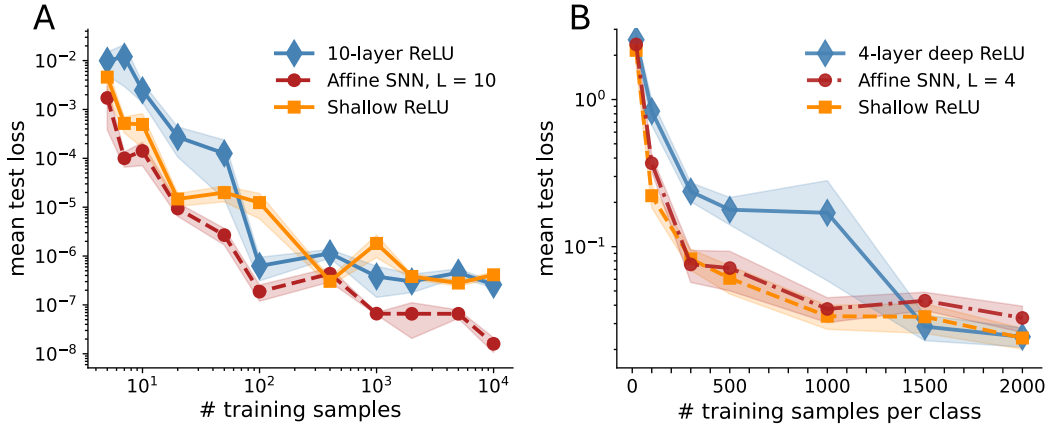


Figure 5: Comparison of affine SNNs and ReLU networks for different tasks. **A** Mean test loss (line) and standard error (shaded) for the regression task. The test loss is reported for the epoch with the lowest training loss (for a maximum of 2000 epochs). Simulations were repeated for five random seeds. **B** Same as A, but for a classification task (Yin-Yang (Kriener et al., 2022)) and a maximum of 10000 epochs.

We further show results for a classification task using the Yin-Yang dataset (Kriener et al., 2022), again for the same three models, but with different parameters: a 4-layer deep ReLU network with 100 neurons per hidden layer, an affine SNN ($d_{\text{in}} = 40, d_0 = 4, d_1 = 3, d_{\text{out}} = 100, L = 4, b = 0.1$) with 100 neurons per hidden layer and trainable delays bounded by $[0, B]$, and a shallow ReLU network with 4000 hidden neurons. In this experiment, we used a learning rate of 10^{-2} (Adam optimizer, batch size of 400), a cross-entropy loss function, and a test set with 1000 samples per class. The test loss of the affine SNN behaves similarly to the equivalent shallow ReLU neural network—although for very small and very large numbers of training samples, all networks perform similarly, with the deep ReLU neural network reaching the lowest test loss.

Finally, we trained an affine SNN ($d_0 = d_{\text{in}} = 784, d_1 = 200$, and $d_{\text{out}} = 10$) on the MNIST and Fashion MNIST task using a cross-entropy loss function, reaching similar performance levels as reported for deep feedforward neural networks in the literature: $96.75^{+0.02}_{-0.08}$ % median test accuracy for MNIST, with upper and lower index being the distance to the third and first quartile, and $87.81^{+0.04}_{-0.60}$ % for Fashion MNIST. The experiments were repeated five times with different random seeds.

The simulation code is available online³.

8. Discussion

In this study, we identify the fact that the weights in neural networks of simple spike-response neurons are allowed to be negative or arbitrarily close to zero as a reason why the parameterization of such SNNs through their weights is not continuous. This includes a larger family of spike models, such as the simple SRMs and variations of the current-based LIF models. We propose a solution to this shortcoming via a modified type of SNNs called affine SNNs, where the spiking neurons have exclusively positive weights, and show that it possesses remarkable approximation properties.

For mathematical tractability and to shift the focus of our analysis to the temporal aspect of SNNs, the simple SRM has been used. This model is, at the time of this study, not found in neuromorphic platforms (Frenkel et al., 2023), which at first glance limits the adaptability of our results to real-world applications. However, it is the limit case of the current-based LIF neuron model for large time constants and negligible leak, thus connecting to one of the currently most used spiking neuron models, both in simulations and neuromorphic implementations. This directly provides a neuromorphic implementation by modeling the neuron using a capacitance with a very large time constant, which features the required linear charging profile in the interval between the initial potential and the threshold potential. To ensure that spike times remain within a given experimental time window, appropriate regularization terms can be added during training. Thus, we are confident that our model can be reproduced in neuromorphic devices, either by approximating the used SRM through IF neurons, or by designing explicit circuits.

In general, our approach highlights the need to identify SNNs that depend continuously on their parameters to ensure stable training. Notably, recent work demonstrates that the Quadratic Integrate-and-Fire (QIF) model (Klos and Memmesheimer, 2025), which features

3. <https://github.com/dodo47/affineSNN>, commit abaa4c0

a self-amplification mechanism, i.e. a quadratic rise that emulates the action potentials of biological neurons, satisfies this condition.

Lastly, our work complements recent results on mapping the parameters of ReLU neural networks to simple SRM SNNs (Stanojevic et al., 2023, 2024), proving that such SNNs not only share many properties of feedforward ANNs, but also possess superior properties.

In conclusion, we believe that the construction presented in this work is a first step toward identifying spiking neuron models and SNN architectures that admit

1. continuous dependence on parameters,
2. no worse approximation performance than deep feedforward neural networks for relevant function classes,
3. better performance than deep feedforward neural networks in some tasks,
4. superior generalization performance (as evidenced by smaller covering numbers) over deep feedforward neural networks.

However, we also note that there are still many challenges ahead, especially in deriving such results for neuron models that are closer to biology or engineering applications (e.g., manufactured in neuromorphic devices), including features such as negative weights, synaptic response functions that are temporally bounded, alternative encodings (e.g. bursts (Payeur et al., 2021) and dendritic spikes (Larkum et al., 2022)) and adaptation mechanisms such as spike-frequency adaptation and short-term plasticity. Certainly, the function classes studied in this work are only a small subset of the set of relevant function classes. Naturally, the approximation properties for the classes $W^{s,\infty}$ for $s > 1$ should be understood next.

Acknowledgements and disclosure of funding

A.M.N. and P.C.P. were supported by the Austrian Science Fund (FWF) Project P-37010. D.D. was supported by the Horizon Europe’s Marie Skłodowska-Curie Actions (MSCA) Project 101103062 (BASE). The authors would like to thank M. Singh, A. Fono, and G. Kutyniok for enlightening discussions on the subject.

References

- Moshe Abeles. Role of the cortical neuron: Integrator or coincidence detector? *Israel journal of medical sciences*, 18(1):83–92, 1982.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, Cambridge, UK, 1999.
- Andrew R Barron. Neural net approximation. In *Proc. 7th Yale workshop on adaptive and learning systems*, volume 1, pages 69–72, 1992.
- Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*. MIT press, Cambridge, MA, USA, 2017.
- Julius Berner, Philipp Grohs, and Arnulf Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657, 2020.
- Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.
- Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto Gonzalez Sanz. Pay attention to your loss: Understanding misconceptions about Lipschitz neural networks. *Advances in Neural Information Processing Systems*, 35:20077–20091, 2022.
- Sander M Bohte, Joost N Kok, and Johannes A La Poutré. Spikeprop: backpropagation for networks of spiking neurons. In *ESANN*, volume 48, pages 419–424, 2000.
- Andrei Caragea, Philipp Petersen, and Felix Voigtlaender. Neural network approximation and estimation of classifiers with classification boundary in a Barron class. *The Annals of Applied Probability*, 33(4):3039–3079, 2023.
- Iulia M Comsa, Krzysztof Potempa, Luca Versari, Thomas Fischbacher, Andrea Gesmundo, and Jyrki Alakuijala. Temporal coding in spiking neural networks with alpha synaptic function. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8529–8533, 2020.
- Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, Cambridge, MA, USA, 2022.
- George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- Simon Davidson and Steve B Furber. Comparison of artificial and spiking neural networks on digital hardware. *Frontiers in Neuroscience*, 15:651141, 2021.
- Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, 2021.
- Dominik Dold, Josep Soler Garrido, Victor Caceres Chian, Marcel Hildebrandt, and Thomas Runkler. Neuro-symbolic computing with spiking neural networks. In *Proceedings of the International Conference on Neuromorphic Systems 2022*, pages 1–4, 2022.

- Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for Barron functions. *Calculus of Variations and Partial Differential Equations*, 61(2):1–37, 2022.
- Jason K Eshraghian, Max Ward, Emre O Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.
- Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445–466, 1961.
- Charlotte Frenkel, David Bol, and Giacomo Indiveri. Bottom-up and top-down approaches for the design of neuromorphic processing systems: Tradeoffs and synergies between natural and artificial intelligence. *Proceedings of the IEEE*, 111(6):623–652, 2023.
- Wulfram Gerstner. Time structure of the activity in neural network models. *Physical review E*, 51(1):738, 1995.
- Wulfram Gerstner and Richard Naud. How good are neuron models? *Science*, 326(5951):379–380, 2009.
- Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, Cambridge, UK, 2014.
- Julian Göltz, Laura Kriener, Andreas Baumbach, Sebastian Billaudelle, Oliver Breitwieser, Benjamin Cramer, Dominik Dold, Akos Ferenc Kungl, Walter Senn, Johannes Schemmel, et al. Fast and energy-efficient neuromorphic deep learning with first-spike times. *Nature machine intelligence*, 3(9):823–835, 2021.
- Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.
- Philipp Grohs and Felix Voigtlaender. Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces. *Foundations of Computational Mathematics*, 24(4):1085–1143, 2024.
- André Grüning and Sander M Bohte. Spiking neural networks: Principles and challenges. In *ESANN*, 2014.
- Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. ReLU deep neural networks and linear finite elements. *Journal of Computational Mathematics*, 38(3):502–527, 2020.
- Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560, 1990.
- Eugene M Izhikevich. Which model to use for cortical spiking neurons? *IEEE transactions on neural networks*, 15(5):1063–1070, 2004.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
- Saeed Reza Kheradpisheh and Timothée Masquelier. Temporal backpropagation for spiking neural networks with one spike per neuron. *International journal of neural systems*, 30(06):2050027, 2020.
- Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on learning theory*, pages 2306–2327, 2020.
- Christian Klos and Raoul-Martin Memmesheimer. Smooth exact gradient descent learning in spiking neural networks. *Physical Review Letters*, 134(2):027301, 2025.
- Laura Kriener, Julian Göltz, and Mihai A Petrovici. The yin-yang dataset. In *Proceedings of the 2022 Annual Neuro-Inspired Computational Elements Conference*, pages 107–111, 2022.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- L. Lapicque. Recherches quantitatives sur l’excitation électrique des nerfs traitée comme une variable physiologique. *Journal de Physiologie et de Pathologie Générale*, 9:620–635, 1907.
- Matthew E Larkum, Jiameng Wu, Sarah A Duverdin, and Albert Gidon. The guide to dendritic spikes of the mammalian cortex in vitro and in vivo. *Neuroscience*, 489:15–33, 2022.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- Andres Felipe Lerma-Pineda, Philipp Petersen, Simon Frieder, and Thomas Lukasiewicz. Dimension-independent learning rates for high-dimensional classification problems. *arXiv preprint arXiv:2409.17991*, 2024.
- Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

