# PROVABLE BENEFITS OF SINUSOIDAL ACTIVATION FOR MODULAR ADDITION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

This paper studies the role of activation functions in learning modular addition with two-layer neural networks. We first show that sine activations achieve better expressiveness than ReLU activations, in the sense that the width of ReLU networks must scale linearly with the number of summands $m$ to interpolate, whereas sine networks need only two neurons. We then provide a novel Natarajan-dimension generalization bound for sine networks, which in turn leads to a nearly optimal sample complexity of $\widetilde{\mathcal{O}}(p)$ for ERM over constant width sine networks, where $p$ is the modulus. We also provide a margin-based generalization for sine networks in the overparametrized regime. We empirically validate the better generalization of sine networks over ReLU networks and our margin theory.

## 1 INTRODUCTION

Most modern neural networks use *nonperiodic* activations such as ReLU or GELU, a choice that is highly effective on vision and language benchmarks. When the target has inherently periodic structure, however, this choice can be statistically and computationally mismatched: approximating periodic functions with nonperiodic networks may require substantially larger width or depth than architectures that encode periodic features or activations (Rahaman et al., 2019; Rahimi & Recht, 2007; Tancik et al., 2020).

We study this mismatch through a standard testbed in deep learning: *modular addition*. Given $m$ input tokens in $\{0, \ldots, p-1\}$, the label is their sum modulo $p$. This task generalizes $k$-parity and is widely used to probe how networks represent and discover algorithms, as well as to study *grokking*—delayed generalization after a long memorization phase (Power et al., 2022). Mechanistic analyses report *Fourier-like* internal circuits for models that solve modular addition, where tokens are embedded as phases and addition is implemented as rotation on the unit circle. Distinct learning procedures ("clock" vs. "pizza") emerge under different hyperparameters and architectures (Nanda et al., 2023; Zhong et al., 2023). These observations suggest a simple design principle: when the task is periodic, an *explicit periodic inductive bias* should help.

Periodic representations already play a central role across machine learning. Sinusoidal positional encodings is historically canonical in Transformers (Vaswani et al., 2017); ROPE encodes positions as complex rotations, mapping offsets to phase differences and imposing a periodic bias preserving attention geometry (Su et al., 2021). Fourier and random features mitigate spectral bias and improve high-frequency fidelity (Rahimi & Recht, 2007; Tancik et al., 2020; Rahaman et al., 2019); and sinusoidal activations (SIREN) enable compact implicit neural representations for images, audio, and PDEs (Sitzmann et al., 2020). In 3D view synthesis (NeRF), Fourier positional encodings are key to recovering fine detail from coordinates (Mildenhall et al., 2020), and spectral parameterizations power operator-learning methods for PDEs (Li et al., 2021). These examples point to a general hypothesis:

*On periodic tasks, periodic bias increases expressivity and makes learning provably easier.*

We formalize and test this hypothesis in a minimal yet nontrivial setting: two-layer multilayer perceptrons (MLPs) trained on modular addition with one-hot encoding and a shared, position-independent input embedding (so the network observes bag-of-tokens counts). We compare standard ReLU activations with periodic sine activations, and analyze 0–1 multiclass classification in both the *underparameterized* and *overparameterized* regimes.

We summarize our contributions below:

1. **Expressivity: compact periodic circuits and a sharp ReLU contrast.** We construct an explicit *two-neuron* sine MLP that computes modular addition exactly under the shared embedding (Theorem 4.1), and we give a high-margin variant with width $d = 2p$. In contrast, any ReLU MLP that realizes modular addition exactly must have width $\Omega(m/p)$, i.e., growing at least linearly with the number of summands $m$ (Theorem 4.2).
2. **Unified underparameterized generalization for broad activations.** Via a multiclass Natarajan-dimension analysis based on pairwise reduction, we prove uniform convergence bounds for two-layer MLPs with a wide family of activations—piecewise-polynomial (incl. ReLU), trigonometric-polynomial (incl. sine), and rational–exponential (incl. sigmoid/SiLU/QuickGELU). The resulting sample complexity is $\widetilde{\Theta}(dp)$ with width $d$ and vocabulary size $p$ (Theorem 5.9; Table 1).
3. **Width-independent margin guarantees for overparameterized networks.** Under spectral- and Frobenius-norm constraints for ReLU and a $\|V\|_{1,\infty}$ constraint for sine, we establish multiclass, width-independent margin generalization bounds. Our sine construction attains large normalized margins, leading to population error $\widetilde{O}(p/\sqrt{n})$ when the normalized margin is $\Omega(1)$ (Theorem 6.2). In contrast, the best known ReLU interpolants achieve normalized margins that decay exponentially with $m$, yielding substantially weaker bounds under comparable norms (Theorem 6.3).
4. **Near-optimal ERM sample complexity for constant-width sine networks.** We prove that any interpolating algorithm over constant-width sine MLPs has sample complexity $\widetilde{\mathcal{O}}(p)$ (Theorem 5.11).
5. **Experiments mirroring the theory.** With matched architectures, datasets, and training budgets, sine networks consistently generalize better than ReLU MLPs on modular addition in both regimes; in the overparameterized regime, improved normalized margins track improved test accuracy (Figures 1–3).

## 2 Related Work

**Modular arithmetic as a probe of algorithmic learning and grokking.** Delayed generalization ("grokking") was popularized on modular arithmetic (Power et al., 2022). Mechanistic analyses have reverse-engineered *Fourier-feature* circuits for this task: token embeddings form phases on the unit circle and addition is implemented as rotation (Nanda et al., 2023), with distinct learned procedures ("clock" vs. "pizza") depending on models and hyperparameters (Zhong et al., 2023). For two-operand modular addition ($m$=2), recent theory supports a transition from an early *kernel* regime to later *feature learning* (Mohamadi et al., 2024), consistent with broader effective-theory accounts of grokking dynamics (Liu et al., 2022a). Margin-based analyses further show that Fourier features naturally emerge on modular addition (Morwani et al., 2024; Li et al., 2025). Together these results indicate that neural networks trained on modular arithmetic often adopt periodic internal representations, motivating architectures with *explicit* periodic bias. Grokking dynamics also interact with optimizer choice and regularization (e.g., "slingshot" instabilities for adaptive methods and optimizer-dependent time-to-grok) (Thilak et al., 2022; Tveit et al., 2025). Beyond algorithmic data, grokking-like phenomena have been observed more broadly (Liu et al., 2022b).

**Periodic representations and encodings.** Periodic structure is a long-standing ingredient in modern architectures. Transformers rely on sinusoidal or rotary positional encodings (Vaswani et al., 2017; Su et al., 2021). Random Fourier features and sinusoidal encodings mitigate spectral bias by turning the effective NTK of coordinate networks into a stationary kernel with a tunable spectrum (Rahimi & Recht, 2007; Tancik et al., 2020; Rahaman et al., 2019). Periodic activations (SIREN) enable implicit neural representations that preserve derivatives for images, audio, and PDEs (Sitzmann et al., 2020), while spectral parameterizations underpin neural operators for PDEs (Li et al., 2021). We bring these periodic ideas to a clean algorithmic setting: under shared embeddings, sine activations align with the periodicity of modular addition, yielding compact exact constructions and improved sample complexity.

**Mechanistic and optimization-centric accounts.** Mechanistic reverse-engineering reveals multiple circuit families that implement modular addition (Nanda et al., 2023; Zhong et al., 2023).

Optimization analyses connect representation choice to margin maximization and feature selection (Morwani et al., 2024; Li et al., 2025), and link late generalization to optimizer dynamics and regularization (Thilak et al., 2022; Abbe et al., 2023; Tveit et al., 2025). Our results formalize the benefit of aligning periodic architectural bias with task structure in a minimal two-layer MLP, providing both constructive and statistical advantages.

**Capacity and generalization of networks.** Across network classes with $L$ layers, $W$ parameters, and $U$ units, upper and lower bounds on capacity arise from separate techniques: lower bounds are obtained via "bit-extraction," while upper bounds follow from growth-function arguments that count sign patterns (Bartlett et al., 2017b). For piecewise-linear activations, one has nearly matching bounds $\Omega\left(WL\log(W/L)\right) \leq \mathrm{VCdim}(\mathcal{F}) \leq \mathcal{O}\left(WL\log W\right)$ (Bartlett et al., 2017b). For piecewise-polynomial activations, classical results give $\mathrm{VCdim}(\mathcal{F}) = O(WL^2 + WL\log W)$ (Anthony & Bartlett, 2009), while refined arguments yield $\mathrm{VCdim}(\mathcal{F}) = O(WU)$ together with the lower bound $\mathrm{VCdim}(\mathcal{F}) = \Omega(WL\log(W/L))$ (Bartlett et al., 2017b). For Pfaffian activations (including sigmoid and $\tanh$), one obtains capacity bounds that are polynomial in $W$ (Karpinski & Macintyre, 1997; Anthony & Bartlett, 2009). The growth-function approach has a long history: bounding the number of sign patterns generated by real polynomials yields VC-style capacity bounds for semi-algebraic hypothesis classes (Warren, 1968; Goldberg & Jerrum, 1995; Anthony & Bartlett, 2009). In the multiclass setting, uniform convergence is governed by the Natarajan-dimension (Natarajan, 1989; Haussler & Long, 1995; Shalev-Shwartz & Ben-David, 2014). We adapt these tools to discrete shared-embedding two-layer MLPs, obtaining width-independent generalization guarantees that cover ReLU and sinusoidal units.

**Learning parity and modular structure with gradient methods.** Parity functions are orthogonal characters and underlie hardness results for Statistical Query algorithms (Kearns, 1998; Blum et al., 1994; Reyzin, 2020; O'Donnell, 2014). Noise-tolerant learning of parity (LPN) appears computationally hard in general; sub-exponential algorithms are known but no polynomial-time algorithm is known (Blum et al., 2003). Recent works connect optimization dynamics to the difficulty of high-order interactions and to the hardness of learning fixed parities with neural networks (Abbe et al., 2023; Vempala & Wilmes, 2019; Shoshani & Shamir, 2025).

**Implicit bias of optimizers.** The optimizer induces implicit regularization (Gunasekar et al., 2018). For AdamW, recent analyses characterize convergence to KKT points of an $\ell_\infty$-constrained problem, leading to $\ell_\infty$-type max-margin geometry (Xie & Li, 2024; Zhang et al., 2024). For Muon, emerging analyses indicate spectral-norm constraints and an associated max-margin bias in spectral geometry, aligning with reports that Muon accelerates grokking (Chen et al., 2025; Fan et al., 2025; Tveit et al., 2025). Guided by this perspective, our overparameterized analysis yields width-free margin guarantees under norms aligned with these optimizer-induced geometries.

**Margin-based generalization guarantee.** A substantial line of work relates generalization in overparameterized networks to empirical margins and layerwise scale, rather than parameter counts (Neyshabur et al., 2018b). Prior results include spectral and entrywise $L_{2,1}$-normalized bounds for networks with Lipschitz activations (Bartlett et al., 2017a); margin bounds normalized by the Frobenius norm for homogeneous activations (Golowich et al., 2017); entrywise $L_{1,\infty}$-normalized margin bounds for Lipschitz activations (Golowich et al., 2017); and a PAC–Bayesian variant robust to weight perturbations (Neyshabur et al., 2018a). Path norms offer rescaling-invariant capacity control and, in two-layer settings, are closely connected to Barron-space viewpoints; Path-SGD encourages small path norms, often associated with larger margins and improved test performance (Neyshabur et al., 2015b; E et al., 2022; Neyshabur et al., 2015a; Gonon et al., 2024). Collectively, these insights motivate our width-independent multiclass margin bounds.

**Gradient descent and empirical margins.** In separable classification with cross-entropy, gradient methods continue decreasing loss primarily by scaling logits, thereby enlarging margins: in linear models the iterates align with the hard-margin solution while norms diverge (Soudry et al., 2018). For positively homogeneous networks, gradient flow maximizes a layer-normalized margin and converges in direction to a KKT point of the corresponding constrained margin problem (Lyu & Li, 2020; Ji & Telgarsky, 2020); mean-field analyses show analogous max-margin behavior in wide two-layer logistic models (Chizat & Bach, 2020). Beyond exact homogeneity, once risk is small, normalized margins still increase and the direction converges to KKT points; scale-normalizing mechanisms such as BatchNorm reintroduce uniform/max-margin biases (Ji & Telgarsky, 2020; Cao et al., 2023; Cai et al., 2025).