- Sasha Luccioni, Boris Gamazaychikov, Sara Hooker, Régis Pierrard, Emma Strubell, Yacine Jernite, and Carole-Jean Wu. Light bulbs have energy ratings—so why can’t AI chatbots? *Nature*, 632(8026):736–738, 2024a.
- Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of AI deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 85–99, 2024b.
- P Lunghi, S Silvestrini, G Meoni, D Dold, A Hadjiivanov, D Izzo, et al. Investigation of low-energy spiking neural networks based on temporal coding for scene classification. In *75th International Astronautical Congress (IAC 2024)*, pages 1–13, 2024.
- Wolfgang Maass. Lower bounds for the computational power of networks of spiking neurons. *Neural computation*, 8(1):1–40, 1996.
- Wolfgang Maass. Noisy spiking neurons with temporal coding have more computational power than sigmoidal neurons. *Advances in Neural Information Processing Systems*, 9: 211–217, 1997a.
- Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997b.
- Wolfgang Maass. To spike or not to spike: That is the question. *Proceedings of the IEEE*, 103(12):2219–2224, 2015.
- Wolfgang Maass and Michael Schmitt. On the complexity of learning for a spiking neuron. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 54–61, 1997.
- Wolfgang Maass and Michael Schmitt. On the complexity of learning for spiking neurons with temporal coding. *Information and Computation*, 153(1):26–46, 1999.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, Cambridge, MA, USA, 2018.
- Hesham Mostafa. Supervised learning based on temporal coding in spiking neural networks. *IEEE transactions on neural networks and learning systems*, 29(7):3227–3235, 2017.
- Jinichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.
- Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.
- Ricardo H Nochetto and Andreas Veiser. Primer of adaptive finite element methods. In *Multiscale and Adaptivity: Modeling, Numerics and Applications: CIME Summer School*, pages 125–225. 2011.
- OpenAI. GPT-4 technical report. *arXiv preprint 2303.08774*, 2023.

- Joost AA Opschoor, Philipp C Petersen, and Christoph Schwab. Deep ReLU networks and high-order finite element methods. *Analysis and Applications*, 18(05):715–770, 2020.
- Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2022.
- Hélène Paugam-Moisy. Spiking neuron networks a survey. *Rapport Technique RR-11, IDIAP*, 2006.
- Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, 24(7):1010–1019, 2021.
- Philipp Petersen and Jakob Zech. Mathematical theory of deep learning, 2024. URL <https://arxiv.org/abs/2407.18384>.
- Philipp Petersen, Mones Raslan, and Felix Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *Foundations of computational mathematics*, 21:375–444, 2021.
- Julian Rossbroich, Julia Gyax, and Friedemann Zenke. Fluctuation-driven initialization for spiking neural network training. *Neuromorphic Computing and Engineering*, 2(4):044016, 2022.
- Itay Safran, Daniel Reichman, and Paul Valiant. How many neurons does it take to approximate the maximum? In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3156–3183, 2024.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of statistics*, 48(4):1875–1897, 2020.
- Michael Schmitt. VC dimension bounds for networks of spiking neurons. In *ESANN*, pages 429–434, 1999.
- Catherine D Schuman, Thomas E Potok, Robert M Patton, J Douglas Birdwell, Mark E Dean, Garrett S Rose, and James S Plank. A survey of neuromorphic computing and neural networks in hardware. *arXiv preprint arXiv:1705.06963*, 2017.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, Cambridge, UK, 2014.
- Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.
- Jonathan W Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.
- Manjot Singh, Adalbert Fono, and Gitta Kutyniok. Expressivity of spiking neural networks. *arXiv preprint arXiv:2308.08218*, 2023.

- Ana Stanojevic, Stanisław Woźniak, Guillaume Bellec, Giovanni Cherubini, Angeliki Pantazi, and Wulfram Gerstner. An exact mapping from ReLU networks to spiking neural networks. *Neural Networks*, 168:74–88, 2023.
- Ana Stanojevic, Stanisław Woźniak, Guillaume Bellec, Giovanni Cherubini, Angeliki Pantazi, and Wulfram Gerstner. High-performance deep spiking neural networks with 0.3 spikes per neuron. *Nature Communications*, 15(1):6793, 2024.
- Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. Deep learning’s diminishing returns: The cost of improvement is becoming unsustainable. *IEEE Spectrum*, 58(10):50–55, 2021.
- Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.
- Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.
- Weizhen Xie, John H Wittig Jr, Julio I Chapeton, Mostafa El-Kalliny, Samantha N Jackson, Sara K Inati, and Kareem A Zaghloul. Neuronal sequences in population bursts encode information in human cortex. *Nature*, 635(8040):935–942, 2024.
- Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.
- Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10):905–913, 2021.
- Friedemann Zenke, Sander M Bohté, Claudia Clopath, Iulia M Comşa, Julian Göltz, Wolfgang Maass, Timothée Masquelier, Richard Naud, Emre O Neftci, Mihai A Petrovici, et al. Visualizing a joint future of neuroscience and neuromorphic engineering. *Neuron*, 109(4):571–575, 2021.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix A. Proofs

A.1 Proof of Lemma 5

Proof Fixing $v \in V \setminus V_{\text{in}}$, we first assume t_u exists whenever $(u, v) \in E$. Then the potential at v is given by (2)

$$P_v(t) = \sum_{(u,v) \in E} w_{(u,v)} \varrho(t - t_u - \theta_{(u,v)}),$$

for $t \in \mathbb{R}$, where, by Definition 2, at least one $w_{(u,v)} > 0$. It follows that P_v crosses the threshold value 1 at a unique time t_v . The proof can now be concluded via an induction pattern. Namely, since G is directed acyclic by assumption, there exists a topological ordering of the nodes, $(v_1, \dots, v_{\#V})$ such that $(v_i, v_j) \notin E$ for all $j \leq i$, see (Cormen et al., 2022, Section 20.4). Since the spike time is given for each input neuron, we deduce that P_v is well-defined for all $v \in V \setminus V_{\text{in}}$ by induction over the nodes in the topological order. \blacksquare

A.2 Proof of Theorem 13

We start with the following auxiliary lemma.

Lemma 35 *Let $\Phi = (G, W, D)$ be a positive SNN. Let $v \in V \setminus V_{\text{in}}$. Let $(t_u)_{(u,v) \in E}$, $(\tilde{t}_u)_{(u,v) \in E}$ denote two different vectors of spike times from all the presynaptic neurons u to v , and let $t_v((t_u)_{(u,v) \in E})$, $t_v((\tilde{t}_u)_{(u,v) \in E})$ denote the respective, corresponding spike times at v . Then it holds that*

$$|t_v((t_u)_{(u,v) \in E}) - t_v((\tilde{t}_u)_{(u,v) \in E})| \leq \|(t_u)_{(u,v) \in E} - (\tilde{t}_u)_{(u,v) \in E}\|_{\ell^\infty}. \quad (32)$$

Proof We explicitly indicate the dependence of the potential P_v on the received spike times $(t_u)_{(u,v) \in E}$ by writing

$$P_v(t; (t_u)_{(u,v) \in E}) = \sum_{(u,v) \in E} w_{(u,v)} \varrho(t - t_u - \theta_{(u,v)}).$$

Suppose $\tilde{t}_u \geq t_u$ for all $(u, v) \in E$. Then it is evident from the construction that P_v satisfies

$$P_v(t; (t_u)_{(u,v) \in E}) \geq P_v(t; (\tilde{t}_u)_{(u,v) \in E}). \quad (33)$$

Furthermore, by definition, $t_v((t_u)_{(u,v) \in E})$, $t_v((\tilde{t}_u)_{(u,v) \in E})$, are determined as the unique times at which

$$\begin{aligned} P_v(t_v((t_u)_{(u,v) \in E}); (t_u)_{(u,v) \in E}) &= 1, \\ P_v(t_v((\tilde{t}_u)_{(u,v) \in E}); (\tilde{t}_u)_{(u,v) \in E}) &= 1. \end{aligned} \quad (34)$$

Set $c = \|(t_u)_{(u,v) \in E} - (\tilde{t}_u)_{(u,v) \in E}\|_{\ell^\infty}$. It follows from the monotonicity depicted in (33) that,

$$\begin{aligned} P_v(t + c; (\tilde{t}_u)_{(u,v) \in E}) &= P_v(t; (\tilde{t}_u - c)_{(u,v) \in E}) \geq P_v(t; (t_u)_{(u,v) \in E}), \\ P_v(t - c; (\tilde{t}_u)_{(u,v) \in E}) &= P_v(t; (\tilde{t}_u + c)_{(u,v) \in E}) \leq P_v(t; (t_u)_{(u,v) \in E}). \end{aligned} \quad (35)$$

In particular, by substituting $t = t_v((t_u)_{(u,v) \in E})$ in (35), and applying (34), we obtain

$$\begin{aligned} P_v(t_v((t_u)_{(u,v) \in E}) - c; (\tilde{t}_u)_{(u,v) \in E}) &\leq 1 \\ &= P_v(t_v((\tilde{t}_u)_{(u,v) \in E}); (\tilde{t}_u)_{(u,v) \in E}) \\ &\leq P_v(t_v((t_u)_{(u,v) \in E}) + c; (\tilde{t}_u)_{(u,v) \in E}). \end{aligned}$$