## 3 MODEL SETUP

**Notation.** For $p \in \mathbb{N}_{\geq 2}$, let $[p] := \{0, \ldots, p-1\}$ and $e_i$ denote the $i$-th standard basis vector in $\mathbb{R}^p$. For nonnegative $f, g$, we write $f(n) = \mathcal{O}(g(n))$ (resp. $f(n) = \Omega(g(n))$) if there exists an absolute constant $C > 0$ such that for all $n \geq 0$, $f(n) \leq Cg(n)$ (resp. $f(n) \geq Cg(n)$). We write $f(n) = \Theta(g(n))$ if both $\mathcal{O}$ and $\Omega$ hold. We write $f(n) = \widetilde{\mathcal{O}}(g(n))$ to suppress absolute constants (independent of the model architecture and data) and polylog factors. The symbols $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ are defined analogously.

**Task and data.** Fix integers $m, p, d \in \mathbb{N}$ with $p \geq 2$ and vocabulary $\mathcal{V} = \{0, 1, \ldots, p-1\}$. Each example is a length-$m$ sequence $s_{1:m} \in [p]^m$ with $s_1, \ldots, s_m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([p])$. We use one-hot encoding and a shared, position-independent input embedding so the network observes only the bag-of-tokens vector

$$x \;=\; \sum_{i=1}^{m} e_{s_i} \;\in \{0, 1, \ldots, m\}^p, \qquad \|x\|_1 = m.$$

The effective instance space is

$$\mathcal{X} \;=\; \big\{x \in \{0, 1, \ldots, m\}^p : \; \|x\|_1 = m\big\}, \qquad |\mathcal{X}| = \binom{m+p-1}{p-1}.$$

Labels are modular sums $y \equiv \big(\sum_{i=1}^{m} s_i\big) \pmod p \in \{0, \ldots, p-1\}$. Let $\mathcal{D}$ denote the induced population distribution on $\mathcal{X} \times [p]$. Training data are

$$S \;=\; \big\{(x^{(i)}, y^{(i)})\big\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathcal{D}^n.$$

**Model.** We study width-$d$ two-layer MLPs with shared input embedding, comparing standard ReLU to periodic sine activations. Let parameters be $\theta = (W, V) \in \Theta := \mathbb{R}^{d \times p} \times \mathbb{R}^{p \times d}$. For activation $\sigma \in \{\mathrm{ReLU}, \sin\}$ applied elementwise,

$$s^\theta(x) \;=\; V\,\sigma(Wx) \in \mathbb{R}^p, \qquad h_\theta(x) \;=\; \arg\max_{\ell \in [p]} s^\theta_\ell(x),$$

with any fixed deterministic tie-break (e.g., smallest index), so $h_\theta$ is well-defined for all $\theta$. The hypothesis class is

$$\mathcal{H}_\Theta \;=\; \{\, s^\theta : \; \theta = (W, V) \in \Theta \,\}.$$

**Training.** We minimize the empirical cross-entropy over $S = \mathcal{D}_{\text{train}}$ using mini-batches, treating $s^\theta(x)$ as logits for the $p$-class problem with labels in $[p]$. Optimization uses AdamW and Muon; implementation details and hyperparameters are provided in Appendix B.

## 4 EXPRESSIVITY OF SINE AND RELU MLPS

We begin by comparing expressivity under our shared, position-independent input embedding. A two-neuron sine MLP computes modular addition exactly, whereas ReLU MLPs require width that grows with the number of summands. Proofs are provided in Section F.1.1 and Section G.

**Theorem 4.1** (Low-width construction for sine MLP)**.** *There exists a construction with hidden dimension $d = 2$ and sine activation that realizes $\sum_{i=1}^{m} s_i \bmod p$ for all $x = (s_1, \cdots, s_m) \in \mathcal{X}$.*

**Theorem 4.2** (Necessary width for modular addition with ReLU)**.** *Let $\sigma(t) = \max\{t, 0\}$. If $h_\theta(x) = \arg\max_\ell s^\theta_\ell(x)$ realizes modular addition exactly on $\mathcal{X}$, then necessarily*

$$d \;\geq\; \frac{m}{p} - 1.$$

However, the remarkable expressivity of sine-activated MLPs does not ensure generalization. In fact, even a constant-size sine-activated MLP realizes a one-parameter hypothesis class on $\mathbb{N}$ with infinite VC-dimension:

**Example 4.3** (Lemma 7.2 (Anthony & Bartlett, 2009); see also Appendix C)**.** *The class $\mathcal{F} = \{x \mapsto \mathrm{sgn}(\sin(ax)) : a \in \mathbb{R}^+\}$ of functions defined on $\mathbb{N}$ has $\mathrm{VCdim}(F) = \infty$.*

Thus, expressivity alone does not imply generalization. Leveraging the structure of our discrete, bounded-input, we establish uniform convergence bounds that scale linearly with parameter counts.

## 5 GENERALIZATION IN THE UNDERPARAMETERIZED REGIME

We provide uniform-convergence guarantees for two-layer MLPs with a broad family of activations under shared embeddings. Informally, our proof counts sign patterns induced by pairwise margins via growth-function bounds in the parameter space, echoing classical arguments of (Warren, 1968; Goldberg & Jerrum, 1995) with the multiclass reduction to the Natarajan-dimension (Shalev-Shwartz & Ben-David, 2014; Anthony & Bartlett, 2009). Intuitively, it determines the richness of the output class of the model. Proof details are in Appendix E.

**Definition 5.1** (Shattering and VC-dimension)**.** Let $\mathcal{B} \subseteq \{-1,+1\}^{\mathcal{Z}}$ be a binary hypothesis class on a domain $\mathcal{Z}$. A finite set $T \subset \mathcal{Z}$ is *shattered* by $\mathcal{B}$ if every labeling of $T$ is realized by some $b \in \mathcal{B}$, i.e., $\mathcal{B}_{|_T} = \{-1,+1\}^T$. The *VC-dimension* of $\mathcal{B}$, denoted $\mathrm{VCdim}(\mathcal{B})$, is

$$\mathrm{VCdim}(\mathcal{B}) \;=\; \sup\big\{\, |T| :\; T \subset \mathcal{Z} \text{ is finite and shattered by } \mathcal{B} \,\big\},$$

with the convention that $\mathrm{VCdim}(\mathcal{B}) = \infty$ if sets of arbitrarily large finite size are shattered.

**Definition 5.2** (Growth function)**.** Let $\mathcal{B} \subseteq \{-1,+1\}^{\mathcal{Z}}$ be a binary hypothesis class on a domain $\mathcal{Z}$. For $m \in \mathbb{N}$, the *growth function* of $\mathcal{B}$ is

$$\Pi_{\mathcal{B}}(m) \;:=\; \max\big\{\, \big|\mathcal{B}_{|_T}\big| : T \subseteq \mathcal{Z},\ |T| = m \,\big\}.$$

**Definition 5.3** (Shattering and Natarajan-dimension)**.** Let $\mathcal{H} \subseteq [p]^{\mathcal{X}}$ be a multiclass hypothesis class. A finite set $S \subset \mathcal{X}$ is *Natarajan-shattered* by $\mathcal{H}$ if there exist $f_1, f_2 \in [p]^S$ with $f_1(x) \neq f_2(x)$ for all $x \in S$, such that for every selector $b : S \to \{1,2\}$ there is $h_b \in \mathcal{H}$ with $h_b(x) = f_{b(x)}(x)$ for all $x \in S$. The *Natarajan-dimension* of $\mathcal{H}$, denoted $\mathrm{Ndim}(\mathcal{H})$, is

$$\mathrm{Ndim}(\mathcal{H}) \;=\; \sup\big\{\, |S| :\; S \subset \mathcal{X} \text{ is finite and Natarajan-shattered by } \mathcal{H} \,\big\},$$

with the convention that $\mathrm{Ndim}(\mathcal{H}) = \infty$ if sets of arbitrarily large finite size are Natarajan-shattered.

**Definition 5.4** (Network class and pairwise reduction)**.** Let $\mathcal{H}_\Theta \subseteq [p]^{\mathcal{X}}$ be a $p$-class network class realized by score vectors $s^\theta(x) = (s_1^\theta(x), \ldots, s_p^\theta(x)) \in \mathbb{R}^p$, $\theta \in \Theta$, $x \in \mathcal{X}$, and a fixed, deterministic tie-breaking rule for $\arg\max$:

$$h_\theta(x) \;=\; \arg\max_{\ell \in [p]} s_\ell^\theta(x).$$

Define the *pairwise reduction* on the domain

$$\mathcal{Z}_{\mathrm{pair}} \;:=\; \mathcal{X} \times \{(i,j) \in [p] \times [p] : i < j\}$$

by the reduction class $\mathcal{G}_\Theta \subseteq \{-1,+1\}^{\mathcal{Z}_{\mathrm{pair}}}$ for $\mathcal{H}_\Theta$ of functions

$$g_\theta(x,i,j) = \mathrm{sgn}\big(s_i^\theta(x) - s_j^\theta(x)\big) = \begin{cases} +1, & \text{if } s_i^\theta(x) \geq s_j^\theta(x), \\ -1, & \text{if } s_i^\theta(x) < s_j^\theta(x) \end{cases}$$

The lemma below connects the Natarajan-dimension to the growth function of the reduction class, which is a key tool in this section.

**Lemma 5.5** (Natarajan shattering and the growth function)**.** *If $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ is Natarajan-shattered by a $p$-class network class $\mathcal{H}_\Theta$, then*

$$2^n \;\leq\; \Pi_{\mathcal{G}_\Theta}\big(n\, p(p-1)/2\big),$$

*where $\mathcal{G}_\Theta$ is the reduction class of $\mathcal{H}_\Theta$.*

**Definition 5.6** (Piecewise-polynomial activation)**.** A function $\sigma : \mathbb{R} \to \mathbb{R}$ is *piecewise polynomial with at most $L \geq 1$ pieces and maximal piece degree $r \geq 1$* if there exist breakpoints

$$-\infty = b_0 < b_1 < \cdots < b_{L-1} < b_L = +\infty$$

and polynomials $P_1, \ldots, P_L$ with $\deg P_\ell \leq r$ such that $\sigma(t) = P_\ell(t)$ for all $t \in (b_{\ell-1}, b_\ell]$, $\ell \in [L]$.

**Definition 5.7** (Trigonometric-polynomial activation)**.** Let $K \in \mathbb{N}_0$. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is a *trigonometric polynomial of degree at most $K$* if

$$\sigma(t) = a_0 + \sum_{k=1}^{K} \big(a_k \cos(kt) + b_k \sin(kt)\big)$$

for some real coefficients $a_0, (a_k)_{k \leq K}, (b_k)_{k \leq K}$.

Table 1: Comparison of existing capacity bounds for two-layer MLPs with $W$ trainable parameters and width $d$. Notation $\widetilde{\Theta}(\cdot)$ hides polylog factors. Input types affect expressiveness because richer inputs permit higher capacity. Bold entries are this paper's contributions; precise references for externally sourced VC-dimension bounds are collected in Appendix C.

| Activation | Input type | VCdim[1] | Ndim |
|---|---|---|---|
| Piecewise linear | real inputs | $\Theta(W \log W)$ | $\widetilde{\boldsymbol{\Theta}}(\boldsymbol{W})$ |
| Piecewise polynomial | real inputs | $\Theta(W \log(W))$ | $\widetilde{\boldsymbol{\Theta}}(\boldsymbol{W})$ |
| Pfaffian, incl. standard sigmoid | real inputs | $\mathcal{O}(d^2W^2)$ | — |
| Standard sigmoid | real inputs | $\Omega(W \log W)$ | $\Omega(WlogW)$ |
| **Standard sigmoid** | **discrete, bounded inputs** | $\Omega(W)$, $\widetilde{\mathcal{O}}(W)$ | $\widetilde{\boldsymbol{\Theta}}(\boldsymbol{W})$ |
| Sine | discrete, unbounded inputs | $\infty$ | $\infty$ |
| **Trigonometric polynomial** | **discrete, bounded inputs** | — | $\widetilde{\boldsymbol{\mathcal{O}}}(\boldsymbol{W})$ |
| **Rational exponential** | **discrete, bounded inputs** | — | $\widetilde{\boldsymbol{\mathcal{O}}}(\boldsymbol{W})$ |

**Definition 5.8** (Polynomial–rational–exponential activation)**.** Fix $k \in \mathbb{R} \setminus \{0\}$, $c \geq 0$, $\tau > 0$, $a, b \in \mathbb{R}$, and a polynomial $P$ with degree $r := \deg P \in \mathbb{N}_0$. Define

$$\sigma(t) = P(t)\frac{ae^{kt} + b}{ce^{kt} + \tau}.$$

**Theorem 5.9** (Uniform convergence for broad activation families)**.** *Let $\sigma$ be one of: piecewise-polynomial (Def. 5.6), trigonometric-polynomial (Def. 5.7), or polynomial–rational–exponential (Def. 5.8). Let $\mathcal{H}_\sigma$ be the corresponding two-layer class. Then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of $\mathcal{D}_{train} \sim \mathcal{D}^n$,*

$$\sup_{h \in \mathcal{H}_\sigma} \left| \mathbb{P}_{(X,Y)\sim\mathcal{D}}(h(X) \neq Y) - \mathbb{P}_{(X,Y)\sim\mathcal{D}_{train}}(h(X) \neq Y) \right| \leq \widetilde{O}\left( \sqrt{\frac{dp + \log(1/\delta)}{n}} \right).$$

As direct corollaries of Theorem 5.9, we have:

**Corollary 5.10.** *Two-layer MLPs with activation ReLU ($\sigma(t) = \max\{0, t\}$), monomial ($\sigma(t) = t^m$), sine ($\sigma(t) = \sin t$), Sigmoid ($\sigma(t) = \frac{e^t}{e^t+1}$), SiLU ($\sigma(t) = t\,\mathrm{sigmoid}(t)$), QuickGELU ($\sigma(t) = t\,\mathrm{sigmoid}(\beta t),\ \beta > 0$) have sample complexity $\widetilde{O}(dp)$.*

**Corollary 5.11** (Sample complexity upper bound for ERM with constant-width sine networks)**.** *Fix a constant width $d \geq 2$. With probability at least $1 - \delta$ over the random draw of $\mathcal{D}_{train} \sim \mathcal{D}^n$, for all interpolating solutions $\theta$,*

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}}\big[h_{\hat{\theta}}(X) \neq Y\big] \leq \widetilde{O}\left( \sqrt{\frac{p + \log(1/\delta)}{n}} \right),$$

*where $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $n$, $m$, and $\delta^{-1}$.*

*Consequently, the sample complexity is $\widetilde{\mathcal{O}}(p)$.*

The bound in Theorem 5.11 is essentially near-optimal. Intuitively, if an algorithm fails to observe a constant fraction of the total $p$ classes, learning is information-theoretically impossible. We formalize this via a PAC lower bound in Theorem D.6, where we apply a uniformly random permutation to the labels so the learner does not know which output index corresponds to which residue class.

## 6 Generalization in overparameterized regime

From a uniform-convergence bound, two-layer sine MLPs admit better generalization guarantees than ReLU networks. However, those bounds scale with the hidden width. A natural question is:

[1] VC-dimension lower bounds are existential: for given size and depth budgets, there exists a network that shatters a set of the claimed cardinality. Upper bounds are universal: they hold for every network in the family.

as the width becomes very large, what happens? We show margin-based bounds that are width-independent.

Let $v_j$ ($j \in [p]$) be the $j$-th row of $V$, we use the norm $\|V\|_{1,\infty} := \max_j \|v_j\|_1$ (maximum row $\ell_1$-norm), $\|V\|_2$ for the spectral norm, and $\|W\|_F$ for the Frobenius norm.

**Definition 6.1** (Empirical margin)**.** For a labeled example $(x, y)$ with $y \in [p]$ and score vector $s^\theta(x) \in \mathbb{R}^p$, define the multiclass margin

$$\gamma_\theta(x, y) \;:=\; s^\theta_y(x) \;-\; \max_{k \neq y} s^\theta_k(x).$$

For a finite sample $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, define its margin as

$$\gamma_\theta(S) \;:=\; \min_{i \in [n]} \gamma_\theta\big(x^{(i)}, y^{(i)}\big).$$

We say the classifier interpolates $S$ if $\gamma_\theta(S) > 0$.

**Theorem 6.2** (Two-layer sin MLP, margin-based generalization)**.** *Consider the two-layer MLP $s^\theta(x) = V \sin(Wx) \in \mathbb{R}^p$ on $\mathcal{X}$. Fix $\delta \in (0, 1)$ and assume $d \geq 2p$. With probability at least $1-\delta$ over the random draw of $\mathcal{D}_{train} \sim \mathcal{D}^n$, for all interpolating solutions $\theta$ with normalized margin $\overline{\gamma}_{\theta,\sin} := \frac{\gamma_\theta(\mathcal{D}_{\text{train}})}{\|V\|_{1,\infty}} = \Omega(1)$, it holds that*

$$\mathbb{P}_{(X,Y)\in\mathcal{D}}\big[h_\theta(X) \neq Y\big] \;\leq\; \widetilde{O}\left(p\sqrt{\frac{1}{n}}\right),$$

*where $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $n$, $m$, and $\delta^{-1}$.*

**Theorem 6.3** (Two-layer ReLU MLP, margin-based generalization)**.** *Assume $p > m$ and $n > m^2$, $n \geq 17$. Fix $\delta \in (0, 1)$. Suppose the width satisfies $d \;\geq\; 64\, p\, m^{\frac{m}{2}+2}\, 4.67^m$. With probability at least $1 - \delta$ over the random draw of $\mathcal{D}_{train} \sim \mathcal{D}^n$, for all interpolating solutions $\theta$ with normalized margin $\overline{\gamma}_{\theta,\mathrm{ReLU}} \;:=\; \frac{\gamma_\theta(\mathcal{D}_{\text{train}})}{\|V\|_2\|W\|_F} = \Omega\left(\frac{1}{\sqrt{p}} \cdot \frac{1}{m^{1.5m+2.5}\, 6.34^m}\right)$, it holds that*

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}}\big[h_\theta(X) \neq Y\big] \;\leq\; \widetilde{O}\left(p\, m^{1.5m+2.5} 6.34^m \sqrt{\frac{m}{n}}\right),$$

*where $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $n$ and $\delta^{-1}$.*

The proofs proceed by first applying an $\ell_\infty$ vector-contraction bound (Foster & Rakhlin, 2019) for the Rademacher complexity. For sine MLPs, we then bound the contracted complexity via the standard Dudley entropy integral, after estimating covering numbers for sine networks. For ReLU MLPs, we invoke a key technical lemma for positively homogeneous activations that enables a layer-wise peeling argument within the Rademacher complexity, following (Golowich et al., 2017). See Appendix H for details.

## 7 Experiments

To empirically investigate and confirm our proposed theory, we conduct experiments with two-layer sine and ReLU MLPs on modular addition under various settings. Full setup details and additional figures are deferred to Appendix B.

**Underparameterized regime.** We evaluate our sample complexity predictions by training matched architectures that differ only in their nonlinearity (sine vs. ReLU), using AdamW with zero weight decay on identical datasets and with identical optimization hyperparameters (Figure 1). Across widths and training sizes, sine networks consistently outperform ReLU in both training and test accuracy, attaining a given accuracy at substantially smaller widths. For a fixed training set size, reducing the width—provided it remains sufficient for optimization—improves test accuracy for both activations, consistent with our uniform convergence guarantee in Section 5.

**Overparameterized regime.** To verify the margin-based bounds in Section 6, we train wide two-layer MLPs with Muon and sweep over decoupled weight decay rates. For sine models we apply weight decay only to the second layer; for ReLU we decay both layers. We report the $0.5\%$-quantile

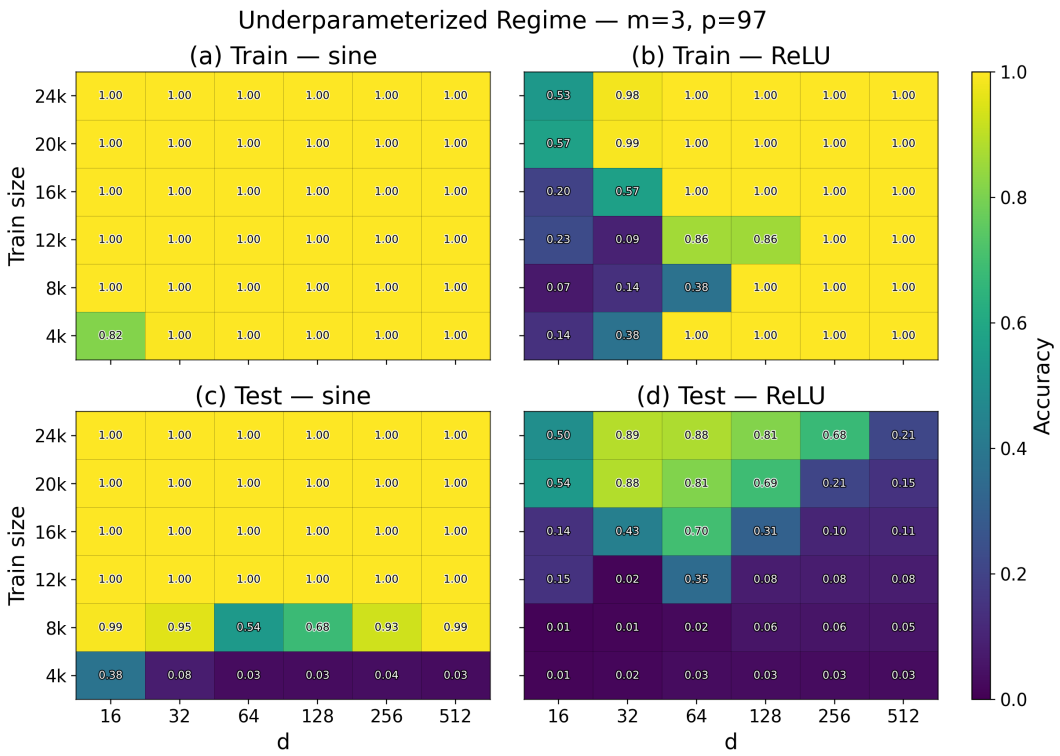


Figure 1: Accuracies for two-layer sine and ReLU MLPs in the underparameterized regime.

rather than the minimum margin because the latter is often dominated by rare outliers; a small quantile yields a stable large-margin proxy and, by Corollary H.4, only adds an additive $0.5\%$ term to the population error. We log training and test accuracies, the 0.5%-quantile of the training margin $\gamma_{\text{train}}^{0.5\%} := \text{Quantile}_{0.005}\left\{\gamma_\theta(x^{(i)}, y^{(i)})\right\}_{i=1}^n$, together with normalized margins that factor out layer scales:

$$\text{ReLU:} \quad \widehat{\gamma}_{\text{ReLU}} = \frac{\gamma_{\text{train}}^{0.5\%}}{\|V\|_2 \, \|W\|_F}, \qquad \text{Sine:} \quad \widehat{\gamma}_{\text{sin}} = \frac{\gamma_{\text{train}}^{0.5\%}}{\|V\|_{1,\infty}}.$$

Figures 2 and 3 show that, as weight decay increases through a moderate range, normalized margins grow and test accuracy improves; with excessively large decay, training accuracy falls and generalization degrades. These trends align with the prediction that, in the overparameterized regime, generalization is governed by effective layer scales and margins.

## 8 Conclusion

We show a provable benefit of learning periodic tasks with sine activations over standard ReLU activations. On modular addition with shared, position-independent embeddings, a width-2 sine MLP exactly implements the task, whereas ReLU MLPs require width that grows at least linearly with $m$, separating the two families in representational efficiency. On the statistical side, our Natarajan-dimension analysis yields uniform convergence bounds of $\widetilde{\Theta}(dp)$ for broad activation families (Theorem 5.9); specialized to sine with constant width and using realizability, any interpolating learner achieves nearly optimal $\widetilde{\Theta}(p)$ sample complexity (Theorem 5.11). In the overparameterized regime, we prove width-independent, margin-based generalization guarantees for sine networks under the natural $\|V\|_{1,\infty}$ scaling; larger-margin interpolants achieve sample complexity $\tilde{O}(p^2)$ (Theorem 6.2). Empirically, we observe that sine networks train and generalize consistently reliably than ReLU MLPs, and larger normalized margins track better generalization (Figures 1–3). Together, these results support a clear design principle: when the target is periodic, encoding periodic structure in the architecture, both increases expressivity and makes learning easier.