By invoking the continuity of P_v with respect to time and the intermediate value theorem, along with the uniqueness of spike times, we obtain

$$t_v((\tilde{t}_u)_{(u,v) \in E}) \in [t_v((t_u)_{(u,v) \in E}) - c, t_v((t_u)_{(u,v) \in E}) + c],$$

which readily implies (32), when $\tilde{t}_u \geq t_u$ for all $(u, v) \in E$. For the general case, we define

$$(t'_u)_{(u,v) \in E} := (t_u + c)_{(u,v) \in E} \quad \text{and} \quad (t''_u)_{(u,v) \in E} := (t_u - c)_{(u,v) \in E}.$$

Evidently, $t''_u \leq t_u, \tilde{t}_u \leq t'_u$. Then by applying the previous reasoning, we deduce that

$$t_v((t''_u)_{(u,v) \in E}) \leq t_v((t_u)_{(u,v) \in E}), \quad t_v((\tilde{t}_u)_{(u,v) \in E}) \leq t_v((t'_u)_{(u,v) \in E}), \quad (36)$$

and that

$$\begin{aligned} |t_v((t_u)_{(u,v) \in E}) - t_v((t'_u)_{(u,v) \in E})| &\leq c, \\ |t_v((t_u)_{(u,v) \in E}) - t_v((t''_u)_{(u,v) \in E})| &\leq c. \end{aligned} \quad (37)$$

Hence, together, (36), (37) imply

$$|t_v((t_u)_{(u,v) \in E}) - t_v((\tilde{t}_u)_{(u,v) \in E})| \leq c = \|(t_u)_{(u,v) \in E} - (\tilde{t}_u)_{(u,v) \in E}\|_{\ell^\infty},$$

as desired. ■

The proof of Theorem 13 is a direct application of this lemma.

Proof [Proof of Theorem 13] Let $(t_u)_{u \in V_{\text{in}}}, (\tilde{t}_u)_{u \in V_{\text{in}}}$ be two tuples of input spike times for Φ . For each $v \in V \setminus V_{\text{in}}$, we denote by t_v, \tilde{t}_v , the corresponding spike times in response to $(t_u)_{u \in V_{\text{in}}}, (\tilde{t}_u)_{u \in V_{\text{in}}}$, respectively. Define $S(v) := \{u : (u, v) \in E\}$. Let $v_1 \in S(v)$ be such that $\|(t_u)_{u \in S(v)} - (\tilde{t}_u)_{u \in S(v)}\|_{\ell^\infty} = |t_{v_1} - \tilde{t}_{v_1}|$. Then it follows from Lemma 35 that

$$|t_v - \tilde{t}_v| \leq |t_{v_1} - \tilde{t}_{v_1}|.$$

Iterating this process, and letting $v_0 := v$, we move to select $v_{i+1} \in S(v_i)$ satisfying $\|(t_u)_{u \in S(v_i)} - (\tilde{t}_u)_{u \in S(v_i)}\|_{\ell^\infty} = |t_{v_{i+1}} - \tilde{t}_{v_{i+1}}|$. Note that, as G is a finite graph, this process yields a directed path of synaptic edges that starts at some $v_N \in V_{\text{in}}$ and ends in v_0 : (v_N, \dots, v_0) . Furthermore, for each $i = 0, \dots, N-1$, it holds that

$$|t_{v_i} - \tilde{t}_{v_i}| \leq |t_{v_{i+1}} - \tilde{t}_{v_{i+1}}|.$$

Therefore,

$$|t_v - \tilde{t}_v| = |t_{v_0} - \tilde{t}_{v_0}| \leq |t_{v_N} - \tilde{t}_{v_N}| \leq \|(t_u)_{u \in V_{\text{in}}} - (\tilde{t}_u)_{u \in V_{\text{in}}}\|_{\ell^\infty}.$$

Choosing $v \in V_{\text{out}}$ arbitrary, we conclude that

$$\begin{aligned} & \|R(\Phi)((t_u)_{u \in V_{\text{in}}}) - R(\Phi)((\tilde{t}_u)_{u \in V_{\text{in}}})\|_{\ell^\infty} \\ &= \|(t_v)_{v \in V_{\text{out}}} - (\tilde{t}_v)_{v \in V_{\text{out}}}\|_{\ell^\infty} \leq \|(t_u)_{u \in V_{\text{in}}} - (\tilde{t}_u)_{u \in V_{\text{in}}}\|_{\ell^\infty}. \end{aligned} \quad (38)$$

Next, let $(t_u)_{u \in V_{\text{in}}} = A_{\text{in}}(x)$, $(\tilde{t}_u)_{u \in V_{\text{in}}} = A_{\text{in}}(\tilde{x})$, for some $x, \tilde{x} \in \mathbb{R}^{d_0}$. Then

$$\begin{aligned} \|(t_u)_{u \in V_{\text{in}}} - (\tilde{t}_u)_{u \in V_{\text{in}}}\|_{\ell^\infty} &\leq \|(t_u)_{u \in V_{\text{in}}} - (\tilde{t}_u)_{u \in V_{\text{in}}}\|_{\ell^2} \\ &\leq \|W_{\text{in}}\|_F \|x - \tilde{x}\|_{\ell^2} \leq d_0^{\frac{1}{2}} \|W_{\text{in}}\|_F \|x - \tilde{x}\|_{\ell^\infty}. \end{aligned} \quad (39)$$

In a similar vein,

$$\begin{aligned} & \|A_{\text{out}}((t_v)_{v \in V_{\text{out}}}) - A_{\text{out}}((\tilde{t}_v)_{v \in V_{\text{out}}})\|_{\ell^\infty} \\ &\leq \|A_{\text{out}}((t_v)_{v \in V_{\text{out}}}) - A_{\text{out}}((\tilde{t}_v)_{v \in V_{\text{out}}})\|_{\ell^2} \\ &\leq \|W_{\text{out}}\|_F \|(t_v)_{v \in V_{\text{out}}} - (\tilde{t}_v)_{v \in V_{\text{out}}}\|_{\ell^2} \\ &\leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}}\|_F \|(t_v)_{v \in V_{\text{out}}} - (\tilde{t}_v)_{v \in V_{\text{out}}}\|_{\ell^\infty}. \end{aligned} \quad (40)$$

Combining (38), (39), (40), we arrive at the conclusion of the theorem. \blacksquare

A.3 Proof of Proposition 14

We begin by establishing a series of helper lemmas. For the forthcoming analysis, we recall that all SNNs considered share the same network graph.

Lemma 36 *Let $\Phi = (G, W, D)$, $\tilde{\Phi} = (G, W, \tilde{D})$ be two positive SNNs. Let $v \in V \setminus V_{\text{in}}$. Suppose D, \tilde{D} only differ in the synaptic delays from all the presynaptic neurons u to v , which are $(\theta_{(u,v)})_{(u,v) \in E}$, $(\tilde{\theta}_{(u,v)})_{(u,v) \in E}$, respectively. Let $t_v((\theta_{(u,v)})_{(u,v) \in E})$, $t_v((\tilde{\theta}_{(u,v)})_{(u,v) \in E})$ denote the respective, corresponding spike times at v in $\Phi, \tilde{\Phi}$. Then*

$$|t_v((\theta_{(u,v)})_{(u,v) \in E}) - t_v((\tilde{\theta}_{(u,v)})_{(u,v) \in E})| \leq \|(\theta_{(u,v)})_{(u,v) \in E} - (\tilde{\theta}_{(u,v)})_{(u,v) \in E}\|_{\ell^\infty}. \quad (41)$$

Proof An important observation to make is that, since D and \tilde{D} only vary in the synaptic delays from all presynaptic neurons u to v , the spike times at the neurons u with $(u, v) \in E$ remain identical for a given set of network input spike times $(t_u)_{u \in V_{\text{in}}}$. We denote these spike times coming to v as $(t_u)_{(u,v) \in E}$. The corresponding spike times at v , $t_v((\theta_{(u,v)})_{(u,v) \in E})$, $t_v((\tilde{\theta}_{(u,v)})_{(u,v) \in E})$, in $\Phi, \tilde{\Phi}$ are then the unique times at which

$$\begin{aligned} \sum_{(u,v) \in E} w_{(u,v)} \varrho(t_v((\theta_{(u,v)})_{(u,v) \in E}) - t_u - \theta_{(u,v)}) &= 1, \\ \sum_{(u,v) \in E} w_{(u,v)} \varrho(t_v((\tilde{\theta}_{(u,v)})_{(u,v) \in E}) - t_u - \tilde{\theta}_{(u,v)}) &= 1, \end{aligned}$$

respectively. It remains to notice that, from this point onward, by employing a similar argument to that used in the proof of Lemma 35, we can establish (41), as desired. \blacksquare

Lemma 37 *Let $\Phi = (G, W, D)$, $\tilde{\Phi} = (G, \tilde{W}, D)$ be two positive SNNs. Let $v \in V \setminus V_{\text{in}}$. Suppose that there exists $b > 0$, such that for every $w_{(u,v)} \in W$, $\tilde{w}_{(u,v)} \in \tilde{W}$,*

$$\min \{w_{(u,v)}, \tilde{w}_{(u,v)}\} \geq b. \quad (42)$$

Furthermore, suppose W, \tilde{W} only differ in the synaptic weights from all the presynaptic neurons u to v , which are $(w_{(u,v)})_{(u,v) \in E}$, $(\tilde{w}_{(u,v)})_{(u,v) \in E}$, respectively. Let $t_v((w_{(u,v)})_{(u,v) \in E})$, $t_v((\tilde{w}_{(u,v)})_{(u,v) \in E})$ denote the respective, corresponding spike times at v in $\Phi, \tilde{\Phi}$. Then

$$|t_v((w_{(u,v)})_{(u,v) \in E}) - t_v((\tilde{w}_{(u,v)})_{(u,v) \in E})| \leq \frac{\|(w_{(u,v)})_{(u,v) \in E} - (\tilde{w}_{(u,v)})_{(u,v) \in E}\|_{\ell^\infty}}{b^2}. \quad (43)$$