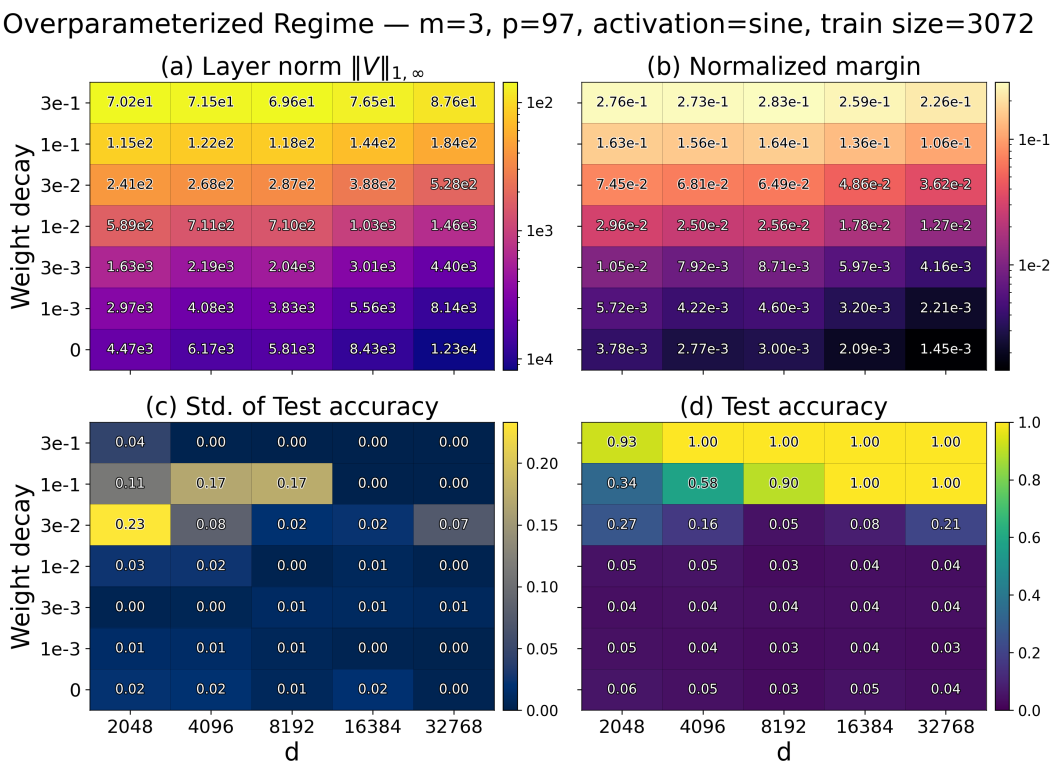


Figure 2: Two-layer sine networks in the overparameterized regime. Clockwise from the top-left: Layer norm, Normalized margin, Test accuracy, Standard deviation of Test accuracy.

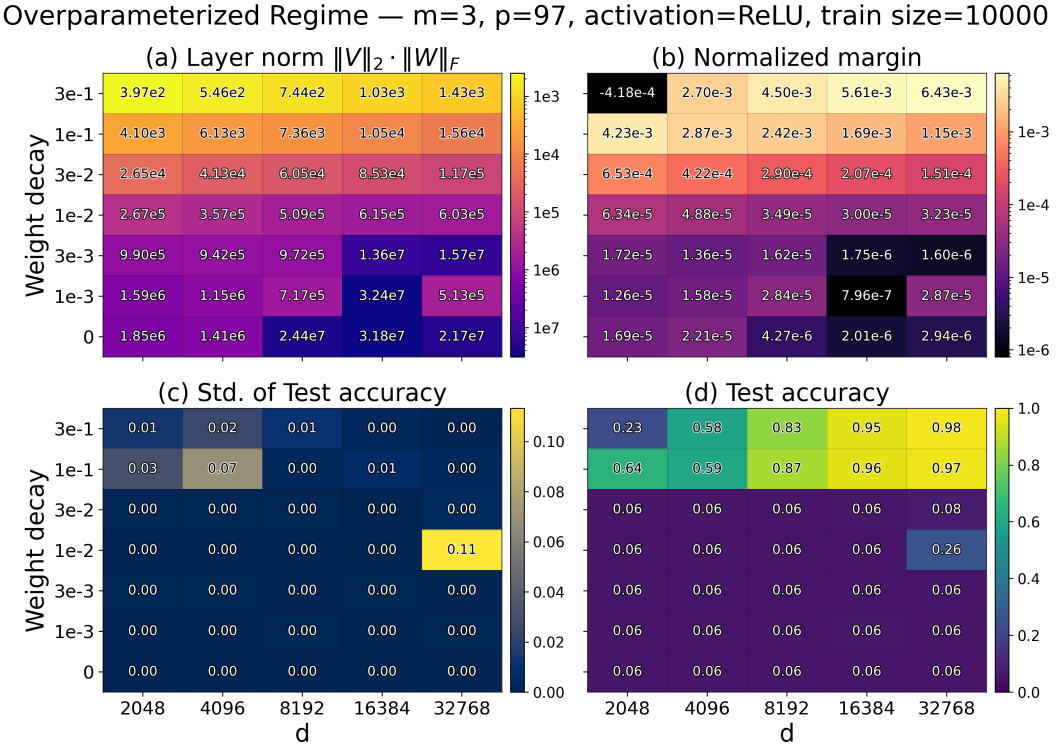


Figure 3: Two-layer ReLU networks in the overparameterized regime (panels as in Fig. 2).

## REFERENCES


Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009. ISBN 978-0-521-11862-0. doi: 10.1017/CBO9780511624216. Paperback reissue of the 1999 edition.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017a.

Peter L. Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *arXiv preprint arXiv:1703.02930*, 2017b. doi: 10.48550/arXiv.1703.02930. URL https://arxiv.org/abs/1703.02930.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing (STOC)*, pp. 421–430. ACM, 1994.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025. URL https://arxiv.org/abs/2502.16075. Also available as arXiv:2502.16075.

Yuan Cao, Difan Zou, Yuanzhi Li, and Quanquan Gu. The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5699–5753. PMLR, Jul 12–15 2023. URL https://proceedings.mlr.press/v195/cao23a.html.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025. URL https://arxiv.org/abs/2506.15054.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1305–1338. PMLR, 2020. URL https://proceedings.mlr.press/v125/chizat20a.html.

Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022. doi: 10.1007/s00365-021-09549-y. URL https://doi.org/10.1007/s00365-021-09549-y.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *Proceedings of the 3rd Workshop on High-dimensional Learning Dynamics (HiLD)*, 2025. URL https://openreview.net/forum?id=tirGweSx3a.

Dylan J. Foster and Alexander Rakhlin. $\ell_\infty$ vector contraction for rademacher complexity, 2019. URL https://arxiv.org/abs/1911.06468.

Paul W. Goldberg and Mark Jerrum. Bounding the vapnik–chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2–3):131–148, 1995. doi: 10.1007/BF00993408.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017. URL https://arxiv.org/abs/1712.06541. submitted 18 December 2017, revised 17 November 2019.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A path-norm toolkit for modern networks: Consequences, promises and challenges. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b3732a13897c4cea145c3bdece80de64-Paper-Conference.pdf.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry, 2018.

David Haussler and Philip M. Long. A generalization of sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995. doi: 10.1016/0097-3165(95)90001-2.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2006.06657.

Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997. doi: 10.1006/jcss.1997.1477.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45 (6):983–1006, 1998. doi: 10.1145/293347.293351.

Chenyang Li, Yingyu Liang, Zhenmei Shi, Zhao Song, and Tianyi Zhou. Fourier circuits in neural networks and transformers: A case study of modular arithmetic with multiple inputs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. arXiv:2402.09469.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/69f283a00ec6cd89e21d9366f7a79cea-Paper-Conference.pdf.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data, oct 2022b. URL https://arxiv.org/abs/2210.01117.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1906.05890.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? A theoretical analysis of grokking modular addition. *arXiv preprint arXiv:2407.12332*, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M. Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, October 1989. doi: 10.1007/BF00114804.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pp. 2422–2430, 2015a. URL https://proceedings.neurips.cc/paper/2015/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401. PMLR, 2015b. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018a. URL https://arxiv.org/abs/1707.09564.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks, 2018b.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, UK, 2014. ISBN 978-1-107-03832-5. doi: 10.1017/CBO9781139814782.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Nasim Rahaman, Devansh Arpit, Aristide Baratin, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Lev Reyzin. Statistical queries and statistical algorithms: Foundations and applications. *CoRR*, abs/2004.00557, 2020. doi: 10.48550/arXiv.2004.00557. URL https://arxiv.org/abs/2004.00557.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Itamar Shoshani and Ohad Shamir. Hardness of learning fixed parities with neural networks, 2025. URL https://arxiv.org/abs/2501.00817. v2, last revised 8 Jan 2025.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19:1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and SQ lower bounds. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 3115–3117. PMLR, 2019. URL https://proceedings.mlr.press/v99/vempala19a.html.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968. doi: 10.1090/S0002-9947-1968-0226281-1.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: $\ell_\infty$ norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. *arXiv preprint arXiv:2406.10650*, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

## A Disclosure of Large Language Model Usage

**Tool and scope.** We used Gemini 2.5 Pro and GPT-5 (high) as general-purpose assist tools for (i) code assistance (e.g., suggesting small snippets, refactoring, debugging hints, writing docstrings/comments, and drafting unit-test scaffolds) and (ii) writing assistance (e.g., copy-editing, grammar/fluency improvements, and localized rephrasing for clarity). Prompts sometimes included short excerpts of our own draft text or code necessary to request the above assistance.

**What the LLM did not do.** The LLM did not originate the paper's core research ideas, hypotheses, methodological designs, experimental protocols, analyses, or conclusions; it did not write sections containing novel scientific claims; and it did not determine which results to report or how to interpret them.

**Verification and oversight.** All LLM-suggested text and code were independently reviewed and edited by the authors. For code, the authors carefully checked and verified correctness (including running and testing LLM-suggested snippets before inclusion). Any factual statements in edited prose were cross-checked by the authors against our own results or appropriate sources. No LLM outputs were accepted without human scrutiny.

**Assessment of significance.** While the LLM provided editing assistance and code-level suggestions, its role does not rise to the level of a contributing author under the ICLR policy. The intellectual contributions (problem formulation, algorithmic design, experiments, and interpretation) are those of the human authors.

**Reproducibility note.** LLM assistance was limited to improving clarity and developer ergonomics; it does not affect the reproducibility of our methods or results. All final code and experiments are authored, verified, and maintained by the authors.

## B Experimental Setup and Additional Results

In this section, we explain the configuration used in all experiments.

### B.1 Experimental Setup

**Data.** For each run we generate a static training set of size $n$ once and reshuffle it every epoch; the test set contains $10{,}000$ i.i.d. samples.

**Initialization and reproducibility.** We fix seeds $\{1337, 1338, 1339\}$ and report averages over seeds for metrics. All weights are initialized i.i.d. $\mathcal{N}(0, 0.01^2)$. This ensures consistent model initializations and that smaller training sets are strict subsets of larger ones within a given sweep.

**Precision and implementation.** All experiments are implemented in PyTorch with TF32 disabled and `float32` throughout. We log with Weights & Biases. Each run uses a single NVIDIA GPU (RTX,A4000/A6000, RTX,5000/6000,Ada, L40S, A100, H100, or H200).

**Optimizers and hyperparameters in underparameterized regime.** We use AdamW with a constant learning rate $10^{-3}$ and *zero* weight decay. All other AdamW hyperparameters are left at their PyTorch defaults (betas $(0.9, 0.999)$, $\varepsilon = 10^{-8}$). We do not use learning-rate schedules, warmup, or gradient clipping.

**Optimizers and hyperparameters in overparameterized regime.** We use Muon with constant learning rate $10^{-3}$ and vary the decoupled weight decay. Momentum, Nesterov, and Newton–Schulz steps are left at the library defaults (momentum $0.95$, Nesterov enabled, $5$ Newton–Schul steps). We do not use learning-rate schedules, warmup, or gradient clipping.

**Batches.** We use mini-batchs and the batch size is $1024$. We train for $300,000$ epochs. Where we compare activations, we match $(m, p, d, n)$ and optimizer settings.

**Training.** We use mini-batch training and the batch size is $1024$. We train for up to $300{,}000$ epochs and report the final metrics after training.

**Metrics.** We report train/test accuracy, the generalization gap, and the $0.5$th-percentile training margin $\gamma_{\text{train}}^{0.5\%}$. We log layer norms $\|W\|_F$, $\|V\|_2$ for ReLU and $\|V\|_{1,\infty}$ for sine models, enabling the normalized margins used in Section 7.

**Normalization choices.** Our denominators follow decoupled weight-decay scales: spectral/Frobenius norms are natural for controlling effective layer size; note $\|A\|_F \leq \sqrt{\operatorname{rank}(A)}\,\|A\|_2 \leq \sqrt{\min\{m,n\}}\,\|A\|_2$ for $A \in \mathbb{R}^{m\times n}$.

**Additional results.** We provide additional figures for our experiments.

Figure 4 and Figure 5 provide multiple plots for the underparameterized sweeps at $(m, p) = (2, 307)$ and $(4, 53)$, respectively. In both cases, sine networks dominate ReLU at matched width and training budget, and the advantage widens as width decreases until optimization begins to fail.

Figure 6–9 provide multiple plots for the overparameterized sweeps at $(m, p) = (2, 307)$ and $(4, 53)$, respectively. In both cases, as weight decay increases through a moderate range, normalized margins grow and test accuracy improves; with excessively large decay, training accuracy falls and generalization degrades. These trends align with the prediction that, in the overparameterized regime, generalization is governed by effective layer scales and margins.

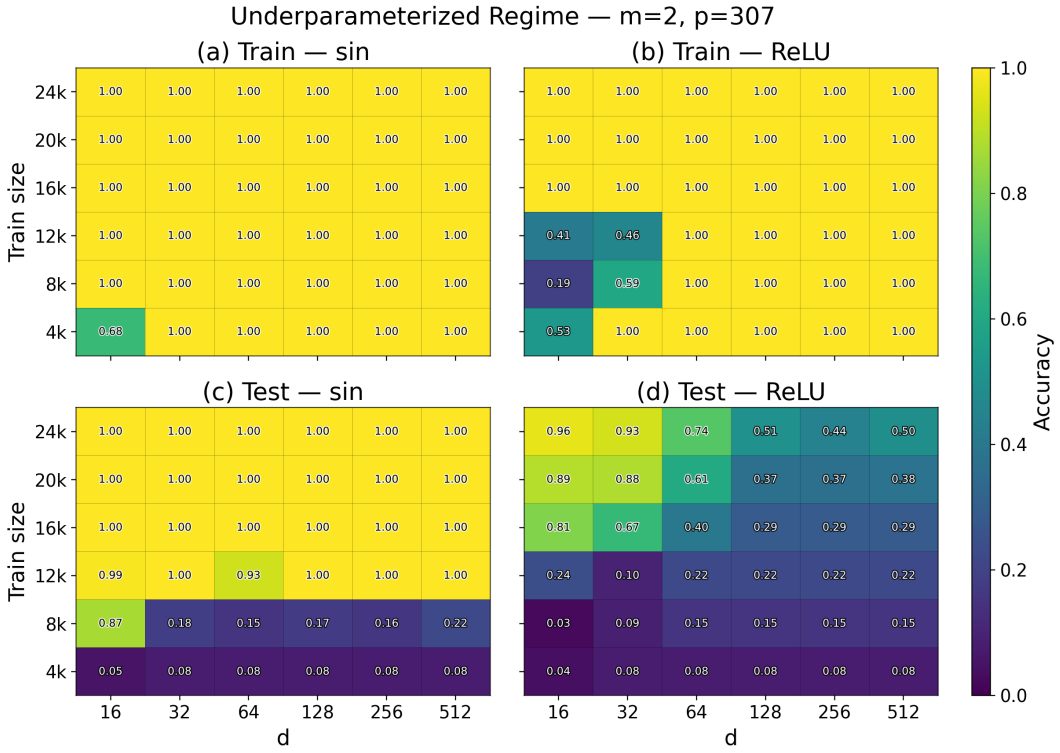


Figure 4: Underparameterized regime ($m = 2$, $p = 307$). Final train/test accuracies for two-layer MLPs with sine vs. ReLU activations under matched budgets.

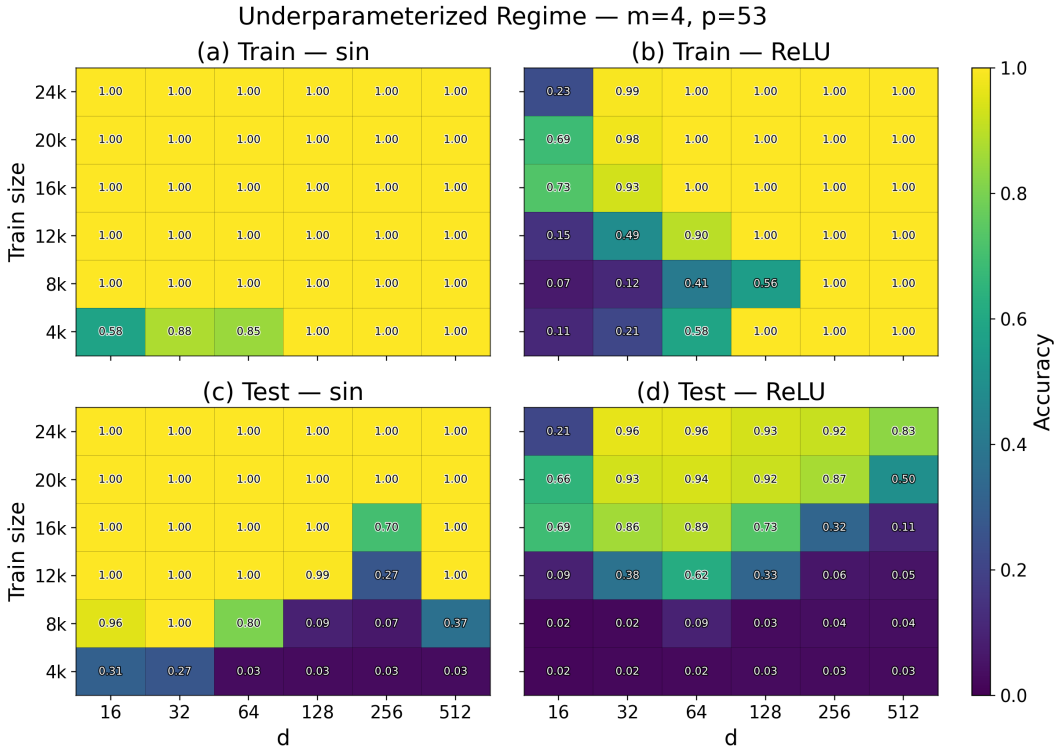


Figure 5: Underparameterized regime ($m = 4$, $p = 53$). Final train/test accuracies for two-layer MLPs with sine vs. ReLU activations under matched budgets.

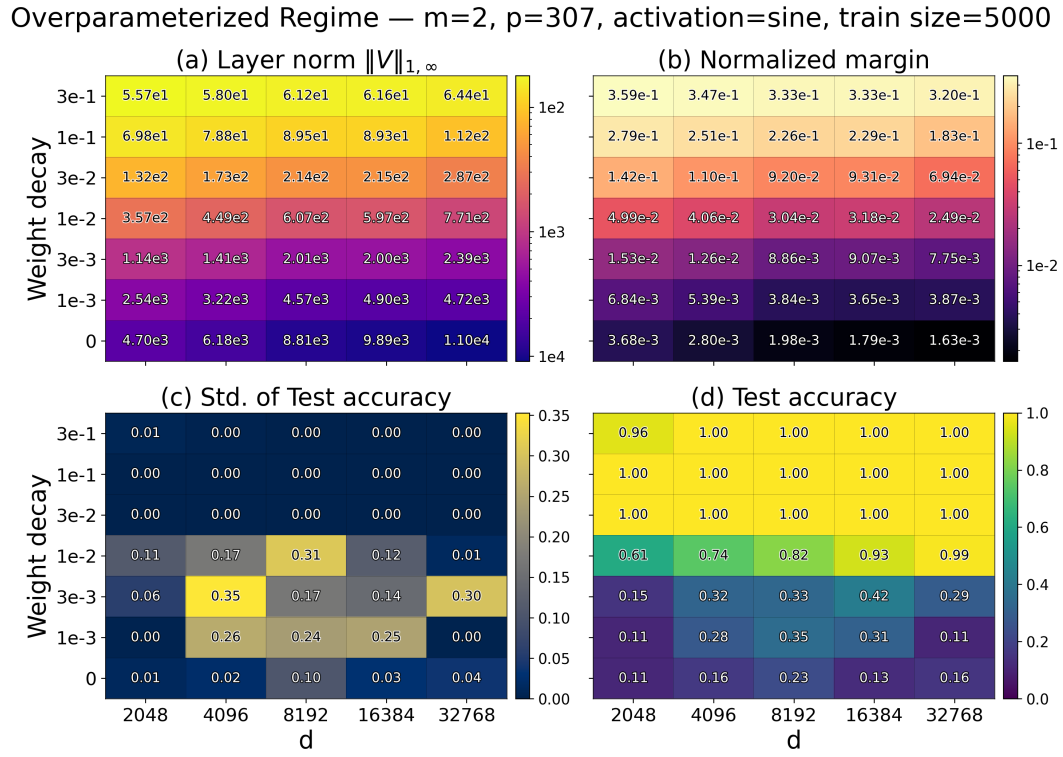


Figure 6: Two-layer sine networks in the overparameterized regime.

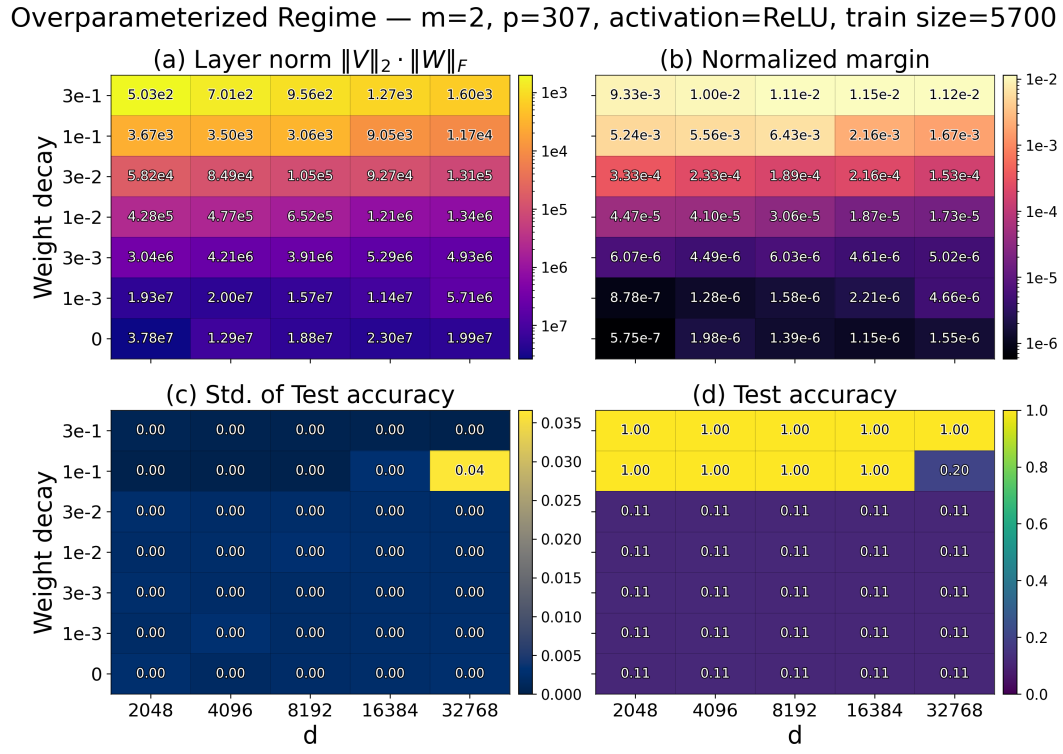


Figure 7: Two-layer ReLU networks in the overparameterized regime.