Proof As with the proof of Lemma 36, considering that W, \tilde{W} only differ in the synaptic weights from all the neurons u presynaptic to v , we conclude that the spike times at these neurons, denoted by $(t_u)_{(u,v) \in E}$, are the same for a given set of network input spike times $(t_u)_{u \in V_{\text{in}}}$. Without loss of generality, we assume $\theta_{(u,v)} = 0$ for all $(u,v) \in E$; otherwise, we can redefine t_u to $t_u + \theta_{(u,v)}$. Next, we denote $t^* := t_v((w_{(u,v)})_{(u,v) \in E})$, $\tilde{t}^* := t_v((\tilde{w}_{(u,v)})_{(u,v) \in E})$, the unique times at which

$$\sum_{(u,v) \in E} w_{(u,v)} \varrho(t^* - t_u) = 1 \quad \text{and} \quad \sum_{(u,v) \in E} \tilde{w}_{(u,v)} \varrho(\tilde{t}^* - t_u) = 1, \quad (44)$$

respectively. We order the presynaptic neurons according to their associated arrival times t_u ; namely, $\{u : (u,v) \in E\} = \{u_1, \dots, u_N\}$, where $t_{u_1} \leq \dots \leq t_{u_N}$, and rewrite (44) as

$$\sum_{i=1}^N w_{(u_i,v)} \varrho(t^* - t_{u_i}) = 1 \quad \text{and} \quad \sum_{i=1}^N \tilde{w}_{(u_i,v)} \varrho(\tilde{t}^* - t_{u_i}) = 1.$$

Suppose $\tilde{w}_{(u_i,v)} \geq w_{(u_i,v)}$, for all $i = 1, \dots, N$. It follows that $\tilde{t}^* \leq t^*$. Let $1 \leq J \leq N$ be the largest index for which $t^* \geq t_{u_J}$, and $1 \leq \tilde{J} \leq N$ the largest index for which $\tilde{t}^* \geq t_{u_{\tilde{J}}}$. Then $\tilde{J} \leq J$. Respectively, the times t^*, \tilde{t}^* are dictated by

$$\sum_{i=1}^J w_{(u_i,v)} (t^* - t_{u_i}) = 1 \quad \text{and} \quad \sum_{i=1}^{\tilde{J}} \tilde{w}_{(u_i,v)} (\tilde{t}^* - t_{u_i}) = 1. \quad (45)$$

Hence,

$$0 \leq \sum_{i=1}^{\tilde{J}} \tilde{w}_{(u_i,v)} (t^* - \tilde{t}^*) = \sum_{i=1}^{\tilde{J}} (\tilde{w}_{(u_i,v)} - w_{(u_i,v)}) (t^* - t_{u_i}) - \sum_{\tilde{J}+1}^J w_{(u_i,v)} (t^* - t_{u_i}),$$

which implies

$$\sum_{i=1}^{\tilde{J}} \tilde{w}_{(u_i,v)} |t^* - \tilde{t}^*| \leq \sum_{i=1}^{\tilde{J}} |\tilde{w}_{(u_i,v)} - w_{(u_i,v)}| |t^* - t_{u_i}|. \quad (46)$$

On the other hand, it follows from (45) that $\max\{|t^\star - t_{u_i}| : i = 1, \dots, \tilde{J}\} \leq \mathbf{b}^{-1}$, where \mathbf{b} is as in (42). Putting this back in (46), we deduce that

$$|t^\star - \tilde{t}^\star| \leq \frac{\sum_{i=1}^{\tilde{J}} |\tilde{\mathbf{w}}_{(u_i, v)} - \mathbf{w}_{(u_i, v)}|}{\mathbf{b} \sum_{i=1}^{\tilde{J}} \tilde{\mathbf{w}}_{(u_i, v)}} \leq \frac{\|(\mathbf{w}_{(u, v)})_{(u, v) \in E} - (\tilde{\mathbf{w}}_{(u, v)})_{(u, v) \in E}\|_{\ell^\infty}}{\mathbf{b}^2},$$

in the case $\tilde{\mathbf{w}}_{(u, v)} \geq \mathbf{w}_{(u, v)}$, for all $(u, v) \in E$, as wanted. For the general case, we proceed similarly to the proof of Lemma 35. Setting $c = \|(\mathbf{w}_{(u, v)})_{(u, v) \in E} - (\tilde{\mathbf{w}}_{(u, v)})_{(u, v) \in E}\|_{\ell^\infty}$, we define $\mathbf{w}'_{(u, v)} := \mathbf{w}_{(u, v)} + c$, $\mathbf{w}''_{(u, v)} := \mathbf{w}_{(u, v)} - c$. Then

$$\mathbf{w}''_{(u, v)} \leq \mathbf{w}_{(u, v)}, \tilde{\mathbf{w}}_{(u, v)} \leq \mathbf{w}'_{(u, v)}, \quad (47)$$

for all $(u, v) \in E$. Therefore

$$\begin{aligned} |t_v((\mathbf{w}_{(u, v)})_{(u, v) \in E}) - t_v((\mathbf{w}'_{(u, v)})_{(u, v) \in E})| &\leq \frac{c}{\mathbf{b}^2}, \\ |t_v((\mathbf{w}_{(u, v)})_{(u, v) \in E}) - t_v((\mathbf{w}''_{(u, v)})_{(u, v) \in E})| &\leq \frac{c}{\mathbf{b}^2}. \end{aligned} \quad (48)$$

For the final step, we combine (47), (48) to obtain (43). \blacksquare

We are approaching the proof of Proposition 14. Observe, although the inputs of $\mathbf{R}(\Phi)$ and $\mathbf{R}(\tilde{\Phi})$ are identical, represented by $t \in \mathbb{R}^{\text{d}_{\text{in}}}$, the spike times at $v \in V \setminus V_{\text{in}}$ will vary in both Φ and $\tilde{\Phi}$ due to their different network parameters, respectively (\mathbf{W}, D) and $(\tilde{\mathbf{W}}, \tilde{D})$. To effectively monitor the changes in spike times as they propagate through Φ and $\tilde{\Phi}$, we employ a graph splitting algorithm that enables us to partition a network graph into a finite sequence of disjoint subgraphs, each having depth 1. The precise statement is provided below.

Lemma 38 *Let $G = (V, E)$ be a network graph with depth \mathbf{L} . Then there exists a finite sequence of network subgraphs $(G^i)_{i=1}^{\mathbf{L}}$ of G , such that the following hold:*

- (i) $E(G) = \bigcup_{i=1}^{\mathbf{L}} E(G^i)$;
- (ii) $E(G^i) \cap E(G^j) = \emptyset$ if $i \neq j$;
- (iii) the depth of G^i is 1;
- (iv) if v is an output node in G^i , then all the synapses $(u, v) \in E(G)$ are included in $E(G^i)$;
- (v) every input node u in G^i is an input node of the network subgraph of G comprising all the edges in $\bigcup_{j=i}^{\mathbf{L}} E(G^j)$.

Proof First, we partition the vertex set $V = \bigcup_{i=0}^{\mathbf{L}} V^i$ following an algorithm akin to the longest path layering algorithm, which we now describe. Define V^i to be the collection of vertices whose longest path from an input node has length i . Since G is directed acyclic, V^i is well-defined. Evidently, $V^0 = V_{\text{in}}$, and i can reach up to \mathbf{L} , where $V^{\mathbf{L}}$ comprises output

Figure 6: A network graph G with four input nodes in burgundy and three output nodes in violet

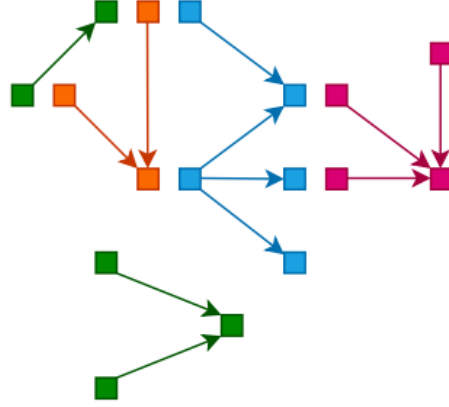
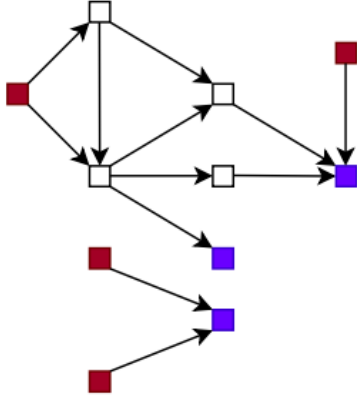


Figure 7: The network graph G split into four network subgraphs with depth of 1: G^1 in green, G^2 in orange, G^3 in light blue and G^4 in pink

nodes whose longest directed path to an input node equals the graph depth L . Moreover, there are only directed edges from vertices in V^i to vertices in V^j , if $i < j$. With $(V^i)_{i=0}^L$ established, we construct G^i , $i = 1, \dots, L$, to be the subgraph of G comprising all incoming edges to vertices in V^i . We refer to Figures 6, 7 for a visualization of this graph splitting process. Then the conditions (i), (ii), (iii), (iv) follow directly from the construction. To see (v), we fix an index i and an input node v in G^i . By way of construction, $v \in V^k$, for some $k < i$. Let \mathcal{G}^i denote the network subgraph comprising all the edges in $\bigcup_{j=i}^L E(G^j)$. Supposing that v is not an input node of \mathcal{G}^i , we can obtain an edge (u, v) in \mathcal{G}^i . Then it must be that $(u, v) \in E(G^k)$, posing a contradiction to (ii). ■

With Lemma 38 established, we advance to the proof of Proposition 14.