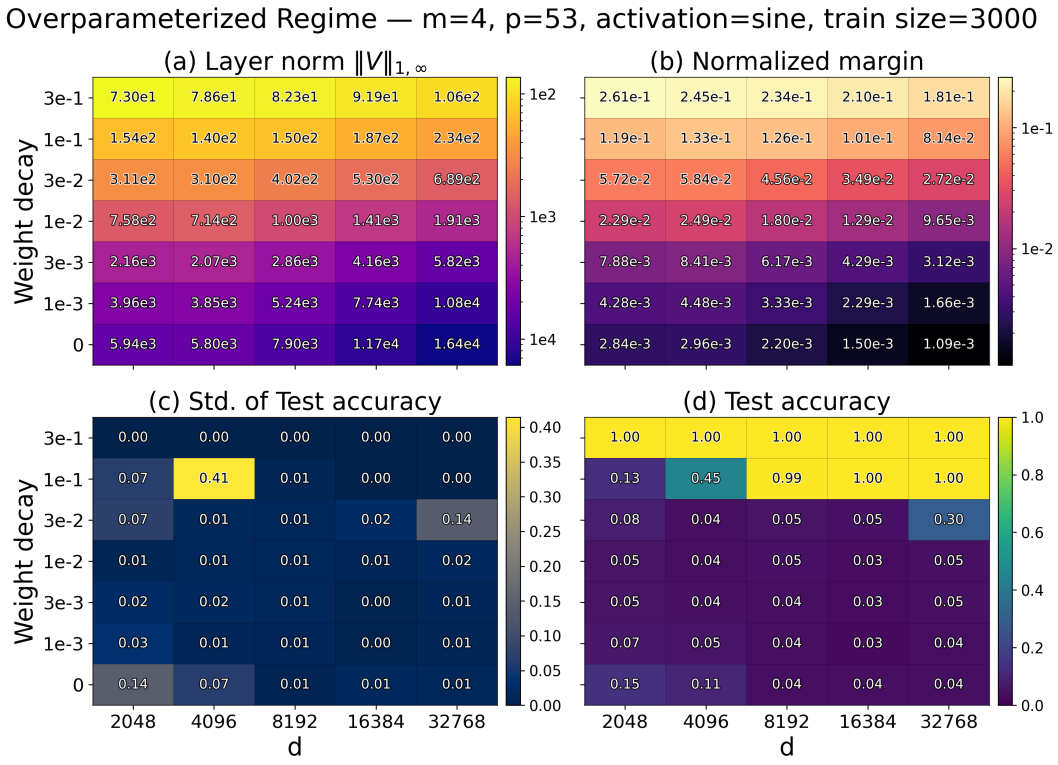


Figure 8: Two-layer sine networks in the overparameterized regime.

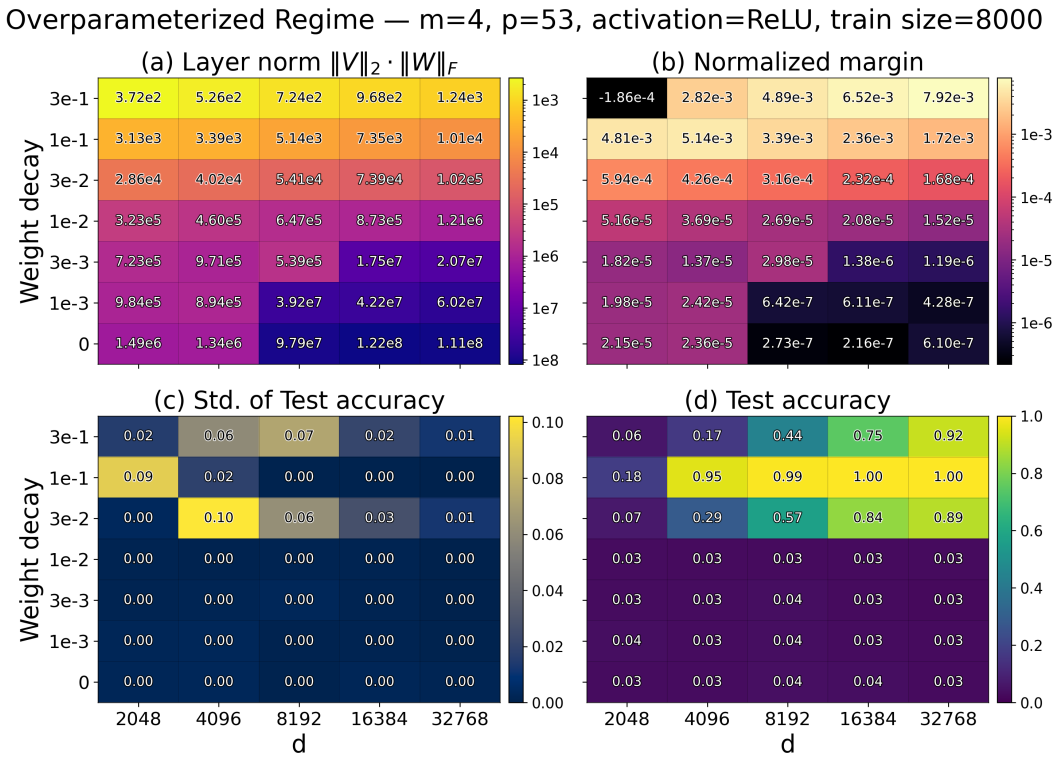


Figure 9: Two-layer ReLU networks in the overparameterized regime.

## C SOURCES FOR CAPACITY BOUNDS IN TABLE 1

**Notation and scope.** All entries are for two-layer MLPs (one hidden layer) with $W$ trainable parameters and width $d$. Throughout, $\widetilde{\Theta}(\cdot)$ hides polylogarithmic factors.

**Piecewise linear, real inputs ($\mathrm{VCdim} = \Theta(W \log W)$).** The nearly-tight bounds for piecewise-linear networks are summarized by (Bartlett et al., 2017b, Eq. (2)); for fixed depth $L = 2$ this specializes to $\Theta(W \log W)$.

**Piecewise-polynomial, real inputs ($\mathrm{VCdim} = \Theta(W \log W)$).** (Anthony & Bartlett, 2009, Thm. 8.8) prove an upper bound of $O(WL \log W + WL^2)$ for networks with piecewise-polynomial activations of bounded degree and a bounded number of pieces; in the depth-2 case this simplifies to $O(W \log W)$. A matching lower bound of $\Omega(W \log W)$ for two-layer linear-threshold networks (a special case with degree 0) appears in (Anthony & Bartlett, 2009, Thm. 6.4). Using a refined bit-extraction technique, (Bartlett et al., 2017b, Thm. 3) further gives an explicit construction achieving $\Omega(WL \log(W/L))$ for ReLU networks, which in particular gives $\Omega(W \log W)$ for depth-2 networks.

**Pfaffian activations (incl. standard sigmoid), real inputs ($\mathrm{VCdim} = \mathcal{O}(d^2W^2)$).** A general upper bound $\mathcal{O}\big(W^2k^2\big)$ for standard sigmoid networks is given in (Anthony & Bartlett, 2009, Thm. 8.13), where $k$ is the number of computation units; in a two-layer networks, $k = d$. The Pfaffian extension follows from Khovanskii's *Fewnomials*: the theorem underlying Lemma 8.15 (see also (Anthony & Bartlett, 2009, §8.6)) bounds the number of connected components (Betti numbers) of semi-Pfaffian sets defined by functions from a fixed Pfaffian chain. Plugging this component bound into the standard growth-function argument used for the exponential case yields the same $\mathcal{O}(d^2W^2)$ VC-dimension bound for networks whose activations lie in a fixed Pfaffian chain (with order/degree independent of the data).

**Standard sigmoid, real inputs ($\mathrm{VCdim} = \Omega(W \log W)$).** The reduction from linear-threshold to smooth sigmoids (Anthony & Bartlett, 2009, Thm. 6.5) implies that the two-layer linear-threshold lower bound (Anthony & Bartlett, 2009, Thm. 6.4) carries over to standard sigmoid networks on the same finite set of inputs. This yields $\Omega(W)$lower bound, and $\Omega(W \log W)$ under the construction of "bit extraction"(Bartlett et al., 2017b, Rmk. 4).

**Standard sigmoid, discrete bounded inputs ($\mathrm{VCdim} = \widetilde{\Theta}(W)$).** For two-layer standard sigmoid networks with discrete inputs and first-layer fan-in $\leq N$, (Anthony & Bartlett, 2009, Thm. 8.11) gives $\mathrm{VCdim} \leq 2W \log_2(60ND) = \widetilde{O}(W)$. The paragraph following the theorem constructs a two-layer linear-threshold network with $\mathrm{VCdim} = \Omega(W)$; (Anthony & Bartlett, 2009, Thm. 6.5) transfers this lower bound to sigmoids. Hence the bound is $\widetilde{\Theta}(W)$.

**Sine, discrete unbounded inputs ($\mathrm{VCdim} = \infty$).** (Anthony & Bartlett, 2009, Lemma 7.2) shows that $\{x \mapsto \mathrm{sgn}(\sin(ax))\}$ has infinite VC-dimension. Thus, restricting to two labels, the corresponding multiclass Natarajan-dimension is also infinite.

**Our Natarajan-dimension results.** The $\widetilde{O}(W)$ upper bounds are established in Theorems E.12, E.14, and E.13, for the piecewise, standard sigmoid, trigonometric, and rational–exponential cases. Moreover, Lemma E.17 shows that the Natarajan-dimension is at least the VC-dimension of the associated binary subclass,

$$\mathrm{VCdim}(\mathcal{M}) \;\leq\; \mathrm{Ndim}(\mathcal{H}),$$

thereby supplying matching lower bounds for the Natarajan-dimension. Because these Natarajan-dimension lower bounds are inherited from VC-dimension lower bounds, they are existential: they assert the existence of networks in the family attaining the stated bounds.

## D A PAC LOWER BOUND FOR MULTI-CLASS CLASSIFICATION

Define the *residue*

$$r(x) \;=\; \sum_{k=0}^{p-1} k\, x_k \pmod p \in [p].$$

A learner that already knows this rule needs essentially no data, as it can compute $r(x)$ from $x$ exactly.

We consider the label-symmetric setting: the learner does not know which output index corresponds to which residue class. Formally, the true rule is composed with an unknown permutation $\pi \in \mathbb{S}_p$. We assume

$$\pi \sim \mathrm{Unif}(\mathbb{S}_p) \quad \text{independently of the data generation,}$$

and define $f_\pi(x) = \pi(r(x))$. A (possibly randomized) learner $\mathcal{A}$ observes $n$ i.i.d. pairs $S = \{(X_i, Y_i)\}_{i=1}^n$ with $X_i$ distributed as above and $Y_i = f_\pi(X_i)$, and outputs a classifier $\widehat{f} = \mathcal{A}(S) : \mathcal{X} \to [p]$. Performance is measured by the population 0–1 risk

$$L(\widehat{f}; \pi) = \mathbb{P}_X\big(\widehat{f}(X) \neq f_\pi(X)\big),$$

where the probability is over an independent test draw $X$, and $\widehat{f}$ (hence $L(\widehat{f}; \pi)$) is random due to $S$, $\pi$, and any internal randomness of $\mathcal{A}$.

**Lemma D.1.** *If $X$ is generated as above, then $r(X) \sim \mathrm{Unif}([p])$. Hence the residues $r(X_1), \ldots, r(X_n)$ in the sample are i.i.d. uniform on $[p]$.*

*Proof.* Write $r(X) = \sum_{i=1}^m s_i \pmod p$. Since $s_1 \sim \mathrm{Unif}([p])$ is independent of $S' = \sum_{i=2}^m s_i \pmod p$, the sum $s_1 + S' \pmod p$ is uniform on the cyclic group $\mathbb{Z}_p$. □

**Lemma D.2.** *Fix any realized training set $S$. Let $R \subseteq [p]$ be the set of residues that appear among $r(X_1), \ldots, r(X_n)$, and let $U = [p] \setminus R$ with $K = |U|$. Conditional on $S$, the restriction $\pi|_U$ is a uniformly random bijection from $U$ to $[p] \setminus \pi(R)$.*

*Proof.* The prior on $\pi$ is uniform over $\mathbb{S}_p$, independent of the data. Observing $S$ reveals $\pi(u)$ for all $u \in R$ (because $r(X_i)$ is computable from $X_i$). Conditioning on these values, all completions of $\pi$ on $U$ are equally likely, and there are $K!$ of them. □

**Lemma D.3** (Risk lower bound via unseen residues)**.** *Let $K$ be the number of unseen residues determined by $S$. For any learner, over the random draw of the training samples,*

$$L(\widehat{f}; \pi) \;\geq\; \frac{(K-1)_+}{p} \qquad \textit{almost surely in } S,$$

*where $(t)_+ = \max\{t, 0\}$.*

*Proof.* Condition on a realized $S$ and its unseen set $U$ (size $K$). By Lemma D.1, $\mathbb{P}(r(X) = u) = 1/p$ for each $u \in [p]$. For any unseen $u \in U$, Lemma D.2 implies $\pi(u)$ is uniform over a set of $K$ labels, independent of $X$ given $r(X) = u$. Thus, for any prediction rule measurable with respect to $S$ and $X$, the success probability at residue $u$ is at most $1/K$, so the misclassification probability is at least $(K-1)/K$. Summing over $u \in U$,

$$L(\widehat{f}; \pi) \;\geq\; \sum_{u \in U} \mathbb{P}(r(X) = u) \cdot \frac{K-1}{K} = \frac{K}{p} \cdot \frac{K-1}{K} = \frac{K-1}{p},$$

□

**Lemma D.4.** *Let $K$ be the number of residues in $[p]$ not hit by $r(X_1), \ldots, r(X_n)$. Then*

$$\mathbb{E}[K] = p\Big(1 - \frac{1}{p}\Big)^n,$$
$$\mathrm{Var}(K) \leq \mathbb{E}[K],$$
$$\mathbb{P}(K \leq \mathbb{E}[K] - t) \leq \exp\left(-\frac{2t^2}{n}\right) \quad \textit{for all } t \geq 0.$$

*Proof.* Let $I_u = \mathbf{1}_{\{\text{residue } u \text{ is unseen}\}}$ for $u \in [p]$. By Lemma D.1, the $n$ residues are i.i.d. uniform, so

$$\mathbb{P}(I_u = 1) = \Big(1 - \frac{1}{p}\Big)^n =: q, \qquad \mathbb{P}(I_u = I_v = 1) = \Big(1 - \frac{2}{p}\Big)^n =: q_2 \quad (u \neq v).$$

Thus $\mathbb{E}[K] = \sum_u \mathbb{E}[I_u] = pq$. Moreover,

$$\mathrm{Var}(K) = \sum_u \mathrm{Var}(I_u) + \sum_{u \neq v} \mathrm{Cov}(I_u, I_v) = pq(1-q) + p(p-1)(q_2 - q^2).$$

Since $(1 - \frac{2}{p}) \leq (1 - \frac{1}{p})^2$, we have $q_2 \leq q^2$, hence $\mathrm{Var}(K) \leq pq(1-q) \leq pq = \mathbb{E}[K]$. For concentration, expose the independent residues $Z_i = r(X_i) \in [p]$. The mapping $(Z_1, \ldots, Z_n) \mapsto K$ is 1-Lipschitz (changing one residue can alter the number of unseen residues by at most 1), so McDiarmid's inequality yields the stated tail bound. □

**Lemma D.5** (A logarithmic inequality)**.** *For $x \in (0,1)$, $\log(1-x) \geq -\frac{x}{1-x}$. Hence, for integers $p \geq 2$ and all $n \geq 0$,*

$$\Big(1 - \frac{1}{p}\Big)^n \;\geq\; \exp\Big(-\frac{n}{p-1}\Big).$$

*Proof.* Define $h(x) = \log(1-x) + \frac{x}{1-x}$. Then $h'(x) = \frac{x}{(1-x)^2} \geq 0$ and $h(0) = 0$, so $h(x) \geq 0$ on $(0,1)$. □

**Theorem D.6** (PAC lower bound)**.** *Fix $\varepsilon \in (0, \frac{1}{2})$ and $\delta \in (0, \frac{1}{2})$. There exists an integer $p_0 = p_0(\varepsilon, \delta)$ such that for all $p \geq p_0$, every learner $\mathcal{A}$ that (with probability at least $1 - \delta$ over the draw of $S$ and the internal randomness) achieves population risk at most $\varepsilon$ must use*

$$n \;\geq\; (p-1)\left(\log\frac{1}{\varepsilon} - 1\right) = \Omega(p).$$

*Equivalently, for every $n \leq (p-1)\big(\log\frac{1}{\varepsilon} - 1\big)$ and every learner,*

$$\mathbb{P}\big(L(\widehat{f};\pi) \leq \varepsilon\big) \;\leq\; \exp\left(-\frac{(e-1)^2}{2\,\log(1/\varepsilon)}\,\varepsilon^2\,p\right), \tag{1}$$

*so $\mathbb{P}(L(\widehat{f};\pi) \leq \varepsilon) \leq \delta$ for all sufficiently large $p$.*

*Proof.* By Lemma D.3, the event $\{L(\widehat{f};\pi) \leq \varepsilon\}$ implies $\{K \leq \varepsilon p + 1\}$. Hence

$$\mathbb{P}\big(L(\widehat{f};\pi) \leq \varepsilon\big) \;\leq\; \mathbb{P}\big(K \leq \varepsilon p + 1\big).$$

Let $\mu = \mathbb{E}[K] = p(1 - \frac{1}{p})^n$. By Lemma D.5,

$$\mu \;\geq\; p\,\exp\Big(-\frac{n}{p-1}\Big).$$

Assume $n \leq (p-1)\big(\log\frac{1}{\varepsilon} - 1\big)$. Then

$$\mu \;\geq\; p\,\exp\Big(-\log\frac{1}{\varepsilon} + 1\Big) \;=\; p\,\varepsilon\,e.$$

Set $t = \mu - (\varepsilon p + 1) \geq p\varepsilon(e-1) - 1$. For all $p$ large enough (depending only on $\varepsilon$), $t \geq \frac{1}{2}\,p\varepsilon(e-1)$. By Lemma D.4 (McDiarmid),

$$\mathbb{P}\big(K \leq \varepsilon p + 1\big) \;=\; \mathbb{P}\big(K \leq \mu - t\big) \;\leq\; \exp\left(-\frac{2t^2}{n}\right) \;\leq\; \exp\left(-\frac{(e-1)^2}{2\,\log(1/\varepsilon)}\,\varepsilon^2\,p\right),$$

where in the last inequality we used $t \geq \frac{1}{2}p\varepsilon(e-1)$ and $n \leq (p-1)\log\frac{1}{\varepsilon} \leq p\log\frac{1}{\varepsilon}$. This proves equation 1. Therefore, the following statement holds since $\exp\Big(-\frac{(e-1)^2}{2\,\log(1/\varepsilon)}\,\varepsilon^2\,p\Big)$ decays exponentially in $p$. □

# E PROOFS IN NATARAJAN-DIMENSION

## E.1 UPPER BOUND OF NATARAJAN-DIMENSION

Let $\mathcal{X}$ be an instance space, let $p \in \mathbb{N}$ with $p \geq 2$, and let $[p] = \{1, \ldots, p\}$ be the label set. Fix a domain $\mathcal{U}$ and a function class $\mathcal{F}$. For a finite $T \subseteq \mathcal{U}$, write $\mathcal{F}_{|_T} := \{ f_{|_T} : f \in \mathcal{F} \}$ for its restriction and $|\mathcal{F}_{|_T}|$ for the number of distinct labelings on $T$ realized by $\mathcal{F}$.

**Definition E.1** (Number of realized multiclass labelings)**.** For $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ and hypothesis class $\mathcal{H} \subseteq [p]^{\mathcal{X}}$, define

$$\Lambda_{\mathcal{H}}(S) \;:=\; \big|\mathcal{H}_{|_S}\big| \;=\; \Big|\Big\{ \big(h(x^{(1)}), \ldots, h(x^{(n)})\big) \in [p]^n \;:\; h \in \mathcal{H} \Big\}\Big|.$$

**Lemma E.2** (Natarajan shattering and labelings)**.** *If a finite set $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ is Natarajan-shattered by a hypothesis $\mathcal{H} \subset [p]^{\mathcal{X}}$, then $\Lambda_{\mathcal{H}}(S) \geq 2^n$.*

*Proof of Lemma E.2.* By Definition 5.3, there exist $f_1, f_2 \in [p]^S$ with $f_1(x) \neq f_2(x)$ for all $x \in S$ such that for every selector $b : S \to \{1, 2\}$ there is $h_b \in \mathcal{H}$ with $h_b(x) = f_{b(x)}(x)$ for all $x \in S$. Define $\Phi : \{1, 2\}^S \to \mathcal{H}_{|_S}$ by $\Phi(b) = h_b|_S$.

If $b \neq b'$, pick $x_0 \in S$ with $b(x_0) \neq b'(x_0)$. Then

$$\Phi(b)(x_0) = h_b(x_0) = f_{b(x_0)}(x_0) \neq f_{b'(x_0)}(x_0) = h_{b'}(x_0) = \Phi(b')(x_0),$$

so $\Phi(b) \neq \Phi(b')$. Thus $\Phi$ is injective and

$$\Lambda_{\mathcal{H}}(S) = |\mathcal{H}_{|_S}| \;\geq\; |\Phi(\{1, 2\}^S)| = |\{1, 2\}^S| = 2^{|S|} = 2^n.$$

□

**Lemma E.3** (Labelings and pairwise reduction)**.** *Fix $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ and a hypothesis class $\mathcal{H}_\Theta \subseteq [p]^{\mathcal{X}}$. Then*

$$\Lambda_{\mathcal{H}_\Theta}(S) \;\leq\; \Pi_{\mathcal{G}_\Theta}\big(n\, p(p-1)/2\big).$$

*Proof of Lemma E.3.* Set

$$T \;:=\; S \times \{(i, j) \in [p] \times [p] :\; i < j\} \;\subset\; \mathcal{Z}_{\text{pair}}.$$

For each $h \in (\mathcal{H}_\Theta)_{|_S}$ define the fiber

$$W(h) := \{\theta \in \Theta :\; h_{\theta|_S} = h\}, \quad \text{and} \quad A(h) := \{\, g_{\theta|_T} :\; \theta \in W(h) \,\} \;\subseteq\; (\mathcal{G}_\Theta)_{|_T}.$$

Now we will show that if $h \neq h'$, then $A(h) \cap A(h') = \emptyset$.

Pick $x \in S$ with $h(x) = i$ and $h'(x) = j \neq i$. Without loss of generality, assume $i < j$. For any $\theta \in W(h)$, the tie-breaking rule implies $s_i^\theta(x) \geq s_j^\theta(x)$, hence $g_\theta(x, i, j) = +1$. For any $\theta' \in W(h')$, we have $s_j^{\theta'}(x) > s_i^{\theta'}(x)$, hence $g_{\theta'}(x, i, j) = -1$. Thus every element of $A(h)$ has $+1$ and every element of $A(h')$ has $-1$ at the coordinate $(x, i, j) \in T$, so $A(h) \cap A(h') = \emptyset$.

Since $|(\mathcal{H}_\Theta)_{|S}| \leq p^n < \infty$ and each $A(h) \neq \emptyset$, fix an arbitrary choice function $\Psi$ selecting one element of $A(h)$ for each $h \in (\mathcal{H}_\Theta)_{|_S}$. Then the map

$$\Psi :\; (\mathcal{H}_\Theta)_{|_S} \;\longrightarrow\; (\mathcal{G}_\Theta)_{|_T}, \qquad h \longmapsto \Psi(h)$$

is well-defined and injective. Therefore,

$$\Lambda_{\mathcal{H}_\Theta}(S) = |(\mathcal{H}_\Theta)_{|_S}| \;\leq\; |(\mathcal{G}_\Theta)_{|_T}| \;\leq\; \Pi_{\mathcal{G}_\Theta}(|T|) \;=\; \Pi_{\mathcal{G}_\Theta}\big(n\tbinom{p}{2}\big) \;=\; \Pi_{\mathcal{G}_\Theta}\big(n\, p(p-1)/2\big).$$

□

Together with Lemmas E.2 and E.3, we have Lemma 5.5.

**Definition E.4** (k-combination of $\operatorname{sgn}(\mathcal{F})$)**.** Let $\mathcal{Z}$ be any domain and let $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}^D \times \mathcal{Z}}$ be a class of real-valued functions of the form $(a, z) \mapsto f(a, z)$, with $a \in \mathbb{R}^D$ and $z \in \mathcal{Z}$. A binary class $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{Z}}$ is a *k-combination of* $\operatorname{sgn}(\mathcal{F})$ if there exist a Boolean map $g : \{-1, +1\}^k \to \{-1, +1\}$ and functions $f_1, \ldots, f_k \in \mathcal{F}$ such that for every $h \in \mathcal{H}$ there is $a \in \mathbb{R}^D$ with

$$h(z) \;=\; g\big(\operatorname{sgn}(f_1(a, z)), \ldots, \operatorname{sgn}(f_k(a, z))\big) \qquad \text{for all } z \in \mathcal{Z}.$$

We say $f \in \mathcal{F}$ is $C^D$ in its parameters if, for every fixed $z$, the map $a \mapsto f(a, z)$ is $C^D$.

**Definition E.5** (Regular zero-set intersections (Def. 7.3 (Anthony & Bartlett, 2009)))**.** For differentiable $f_1, \ldots, f_k : \mathbb{R}^D \to \mathbb{R}$, the family $\{f_1, \ldots, f_k\}$ has *regular zero-set intersections* if for every nonempty $I \subseteq \{1, \ldots, k\}$, the Jacobian of $(f_i)_{i \in I}$ has full row rank $|I|$ at every $a$ with $f_i(a) = 0$ for all $i \in I$.