Proof [Proof of Proposition 14] Consider the decomposition of G into G^1, G^2, \dots, G^L according to Lemma 38. Let W^i, \tilde{W}^i be the respective restrictions of W, \tilde{W} onto $E(G^i)$, and D^i, \tilde{D}^i be the respective restrictions of D, \tilde{D} onto $E(G^i)$. We construct the following $2L$ positive SNNs,

$$\Phi^i = (G^i, W^i, D^i) \quad \text{and} \quad \tilde{\Phi}^i = (G^i, \tilde{W}^i, \tilde{D}^i),$$

for $i = 1, \dots, L$. Let V_{in}^i be the set of input nodes of G^i and V_{out}^i be the corresponding set of output nodes. Let $t \in \mathbb{R}^{\text{din}}$ be a tuple of input spike times for the positive SNNs $\Phi, \tilde{\Phi}$. For each $i = 1, \dots, L$, we write $t_{V_{\text{in}}^i}, \tilde{t}_{V_{\text{in}}^i}$, to denote the input tuples to $\Phi^i, \tilde{\Phi}^i$, respectively, and $t_{V_{\text{out}}^i}, \tilde{t}_{V_{\text{out}}^i}$ to denote the corresponding output tuples. Note that these tuples depend

on the given input tuple $t \in \mathbb{R}^{\text{din}}$. Further,

$$\begin{aligned} & \left\| \mathbf{R}(\Phi^i)(t_{V_{\text{in}}^i}) - \mathbf{R}(\tilde{\Phi}^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty} \\ & \leq \left\| \mathbf{R}(\Phi^i)(t_{V_{\text{in}}^i}) - \mathbf{R}(\Phi^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty} + \left\| \mathbf{R}(\Phi^i)(\tilde{t}_{V_{\text{in}}^i}) - \mathbf{R}(\tilde{\Phi}^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty}. \end{aligned} \quad (49)$$

With the graph depth of G^i being 1 (by Lemma 38), we can apply Lemmas 36, 37 to the second term on the RHS of (49), resulting in

$$\left\| \mathbf{R}(\Phi^i)(\tilde{t}_{V_{\text{in}}^i}) - \mathbf{R}(\tilde{\Phi}^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty} \leq \left(1 + \frac{1}{b^2}\right) \|\Phi^i - \tilde{\Phi}^i\|_{\ell^\infty}. \quad (50)$$

For the first term on the RHS of (49), we use Lemma 35 to obtain

$$\left\| \mathbf{R}(\Phi^i)(t_{V_{\text{in}}^i}) - \mathbf{R}(\Phi^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty} \leq \|t_{V_{\text{in}}^i} - \tilde{t}_{V_{\text{in}}^i}\|_{\ell^\infty}. \quad (51)$$

By construction, an input node of G^i either originates from an output node of G^j , for some $j < i$, or an input node of G (see for instance, the graph G^4 depicted in Figure 7). Subsequently, each component in $t_{V_{\text{in}}^i}$ is either a component in $t_{V_{\text{out}}^j}$, or a component in the network input t . A similar assertion applies to the corresponding components in $\tilde{t}_{V_{\text{in}}^i}$. We can then deduce that the RHS of (51) is bounded above by

$$\left\| t_{V_{\text{out}}^j} - \tilde{t}_{V_{\text{out}}^j} \right\|_{\ell^\infty} = \left\| \mathbf{R}(\Phi^{i-1})(t_{V_{\text{in}}^j}) - \mathbf{R}(\tilde{\Phi}^{i-1})(\tilde{t}_{V_{\text{in}}^j}) \right\|_{\ell^\infty},$$

for some $j < i$. However, this quantity resembles the LHS of (49). Hence, by replicating the provided reasoning inductively, and combining (49), (50), we derive

$$\begin{aligned} \left\| t_{V_{\text{out}}^i} - \tilde{t}_{V_{\text{out}}^i} \right\|_{\ell^\infty} &= \left\| \mathbf{R}(\Phi^i)(t_{V_{\text{in}}^i}) - \mathbf{R}(\tilde{\Phi}^i)(\tilde{t}_{V_{\text{in}}^i}) \right\|_{\ell^\infty} \\ &\leq \sum_{k=1}^i \left(1 + \frac{1}{b^2}\right) \|\Phi^k - \tilde{\Phi}^k\|_{\ell^\infty}. \end{aligned} \quad (52)$$

Let $v \in V_{\text{out}}$ be an output node of G . Then v is an output node in one of the G^i , $i = 1, \dots, L$. By Lemma 38, the edge sets $E(G^i)$ are mutually disjoint, and collectively, they exhaust the entire edge set $E(G)$. Therefore, from (52) we can infer

$$|t_v - \tilde{t}_v| \leq \sum_{i=1}^L \left(1 + \frac{1}{b^2}\right) \|\Phi^i - \tilde{\Phi}^i\|_{\ell^\infty} \leq L \cdot \left(1 + \frac{1}{b^2}\right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty}, \quad (53)$$

where $t_v = t_v(t)$, $\tilde{t}_v = \tilde{t}_v(t)$ denote the output spike times at v in Φ , $\tilde{\Phi}$, respectively, given $t \in \mathbb{R}^{\text{din}}$. The conclusion (4) now follows from (53), completing the proof. \blacksquare

A.4 Proof of Theorem 18

In the subsequent discussion, the dimension of the zero vector $\vec{0}$ can be inferred from the context. As usual, we start with a necessary lemma.

Lemma 39 *Let $\Phi = (G, W, D)$ be a positive SNN. Let $\mathbf{b}, \mathbf{B} \in (0, \infty)$ be such that for all $(u, v) \in E$, $w_{(u,v)} \geq \mathbf{b}$ and $\theta_{(u,v)} \leq \mathbf{B}$. Let \mathbf{L} be the graph depth of G . Then*

$$\|\mathbf{R}(\Phi)(\vec{0})\|_{\ell^\infty} \leq \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right). \quad (54)$$

Proof To begin, we consider the case where $\theta_{(u,v)} = 0$ for all $(u, v) \in E$, and we refer to this specific SNN as Φ_0 . We make the following observation. Let $v \in V \setminus V_{\text{in}}$. Suppose $t_u = 0$ for all $(u, v) \in E$, and let t_v be the corresponding spike time. Then by definition,

$$P_v(t_v) = \sum_{(u,v) \in E} w_{(u,v)} \varrho(t_v) = 1,$$

from which we obtain

$$t_v = \frac{1}{\sum_{(u,v) \in E} w_{(u,v)}}. \quad (55)$$

Using this, we proceed to derive (54) for $\Phi = \Phi_0$. For each $u \in V \setminus V_{\text{in}}$, let t_u be the corresponding spike time at u when the input spike times of Φ_0 are set to $\vec{0}$. Let $v \in V_{\text{out}}$. Then we can also interpret $t_v = t_v((t_u)_{(u,v) \in E})$. Therefore, an application of Lemma 35 and the base case result (55) grants us

$$\begin{aligned} |t_v((t_u)_{(u,v) \in E})| &\leq |t_v(\vec{0})| + |t_v((t_u)_{(u,v) \in E}) - t_v(\vec{0})| \\ &\leq \frac{1}{\sum_{(u,v) \in E} w_{(u,v)}} + \|(t_u)_{(u,v) \in E}\|_{\ell^\infty}. \end{aligned} \quad (56)$$

Continuing as in the proof of Theorem 13, we can find a directed path of synaptic edges, (v_N, \dots, v_0) , such that $v_0 = v$, $v_N \in V_{\text{in}}$, and for each $i = 0, \dots, N-1$, it holds that

$$\|(t_u)_{(u,v_i) \in E}\|_{\ell^\infty} = |t_{v_{i+1}}|.$$

By inductively employing the reasoning in (56) along this path, bearing in mind that the input spike times of Φ_0 are $\vec{0}$, we acquire

$$|t_{v_0}| = |t_v| \leq \sum_{i=0}^{N-1} \frac{1}{\sum_{(u,v_i) \in E} w_{(u,v_i)}} \leq \frac{\mathbf{L}}{\mathbf{b}},$$

which in turn implies

$$\|\mathbf{R}(\Phi_0)(\vec{0})\|_{\ell^\infty} \leq \frac{\mathbf{L}}{\mathbf{b}}. \quad (57)$$

Subsequently, in the case of general Φ , it can be deduced from (57) and the argument presented in the proof of Proposition 14 (with the synaptic weights coinciding) that

$$\begin{aligned} \|\mathbf{R}(\Phi)(\vec{0})\|_{\ell^\infty} &\leq \|\mathbf{R}(\Phi_0)(\vec{0})\|_{\ell^\infty} + \|\mathbf{R}(\Phi)(\vec{0}) - \mathbf{R}(\Phi_0)(\vec{0})\|_{\ell^\infty} \\ &\leq \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \|\Phi - \Phi_0\|_{\ell^\infty} \right) \leq \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right), \end{aligned}$$

as wanted. \blacksquare

Proof [Proof of Theorem 18] Let $x \in \mathbb{R}^{d_0}$. Then the inputs of Φ and $\tilde{\Phi}$ are $t = (t_u)_{u \in V_{\text{in}}} = A_{\text{in}}(x)$ and $\tilde{t} = (\tilde{t}_u)_{u \in V_{\text{in}}} = \tilde{A}_{\text{in}}(x)$, respectively, where

$$A_{\text{in}}(x) = W_{\text{in}}x + b_{\text{in}} \quad \text{and} \quad \tilde{A}_{\text{in}}(x) = \widetilde{W}_{\text{in}}x + \tilde{b}_{\text{in}}. \quad (58)$$

It follows from (58) and the estimate (38), as well as Proposition 14, that

$$\begin{aligned} \|\mathbf{R}(\Phi)(t) - \mathbf{R}(\tilde{\Phi})(\tilde{t})\|_{\ell^\infty} &\leq \|\mathbf{R}(\Phi)(t) - \mathbf{R}(\Phi)(\tilde{t})\|_{\ell^\infty} + \|\mathbf{R}(\Phi)(\tilde{t}) - \mathbf{R}(\tilde{\Phi})(\tilde{t})\|_{\ell^\infty} \\ &\leq \|t - \tilde{t}\|_{\ell^\infty} + \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty} \\ &\leq \|W_{\text{in}} - \widetilde{W}_{\text{in}}\|_F \|x\|_{\ell^2} + \|b_{\text{in}} - \tilde{b}_{\text{in}}\|_{\ell^\infty} + \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty} \\ &\leq d_0^{\frac{1}{2}} \|W_{\text{in}} - \widetilde{W}_{\text{in}}\|_F \|x\|_{\ell^\infty} + \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty} + \|b_{\text{in}} - \tilde{b}_{\text{in}}\|_{\ell^\infty} \end{aligned} \quad (59)$$

Denote $z = \mathbf{R}(\Phi)(t)$, $\tilde{z} = \mathbf{R}(\tilde{\Phi})(\tilde{t})$. We can write

$$\begin{aligned} \|\mathbf{R}(\Psi)(x) - \mathbf{R}(\tilde{\Psi})(x)\|_{\ell^\infty} &= \|A_{\text{out}}(z) - \tilde{A}_{\text{out}}(\tilde{z})\|_{\ell^\infty} \\ &\leq \|A_{\text{out}}(z) - A_{\text{out}}(\tilde{z})\|_{\ell^\infty} + \|A_{\text{out}}(\tilde{z}) - \tilde{A}_{\text{out}}(\tilde{z})\|_{\ell^\infty} \\ &= T_1 + T_2, \end{aligned} \quad (60)$$

where $T_1 := \|A_{\text{out}}(z) - A_{\text{out}}(\tilde{z})\|_{\ell^\infty}$, $T_2 := \|A_{\text{out}}(\tilde{z}) - \tilde{A}_{\text{out}}(\tilde{z})\|_{\ell^\infty}$. Applying (59) to T_1 , we get

$$T_1 \leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}}\|_F \left(d_0^{\frac{1}{2}} \|W_{\text{in}} - \widetilde{W}_{\text{in}}\|_F \|x\|_{\ell^\infty} + \mathbf{L} \cdot \left(1 + \frac{1}{\mathbf{b}^2} \right) \|\Phi - \tilde{\Phi}\|_{\ell^\infty} + \|b_{\text{in}} - \tilde{b}_{\text{in}}\|_{\ell^\infty} \right).$$

Further, given

$$A_{\text{out}}(x) = W_{\text{out}}x + b_{\text{out}} \quad \text{and} \quad \tilde{A}_{\text{out}}(x) = \widetilde{W}_{\text{out}}x + \tilde{b}_{\text{out}},$$

we can also similarly obtain

$$\begin{aligned} T_2 &\leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}} - \widetilde{W}_{\text{out}}\|_F \|\tilde{z}\|_{\ell^\infty} + \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty} \\ &\leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}} - \widetilde{W}_{\text{out}}\|_F \left(\|\mathbf{R}(\tilde{\Phi})(\tilde{t}) - \mathbf{R}(\tilde{\Phi})(\vec{0})\|_{\ell^\infty} + \|\mathbf{R}(\tilde{\Phi})(\vec{0})\|_{\ell^\infty} \right) + \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty} \\ &\leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}} - \widetilde{W}_{\text{out}}\|_F \left(\|\tilde{t}\|_{\ell^\infty} + \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right) \right) + \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty} \\ &\leq d_{\text{out}}^{\frac{1}{2}} \|W_{\text{out}} - \widetilde{W}_{\text{out}}\|_F \left(d_0^{\frac{1}{2}} \|\widetilde{W}_{\text{in}}\|_F \|x\|_{\ell^\infty} + \|\tilde{b}_{\text{in}}\|_{\ell^\infty} + \mathbf{L} \cdot \left(\frac{1}{\mathbf{b}} + \mathbf{B} \right) \right) + \|b_{\text{out}} - \tilde{b}_{\text{out}}\|_{\ell^\infty}, \end{aligned}$$

where we have used (38) and Lemma 39 in the third inequality above. Substituting these estimates back into (60), we acquire the desired conclusion for the theorem. \blacksquare

A.5 Proof of Theorem 21

A standard learning result employing covering numbers is as follows.

Theorem 40 ((Mohri et al., 2018, Exercise 3.31)) *Let $d_0, m \in \mathbb{N}$. Let \mathcal{H} be a hypothesis set of functions from $[0, 1]^{d_0}$ to $[c_{\min}, c_{\max}]$, where $c_{\min}, c_{\max} \in \mathbb{R}$. Let \mathcal{D} be a distribution on $[0, 1]^{d_0} \times [c_{\min}, c_{\max}]$. Let, for $m \in \mathbb{N}$, $S \sim \mathcal{D}^m$ be a sample. Then, for all $\varepsilon > 0$*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |\mathcal{R}(h) - \widehat{\mathcal{R}}_S(h)| \geq \varepsilon\right) \leq 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8(c_{\max} - c_{\min})}, L^\infty([0, 1]^{d_0})\right) \exp\left(\frac{-m\varepsilon^2}{2(c_{\max} - c_{\min})^4}\right).$$

We show below that Theorem 40 and (13) imply Theorem 21.

Proof [Proof of Theorem 21] Set, for $m \in \mathbb{N}$ satisfying (14),

$$\varepsilon(\delta) := \sqrt{\frac{2(\mathbf{M} \log(m \lceil 16\mathbf{B}C_{\text{Lip}}^* \rceil) + \log(2/\delta))}{m}} \leq 1,$$

then by Theorem 40 the statement of this theorem follows if

$$2\mathcal{N}\left(\mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \frac{\varepsilon(\delta)}{8}, L^\infty([0, 1]^{d_0})\right) \exp\left(\frac{-m\varepsilon(\delta)^2}{2}\right) \leq \delta.$$

Since, by definition

$$\exp\left(\frac{-m\varepsilon(\delta)^2}{2}\right) = \frac{\delta}{2} \exp\left(-\mathbf{M} \log\left(m \lceil 16\mathbf{B}C_{\text{Lip}}^* \rceil\right)\right),$$

we can conclude the theorem if

$$\mathcal{N}\left(\mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \frac{\varepsilon(\delta)}{8}, L^\infty([0, 1]^{d_0})\right) \leq \exp\left(\mathbf{M} \log\left(m \lceil 16\mathbf{B}C_{\text{Lip}}^* \rceil\right)\right). \quad (61)$$

Let us have a look at the left-hand side of (61). Building upon the rationale given in (13), it holds that

$$\begin{aligned} \log\left(\mathcal{N}\left(\mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \frac{\varepsilon(\delta)}{8}, L^\infty([0, 1]^{d_0})\right)\right) &\leq \mathbf{M} \log\left(\left\lceil \frac{16\mathbf{B}C_{\text{Lip}}^*}{\varepsilon(\delta)} \right\rceil\right) \\ &\leq \mathbf{M} \log\left(m \lceil 16\mathbf{B}C_{\text{Lip}}^* \rceil\right), \end{aligned} \quad (62)$$

where we have used that $\lceil 1/(\varepsilon(\delta)) \rceil \leq \lceil \sqrt{m} \rceil \leq m$. Exponentiating (62) yields (61) and thus completes the proof. \blacksquare

A.6 Proof of Theorem 22

Proof Let Ψ_m satisfy (15). By setting $\delta = \varepsilon$ in the statement of (Schmidt-Hieber, 2020, Lemma 4), it follows directly that, for $\varepsilon \in (0, 1]$,

$$\begin{aligned} &\mathbb{E}(\widehat{\mathcal{R}}_S(\mathcal{R}_{[0,1]}(\Psi_m))) \\ &\leq 4 \left(\inf_{\Psi \in \mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})} \mathcal{R}(\mathcal{R}_{[0,1]}(\Psi)) + \frac{18\mathcal{N}(\mathcal{R}_{[0,1]}(\mathcal{P}_{\text{SNN}}^*(G, \vec{d}; \mathbf{b}, \mathbf{B})), \varepsilon, L^\infty(\mathbb{R}^{d_0})) + 73}{m} + 32\varepsilon \right). \end{aligned}$$

Subsequently, invoking (13) yields

$$\begin{aligned} & \mathbb{E}(\hat{\mathcal{R}}_S(\mathbf{R}_{[0,1]}(\Psi_m))) \\ & \leq 4 \left(\inf_{\Psi \in \mathcal{P}_{\text{SNN}}^*(G, \vec{\mathbf{d}}; \mathbf{b}, \mathbf{B})} \mathcal{R}(\mathbf{R}_{[0,1]}(\Psi)) + \frac{18M \log \left(\left\lceil \frac{2\text{BC}_{\text{Lip}}^*}{\varepsilon} \right\rceil \right) + 73}{m} + 32\varepsilon \right). \end{aligned} \quad (63)$$

Now, it is straightforward to verify that the function

$$g(z_1, \dots, z_m) := \frac{1}{m} \sum_{i=1}^m |\mathbf{R}_{[0,1]}(\Psi_m)(z_i) - f_0(z_i)|^2$$

is of bounded variation, with

$$|g(z_1, \dots, z_i, \dots, z_m) - g(z_1, \dots, z'_i, \dots, z_m)| \leq \frac{\sqrt{2}}{m}. \quad (64)$$

Thus, applying McDiarmid's inequality ((Mohri et al., 2018, Theorem D.3)) to g , and combining with (63), (64), we obtain the bound in (16). \blacksquare

A.7 Proof of Lemma 23

Proof We begin by constructing $\Phi_\varepsilon^{\min} = (G, \mathbf{W}_\varepsilon^{\min}, D_\varepsilon^{\min})$ as follows. We take G to be a graph with d_0 input nodes u_1, \dots, u_{d_0} and one output node v . Let $\mathbf{W}_\varepsilon^{\min}$ consist of $w_{(u_i, v)} = \varepsilon^{-1}$ for $i = 1, \dots, d_0$. Let all the delays be zero, i.e. $\theta_{(u_i, v)} = 0$, for $i = 1, \dots, d_0$. Finally, we formalize the structure of Ψ_ε^{\min} by defining $A_{\text{in}} : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_0}$, $A_{\text{out}} : \mathbb{R} \rightarrow \mathbb{R}$ as $A_{\text{in}} = \text{Id}^{d_0 \times d_0}$, $A_{\text{out}} = \text{Id}^{1 \times 1}$, respectively.

To show that Ψ_ε^{\min} fulfills (17), we first calculate the output spike time t_v by Φ_ε^{\min} from the input spike times t_{u_i} . Suppose, without loss of generality, that

$$t_{u_1} = \min\{t_{u_i} : i = 1, \dots, d_0\}.$$

The potential at v then takes the form (2)

$$P_v(t) = \sum_{i=1}^{d_0} \varepsilon^{-1} \varrho(t - t_{u_i}). \quad (65)$$

Since all the terms on the right-hand side of (65) are non-negative, we deduce

$$P_v(t_{u_1} + \varepsilon) \geq \varepsilon^{-1} \varrho(\varepsilon) = 1.$$

It follows that the spike time t_v of v is not larger than $t_{u_1} + \varepsilon$. On the other hand, $P_v(t_{u_1}) = 0$. Therefore, $t_v \in (t_{u_1}, t_{u_1} + \varepsilon]$, by the continuity of P_v and the intermediate value theorem. We conclude that, for $x_1, \dots, x_{d_0} \in \mathbb{R}$,

$$\begin{aligned} & |\mathbf{R}(\Psi_\varepsilon^{\min})(x_1, \dots, x_{d_0}) - \min\{x_1, \dots, x_{d_0}\}| \\ & = |(A_{\text{out}} \circ \Phi_\varepsilon^{\min} \circ A_{\text{in}})(x_1, \dots, x_{d_0}) - \min\{x_1, \dots, x_{d_0}\}| \leq \varepsilon, \end{aligned}$$

which shows (17). It remains to observe that $\text{Size}(\Psi_\varepsilon^{\min}) = 2d_0 + 1$, as per the construction. \blacksquare

A.8 Proof of Lemma 25

We create $\Psi_{a,b,c,d,\varepsilon} = (A_{\text{in}}, \Phi, A_{\text{out}})$ as follows. We define $A_{\text{in}}: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^2$ to be such that

$$A_{\text{in}}(x) = \begin{pmatrix} -a^\top x - b \\ 0 \end{pmatrix}, \quad (66)$$

and $A_{\text{out}}: \mathbb{R} \rightarrow \mathbb{R}$ such that $A_{\text{out}}(x) = -cx + d$. In constructing $\Phi = (G, W, D)$, we follow the structure of Φ_ε^{\min} in the proof of Lemma 23. Namely, we let G consist of two synapses, (u_1, v) , (u_2, v) , with $w_{(u_i, v)} = \varepsilon^{-1}$ and $\theta_{(u_i, v)} = 0$. To see that $\Psi_{a,b,c,d,\varepsilon}$ satisfies (18), we note, as argued in the proof of Lemma 23,

$$|\mathbf{R}(\Phi)(t_1, t_2) - \min\{t_1, t_2\}| \leq \varepsilon;$$

therefore,

$$\begin{aligned} & \left| A_{\text{out}} \circ \mathbf{R}(\Phi)(t_1, t_2) - \left(-c \cdot (\min\{t_1, t_2\}) + d \right) \right| \\ &= \left| A_{\text{out}} \circ \mathbf{R}(\Phi)(t_1, t_2) - \left(c \cdot (\max\{-t_1, -t_2\}) + d \right) \right| \leq |c|\varepsilon. \end{aligned} \quad (67)$$

Combining (66), (67), we obtain (18).

Finally, it is straightforward to see that $\text{Size}(\Psi_{a,b,c,d,\varepsilon}) \leq d_0 + 5$. Moreover, the weights of $\Psi_{a,b,c,d,\varepsilon}$ are bounded above in absolute value by $\max\{1/\varepsilon, \|a\|_{\ell^\infty}, |b|, |c|, |d|\}$, and the synaptic weights are bounded below by $1/\varepsilon$.

A.9 Proof of Theorem 26

Proof For $M \in \mathbb{N}$, recall the set H_M defined in (19). Then by the well-known result (Leshno et al., 1993, Theorem 1), $\bigcup_{M \in \mathbb{N}} H_M$ is dense in $\mathcal{C}(\Omega)$. On the other hand, it is evident from Lemma 25 and Definition 11 that each H_M is contained in the closure of the set of all realizations of affine SNNs. The result now follows. \blacksquare

A.10 Proof of Theorem 29

We first prove that affine SNNs with relatively moderate size can effectively approximate the basis elements ϕ_η of $V_{\mathcal{T}}$, in the following lemma.

Lemma 41 *Let $d_0 \in \mathbb{N}$, and let $\Omega \subset \mathbb{R}^{d_0}$ be compact. Let \mathcal{T} be a regular triangulation of Ω with node set \mathcal{N} . Let $\varepsilon > 0$. Then for each node $\eta \in \mathcal{N}$ such that $G(\eta)$ is convex, there exists an SNN Ψ_ε^η such that, for all $x \in \Omega$*

$$|\mathbf{R}(\Psi_\varepsilon^\eta)(x) - \phi_\eta(x)| \leq \varepsilon.$$

Moreover,

$$\text{Size}(\Psi_\varepsilon^\eta) \leq (d_0 + 2)\#T(\eta) + 6,$$

all the weights in Ψ_ε^η are bounded above in absolute value by

$$\max\{1, 1/h_{\min}(\mathcal{T}), Cd_0/h_{\min}(\mathcal{T}), 3/\varepsilon\}$$

for some $C = C(\Omega) > 0$, and all the synaptic weights are bounded below by $\min\{1, 3/\varepsilon\}$.

Proof Since for each $\tau \in T(\eta)$, g_τ is globally affine, it takes the form, $g_\tau(x) = a_\tau^\top x + b_\tau$, where $a_\tau \in \mathbb{R}^{d_0}$, $b_\tau \in \mathbb{R}$. We build an affine SNN $\Psi_{3\varepsilon}^\eta = (A_{\text{in}}, \Phi, A_{\text{out}})$ as follows. Let $A_{\text{out}}: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that, $A_{\text{out}}(x, y) = x - 1 - y$, and $A_{\text{in}}: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{\#T(\eta)+1}$ such that

$$A_{\text{in}}(x) = \begin{pmatrix} (a_\tau^\top x + b_\tau)_{\tau \in T(\eta)} \\ 0 \end{pmatrix}.$$

Next, we consider $\Phi = (G, W, D)$, a positive SNN where G comprises $\#T(\eta) + 1$ input nodes, labeled u_0, u_τ , for $\tau \in T(\eta)$, 2 output nodes, labeled v_1, v_2 , and 1 intermediate node w , along with $\#T(\eta) + 3$ synaptic edges. The synaptic edges and their weights are,

$$w_{(u_\tau, w)} = w_{(w, v_2)} = w_{(u_0, v_2)} = \varepsilon^{-1}, \quad \text{and} \quad w_{(w, v_1)} = 1. \quad (68)$$

Let all the corresponding delays be zero. Then, given input spike times $((t_{u_\tau})_{\tau \in T(\eta)}, t_{u_0})$, we can deduce from (68) and the reasoning provided in the proof of Lemma 23 that

$$|t_w - \min\{t_{u_\tau} : \tau \in T(\eta)\}| \leq \varepsilon,$$

and $|t_{v_2} - \min\{t_w, t_{u_0}\}| \leq \varepsilon$. Therefore,

$$\begin{aligned} |t_{v_1} - (\min\{t_{u_\tau} : \tau \in T(\eta)\} + 1)| &\leq \varepsilon, \\ |t_{v_2} - \min\{t_{u_0}, \min\{t_{u_\tau} : \tau \in T(\eta)\}\}| &\leq 2\varepsilon. \end{aligned} \quad (69)$$

Integrating (69) with the specifics given to $A_{\text{in}}, A_{\text{out}}$, we conclude for $x \in \Omega$,

$$|\mathbf{R}(\Psi_{3\varepsilon}^\eta)(x) - (\min_{\tau \in T(\eta)} (a_\tau^\top x + b_\tau) - \min\{0, \min_{\tau \in T(\eta)} (a_\tau^\top x + b_\tau)\})| \leq 3\varepsilon.$$

It should be now routine to check that $\text{Size}(\Psi_{3\varepsilon}^\eta) \leq (d_0 + 2)\#T(\eta) + 6$ and that every synaptic weight in $\Psi_{3\varepsilon}^\eta$ is bounded below by $\min\{1, 1/\varepsilon\}$. As for the remaining weights, it can be inferred from (20) and the fact that $g_\tau = \phi_\eta$ on τ that

$$\|a_\tau\|_{\ell^\infty} \leq 1/h_{\min}(\mathcal{T}). \quad (70)$$

Additionally, it also follows from (20) that $g_\tau(x_\tau^*) = 0$ for some $x_\tau^* \in \Omega$. Thus

$$|b_\tau| \leq |a_\tau^\top x_\tau^*| \leq d_0 \sup\{|x| : x \in \Omega\}/h_{\min}(\mathcal{T}). \quad (71)$$

Therefore, by combining (70), (71), we conclude the final assertion of the lemma with $1/\varepsilon$ instead of $3/\varepsilon$. Finally, the result follows by substituting ε by $\varepsilon/3$. \blacksquare

Theorem 29 now follows as an immediate consequence of Lemma 41.

Proof [Proof of Theorem 29] Let $f \in V_{\mathcal{T}}$. Since $(\phi_\eta)_{\eta \in \mathcal{N}}$ is a basis of $V_{\mathcal{T}}$ satisfying (20), we can write

$$f(x) = \sum_{\eta \in \mathcal{N}} f(\eta) \phi_\eta(x).$$

For each $\eta \in \mathcal{N}$, following an argument presented in the proof of Lemma 41, we construct an affine SNN Ψ_ε^η that guarantees for all $x \in \Omega$,

$$|\mathcal{R}(\Psi_\varepsilon^\eta)(x) - f(\eta)\phi_\eta(x)| \leq |f(\eta)|\varepsilon.$$

Let $\Psi_\varepsilon^f := \bigoplus_{\eta \in \mathcal{N}} \Psi_\varepsilon^\eta$. The theorem now follows directly from an application of Lemma 12 and the triangle inequality. \blacksquare

A.11 Proof of Theorem 31

We cite the following known result, which we will use to demonstrate Theorem 31.

Proposition 42 ((Nochetto and Veese, 2011, Proposition 1)) *Let $d_0 \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^{d_0}$ be a compact domain. Let $s \in \{1, 2\}$. Let \mathcal{T} be a regular triangulation of Ω . Then for every $f \in W^{s,\infty}(\Omega)$, there exist $g = \sum_{\eta \in \mathcal{N}} c_\eta \phi_\eta$ and constants $C_1, C_2 = C_2(\mathcal{T}, d_0) > 0$, such that $|c_\eta| \leq C_1 \|f\|_{L^\infty(\Omega)}$, and that*

$$\|f - g\|_{L^\infty(\Omega)} \leq C_2 \cdot (h_{\max}(\mathcal{T}))^s \|f\|_{W^{s,\infty}(\Omega)}.$$

In addition, the constant $C_2(\mathcal{T}, d_0)$ depends on \mathcal{T} only through the so-called shape coefficient $h_{\max}(\mathcal{T})/h_{\min}(\mathcal{T})$, i.e., $C_2(\mathcal{T}, d_0) = C_2(h_{\max}(\mathcal{T})/h_{\min}(\mathcal{T}), d_0)$.

Proof [Proof of Theorem 31] In what follows, the majorant constant C is subject to change meaning from one instance to the next, and its parametric dependence, if present, will be explicitly specified. Since Ω is admissible, we can choose for $N \in \mathbb{N}$ a triangulation \mathcal{T}_N for which (22) holds. Therefore, by Proposition 42, there exists $g = \sum_{\eta \in \mathcal{N}} c_\eta \phi_\eta$ satisfying

$$\|f - g\|_{L^\infty(\Omega)} \leq CN^{-s/d_0} \|f\|_{W^{s,\infty}(\Omega)}, \quad (72)$$

where $C = C(d_0)$ only. In turn, since $g \in V_{\mathcal{T}_N}$, Theorem 29 implies the existence of an affine SNN Ψ_N^g such that

$$\|\mathcal{R}(\Psi_N^g) - g\|_{L^\infty(\Omega)} \leq CN \max_{\eta \in \mathcal{N}(\mathcal{T}_N)} |c_\eta| N^{-s/d_0-1} \leq CN^{-s/d_0} \|f\|_{L^\infty(\Omega)}. \quad (73)$$

By renaming Ψ_N^g to Ψ_N^f and combining (72), (73), we obtain (23). The size of Ψ_N^f and its associated weights can now be determined by applying the conclusions of Theorem 29 with ε in place of N^{-s/d_0-1} . \blacksquare

A.12 Proof of Theorem 33

Before proceeding, we introduce a key result derived from (Barron, 1992, Theorem 1) and (Caragea et al., 2023, Proposition 2.2), which sets the stage for our analysis.

Theorem 43 *Let $d_0 \in \mathbb{N}$. There is a universal constant $\kappa > 0$ such that the following holds. For every $K > 0$, every $f \in \Gamma_K$, and every $M \in \mathbb{N}$, there exists an element $g \in H_{8M}$ (as in (19)) such that*

$$\sup_{x \in \overline{B(0,1)}} |f(x) - g(x)| \leq \frac{\kappa d_0^{\frac{1}{2}} K}{\sqrt{M}},$$

where $\overline{B(0,1)}$ denotes the closed unit ball in \mathbb{R}^{d_0} . Furthermore, for all $i = 1, \dots, 8M$, $\|a_i\|_{\ell^\infty}, |b_i|, |c_i|, |d_i|$ in the definition of g via (19) can be chosen so that

$$\|a_i\|_{\ell^\infty}, |b_i|, |c_i|, |d_i| \leq C\sqrt{K},$$

for some $C > 0$.

In the forthcoming discussion, all majorant constants are universal, and the meaning of the analytic constants $C, c > 0$ may vary between different instances.

Proof [Proof of Theorem 33] By Theorem 43, there exist constant $\kappa > 0$ and $a_i \in \mathbb{R}^{d_0}, b_i, c_i, d_i \in \mathbb{R}$ for $i = 1, \dots, 8M$ such that $\|a_i\|_{\ell^\infty}, |b_i|, |c_i|, |d_i| \leq C\sqrt{K}$, for some $C > 0$, and

$$\sup_{x \in \overline{B(0,1)}} \left| f(x) - \sum_{i=1}^{8M} c_i \max\{a_i^\top x + b_i, 0\} + d_i \right| \leq \frac{\kappa d_0^{\frac{1}{2}} K}{\sqrt{M}}. \quad (74)$$

On the other hand, by using Lemma 25 and performing repeated additions of affine SNNs, we can construct an affine SNN Ψ such that

$$\sup_{x \in \overline{B(0,1)}} \left| \Psi(x) - \sum_{i=1}^{8M} c_i \max\{a_i^\top x + b_i, 0\} + d_i \right| \leq \frac{\kappa K}{\sqrt{M}}. \quad (75)$$

Letting $\Psi_M^f = \Psi$, we infer from (74), (75) that

$$\sup_{x \in \overline{B(0,1)}} \left| \Psi_M^f(x) - f(x) \right| \leq \frac{\kappa K \cdot (d_0^{\frac{1}{2}} + 1)}{\sqrt{M}} \leq \frac{2\kappa d_0^{\frac{1}{2}} K}{\sqrt{M}}.$$

Hence, we can set $\nu = 2\kappa$. In addition, it is readily seen from the conclusions of Lemmas 12, 25 and (75) that $\text{Size}(\Psi_M^f) \leq C d_0 M$, that all the weights in Ψ_M^f can be bounded above by $C \cdot (M^{3/2}/\sqrt{K} + \sqrt{K})$, and all of its synaptic weights can be bounded below by $cM^{3/2}/\sqrt{K}$, for some $C, c > 0$. This completes the proof. \blacksquare

A.13 Proof of Theorem 34

Proof It follows directly from (7), (26) that

$$\inf_{g \in \mathcal{H}(m^{-1/(\kappa_M+2)})} \mathcal{R}(g) \leq m^{-2/(\kappa_M+2)}.$$

Therefore, since an empirical risk minimizer is also an empirical risk minimizer, Theorem 22 and its proof apply to g_m ; that is,

$$\mathcal{R}(g_m) \leq 4 \left(\varepsilon^2 + \frac{18M(\varepsilon) \log \left(\left\lceil \frac{2B(\varepsilon)C_{\text{Lip}}^*(\varepsilon)}{\varepsilon^2} \right\rceil \right) + 72}{m} + 32\varepsilon^2 \right) + \sqrt{\frac{\log(1/\delta)}{m}}, \quad (76)$$

with $\varepsilon = m^{-1/(\kappa_M+2)}$, which happens with probability at least $1 - \delta$. To evaluate this expression and obtain (29), we proceed to estimate

$$18M(\varepsilon) \log \left(\left\lceil \frac{2B(\varepsilon)C_{\text{Lip}}^*(\varepsilon)}{\varepsilon^2} \right\rceil \right) + 72.$$

Recall from definition (12) that

$$C_{\text{Lip}}^*(\varepsilon) = d_{\text{out}}(\varepsilon) \left(2B(\varepsilon) \left(d_{\text{in}}^{\frac{1}{2}}(\varepsilon) d_0 + 1 \right) + B(\varepsilon) L(\varepsilon) \left(2 + \frac{1}{b^2(\varepsilon)} \right) + \frac{L(\varepsilon)}{b(\varepsilon)} \right) + 1,$$

where $L(\varepsilon)$ represents the graph depth $G(\varepsilon)$. Thus, using

$$\max\{d_{\text{out}}(\varepsilon), d_{\text{in}}^{\frac{1}{2}}(\varepsilon) d_0, L(\varepsilon)\} \leq M(\varepsilon) \leq \varepsilon^{-\kappa_M},$$

and $\varepsilon^{\kappa_B} \leq b(\varepsilon) \leq 1 \leq B(\varepsilon) \leq \varepsilon^{-\kappa_B}$, we deduce

$$C_{\text{Lip}}^*(\varepsilon) \leq C m^{(2\kappa_M+3\kappa_B)/(\kappa_M+2)},$$

which in turn implies that

$$18M(\varepsilon) \log \left(\left\lceil \frac{2B(\varepsilon)C_{\text{Lip}}^*(\varepsilon)}{\varepsilon^2} \right\rceil \right) + 72 \leq c \varepsilon^{-\kappa_M} \max\{1, \kappa_B/\kappa_M\} \log(m) \quad (77)$$

for a universal constant $c > 0$. Lastly, for $\varepsilon = m^{-1/(\kappa_M+2)}$, we observe

$$\frac{\varepsilon^{-\kappa_M}}{m} = m^{\frac{\kappa_M}{\kappa_M+2}-1} = m^{-\frac{2}{\kappa_M+2}} = \varepsilon^2. \quad (78)$$

The proof is now completed by substituting (77), (78), along with the expression of ε into (76). \blacksquare

Appendix B. SNNs with general synaptic weights

In the following, we consider a modified version of the model (2) where synaptic weights are allowed to take real values. The first example illustrates discontinuity in terms of input spike times, while the second illustrates discontinuity in terms of network parameters.

Example B.1 *Consider an SNN with three input neurons u_1, u_2, u_3 presynaptic to one output neuron v . Let $\mathbf{w}_{(u_1,v)} = \mathbf{w}_{(u_3,v)} = 1$ and $\mathbf{w}_{(u_2,v)} = -1$. Further, let $\theta_{(u_1,v)} = \theta_{(u_2,v)} = \theta_{(u_3,v)} = 0$. Then for every $t \in \mathbb{R}$*

$$P_v(t) = \varrho(t - t_{u_1}) - \varrho(t - t_{u_2}) + \varrho(t - t_{u_3}).$$

Observe that

$$\sum_{i=1}^3 \mathbf{w}_{(u_i, v)} > 0,$$

which is a condition that guarantees the existence of network output spike times, first identified in (Singh et al., 2023). Thus, t_v exists, as a function of $(t_{u_i})_{i=1}^3$. However, $t_v(t_{u_1}, t_{u_2}, t_{u_3})$ is discontinuous at $(t_{u_1}, t_{u_2}, t_{u_3}) = (0, 1, 2)$. Indeed, for $\varepsilon > 0$, the spike time at $(t_{u_1}, t_{u_2}, t_{u_3}) = (0, 1 + \varepsilon, 2)$ is $t_v = 1$, while the spike time at $(t_{u_1}, t_{u_2}, t_{u_3}) = (0, 1 - \varepsilon, 2)$ is $2 + \varepsilon$.

Example B.2 Consider an SNN with three input neurons u_1, u_2, u_3 presynaptic to one output neuron v . Let $\mathbf{w}_{(u_1, v)} = \mathbf{w}_{(u_3, v)} = 1$ and $\mathbf{w}_{(u_2, v)} = -1$. Further, for $s \in \mathbb{R}$, let $\theta_{(u_1, v)} = 0$, $\theta_{(u_2, v)} = 1 + s$, and $\theta_{(u_3, v)} = 2$. Let $t_v(t_{u_1}, t_{u_2}, t_{u_3}; s)$ denote the output spike time at v given the parameter s . Then it is not hard to see that for every $t \in \mathbb{R}$

$$t_v(t, t, t; s) = \begin{cases} t + 2 - s & \text{if } s < 0 \\ t + 1 & \text{if } s \geq 0. \end{cases}$$

We observe that

$$\|t_v(t, t, t; s) - t_v(t, t, t; -s)\|_{L^\infty} \geq 1$$

for all $s \in \mathbb{R}$. Letting $s \rightarrow 0$ demonstrates that $t_v(t_{u_1}, t_{u_2}, t_{u_3}; s)$ does not depend continuously on its parameters. Hence, we conclude that, contrary to feedforward neural networks, gradient-based training is generally not well-defined for SNNs with negative weights.