**Definition E.6** (Solution set components bound (Def. 7.5 (Anthony & Bartlett, 2009)))**.** Let $\mathcal{G}$ be a set of real-valued functions on $\mathbb{R}^D$. We say $\mathcal{G}$ has *solution set components bound* $B$ if for any $1 \le k \le D$ and any $\{f_1, \ldots, f_k\} \subseteq \mathcal{G}$ that has regular zero-set intersections,

$$\operatorname{CC}\left(\bigcap_{i=1}^{k} \{\, a \in \mathbb{R}^D : f_i(a) = 0 \,\}\right) \;\le\; B,$$

where $\operatorname{CC}(\cdot)$ is the number of connected components.

**Theorem E.7** (General Growth function upper bound (Thm. 7.6 (Anthony & Bartlett, 2009)))**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}^D \times \mathcal{Z}}$ be closed under addition of constants, assume every $f \in \mathcal{F}$ is $C^D$ in $a$, and let*

$$\mathcal{G} \;:=\; \big\{\, a \mapsto f(a, z) : f \in \mathcal{F},\; z \in \mathcal{Z} \,\big\}.$$

*If $\mathcal{G}$ has a solution set components bound $B$ and $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{Z}}$ is a $k$-combination of $\operatorname{sgn}(\mathcal{F})$, then for all $N \ge D/k$,*

$$\Pi_{\mathcal{H}}(N) \;\le\; B \sum_{i=0}^{D} \binom{Nk}{i} \;\le\; B\Big(\tfrac{e\,Nk}{D}\Big)^D.$$

**Theorem E.8** (General Growth function upper bound (Thm. 8.3 (Anthony & Bartlett, 2009)))**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}^D \times \mathcal{Z}}$ be a class of functions mapping from $\mathbb{R}^D \times \mathcal{Z}$ to $\mathbb{R}$ such that, for all $z \in \mathcal{Z}$ and $f \in \mathcal{F}$, the map $a \mapsto f(a, z)$ is a polynomial on $\mathbb{R}^D$ of degree at most $r$. Suppose that $\mathcal{H}$ is a $k$-combination of $\operatorname{sgn}(\mathcal{F})$. Then, if $N \ge D/k$,*

$$\Pi_{\mathcal{H}}(N) \;\le\; 2\left(\frac{2eNkr}{D}\right)^D.$$

*Remark* E.9. If $N < D/k$, we have a trivial bound $\Pi_{\mathcal{H}}(N) \le 2^N < 2^{D/k} \le 2^D$, so for all $N \in \mathbb{N}$, $\Pi_{\mathcal{H}}(N) \le \max\left\{2^D, 2\left(\frac{2eNkr}{D}\right)^D\right\}$.

**Lemma E.10** (Absorbing $\log n$ (Lem. A.2 (Shalev-Shwartz & Ben-David, 2014)))**.** *Let $A \ge 1$, $B \ge 0$, and $u > 0$. If $u < A \log u + B$, then*

$$u \;<\; 4A\,\log\big(2A\big) \;+\; 2B.$$

**Lemma E.11** (Trigonometric Sum Polynomialization)**.** *Let $p \ge 1$ and $m \ge 0$ be integers. For any vector of non-negative integers $x = (x_1, \ldots, x_p)$ such that $\sum_{v=1}^{p} x_v = m$, there exist polynomials*

$$S_x, C_x \in \mathbb{Z}[c_1, s_1, \ldots, c_p, s_p]$$

*of total degree at most $m$ that satisfy*

$$S_x(c_1, s_1, \ldots, c_p, s_p) = \sin\Big(\sum_{v=1}^{p} x_v \alpha_v\Big) \quad \textit{and} \quad C_x(c_1, s_1, \ldots, c_p, s_p) = \cos\Big(\sum_{v=1}^{p} x_v \alpha_v\Big)$$

*for all real angles $\alpha_1, \ldots, \alpha_p$, where $c_v := \cos(\alpha_v)$ and $s_v := \sin(\alpha_v)$.*

*Proof of Lemma E.11.* We prove the lemma by induction on $m = \sum_{v=1}^{p} x_v$. The uniqueness of the polynomials is guaranteed by the deterministic recursive construction.

Base Case ($m = 0$):
If $m = 0$, then $x = \mathbf{0}$ is the only possible vector. The sum of angles is $\sum x_v \alpha_v = 0$. The defined polynomials are $S_{\mathbf{0}} = 0$ and $C_{\mathbf{0}} = 1$. These are integer-coefficient polynomials of degree 0. They correctly evaluate to $\sin(0) = 0$ and $\cos(0) = 1$.

Inductive Step:
Assume the claim holds for all vectors $y$ with component sum $m - 1$. Let $x$ be a vector with component sum $m$. Let $u = \min\{v \mid x_v > 0\}$ and define $y = x - e_u$. The components of $y$ sum to $m - 1$. By the induction hypothesis, there exist polynomials $S_y$ and $C_y$ with integer coefficients and degree at most $m - 1$ that represent $\sin(\sum y_v \alpha_v)$ and $\cos(\sum y_v \alpha_v)$.

We define $S_x$ and $C_x$ as per the recursion:

$$S_x := s_u C_y + c_u S_y \qquad C_x := c_u C_y - s_u S_y$$

1. Coefficients and Degree: Since $S_y$ and $C_y$ have integer coefficients, and $s_u, c_u$ are variables, $S_x$ and $C_x$ are also polynomials with integer coefficients. Their total degrees are bounded by:

$$\deg(S_x) \le 1 + \max(\deg(C_y), \deg(S_y)) \le 1 + (m - 1) = m$$

   The same bound holds for $\deg(C_x)$.

2. Trigonometric Identity: By the angle addition formulas and the induction hypothesis:

$$\begin{aligned} S_x &= \sin(\alpha_u) \cos\Big(\sum_{v=1}^{p} y_v \alpha_v\Big) + \cos(\alpha_u) \sin\Big(\sum_{v=1}^{p} y_v \alpha_v\Big) \\ &= \sin\Big(\alpha_u + \sum_{v=1}^{p} y_v \alpha_v\Big) = \sin\Big(\sum_{v=1}^{p} x_v \alpha_v\Big) \end{aligned}$$

   Similarly,

$$\begin{aligned} C_x &= \cos(\alpha_u) \cos\Big(\sum_{v=1}^{p} y_v \alpha_v\Big) - \sin(\alpha_u) \sin\Big(\sum_{v=1}^{p} y_v \alpha_v\Big) \\ &= \cos\Big(\alpha_u + \sum_{v=1}^{p} y_v \alpha_v\Big) = \cos\Big(\sum_{v=1}^{p} x_v \alpha_v\Big) \end{aligned}$$

This completes the induction. □

#### E.1.1 PIECEWISE-POLYNOMIAL ACTIVATIONS

**Theorem E.12** (Two-layer piecewise-polynomial activations)**.** *Let $\sigma$ be as in Definition 5.6. For the two-layer MLP defined in the model setup,*

$$\mathrm{Ndim}(\mathcal{H}_\Theta) \le 2dp\,(6\log(6dp) + \log(2eL) + 2\log(epr)) = \widetilde{O}(dp)$$

*Proof of Theorem E.12.* Let $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ be Natarajan-shattered. By Lemma 5.5, this implies $2^n \le \Pi_{\mathcal{G}_\Theta}\big(n\binom{p}{2}\big)$. The parameter space $W \in \mathbb{R}^{dp}$ is partitioned into regions by the zero sets of $\{w_i x^{(j)} - b_\ell\}$ for $j \in [n], i \in [d], \ell \in [L-1]$. The number of regions, $R_S$, is the number of sign patterns on a sample of size $m = nd(L-1)$ by affine functions of $W$.

Notice that $\big\{w \mapsto \mathrm{sgn}\,(wx - b_\ell) : w \in \mathbb{R}^{1\times p}, x \in S, \ell \in [L-1]\big\} \subset \{-1, +1\}^{\mathcal{X}}$ is a 1-combination of $\mathrm{sgn}\,\big(\big\{w \mapsto (wx - b_\ell) : w \in \mathbb{R}^{1\times p}, x \in S, \ell \in [L-1]\big\} \subset \mathbb{R}^{\mathcal{X}}\big)$.

By Theorem E.8, $R_S \le \max\left\{2^{dp}, 2\left(\frac{2e\cdot nd(L-1)}{dp}\right)^{dp}\right\}$. Within each region, $s_i^\theta(x) - s_j^\theta(x)$ is a polynomial in $\theta \in \mathbb{R}^{2dp}$ of degree at most $r + 1$. Let $N = n\binom{p}{2}$, $D = 2dp$. The growth function is

bounded by the product of the number of regions and the maximum growth function within a region. Applying Theorem E.8 in each region:

$$\Pi_{\mathcal{G}_\Theta}(N) \le R_S \cdot \max_R \Pi_R(N) \le \max\left\{2^{dp}, 2\left(\frac{2e\cdot nd(L-1)}{dp}\right)^{dp}\right\} \max\left\{2^{2dp}, 2\left(\frac{eN(r+1)}{dp}\right)^{2dp}\right\}$$

Substituting $N = \frac{np(p-1)}{2}$ into the inequality $2^n \le \Pi_{\mathcal{G}_\Theta}(N)$ gives $2^n \le (2enL)^{dp}\,(enpr)^{2dp}$. Taking the logarithm of both sides yields

$$n \le 3dp\log(n)/\log(2) + dp(\log(2eL) + 2\log(epr))/\log(2).$$

Lemma E.10 implies that

$$n \le 2dp\,(6\log(6dp) + \log(2eL) + 2\log(epr))\,/\log(2) = \widetilde{O}(dp).$$

Take supreum over $S$ yields the result.

□

### E.1.2 Trigonometric-Polynomial Activations

**Theorem E.13** (Two-layer trigonometric-polynomial activations)**.** *Let $\sigma$ be as in Definition 5.7. For the two-layer MLP from the model setup,*

$$\mathrm{Ndim}(\mathcal{H}_\Theta) \le 2dp\Big(6\log(6dp) + 2\log\big(ep(Km+1)\big)\Big) = \widetilde{O}(dp).$$

*Proof of Theorem E.13.* Let $S = \{x^{(1)}, \dots, x^{(n)}\} \subset \mathcal{X}$ be Natarajan-shattered. By Lemma 5.5, this implies $2^n \le \Pi_{\mathcal{G}_\Theta}\big(n\binom{p}{2}\big)$.

For $j \in [d], v \in [p]$ set $c_{j,v} := \cos(w_{j,v})$ and $s_{j,v} := \sin(w_{j,v})$, and regard

$$a := \big(V, (c_{j,v})_{j,v}, (s_{j,v})_{j,v}\big) \in \mathbb{R}^{3dp}$$

as the (relaxed) parameter vector; ignoring the constraints $c_{j,v}^2 + s_{j,v}^2 = 1$ can only increase the growth function. For any $(i, y \ne y')$,

$$s_y^\theta(x^{(i)}) - s_{y'}^\theta(x^{(i)}) = \sum_{j=1}^{d}(V_{yj} - V_{y'j})\sigma(\langle w_j, x^{(i)}\rangle).$$

Writing $\sigma$ as in Definition 5.7 and applying Lemma E.11 to $kx^{(i)}$ shows that each term $\cos(k\langle w_j, x^{(i)}\rangle)$ and $\sin(k\langle w_j, x^{(i)}\rangle)$ is a polynomial in $\big((c_{j,v})_v, (s_{j,v})_v\big)$ of degree at most $km \le Km$. Hence every pairwise margin is a polynomial in $a$ of degree at most $Km+1$.

The reduction class $\mathcal{G}_\Theta$ is a 1-combination of $\mathrm{sgn}(\mathcal{F})$ with $\mathcal{F}$ being a family of polynomials of degree at most $Km+1$ in $D = 3dp$ parameters. Applying Theorem E.8 with $N = n\binom{p}{2}$, $k = 1$, $r = Km+1$,

$$\Pi_{\mathcal{G}_\Theta}(N) \le \max\Big\{2^D, 2\Big(\frac{2eNr}{D}\Big)^D\Big\} \le 2\,(pKnm)^{3dp}.$$

Combine with $2^n \le \Pi_{\mathcal{G}_\Theta}(N)$, take logs: $n\log(2) \le \log(2) + 3dp\log(pKm) + 3dp\log(n)$. Use Lemma E.10 to absorb the $\log n$ term, yielding

$$n \le 12dp\log(6dp)/\log(2) + 2\,(\log(2) + 3dp\log(pKm))\,/\log(2) = \widetilde{O}\,(dp)$$

Taking the supremum over shattered $S$ gives the claim. □

### E.1.3 Rational-Exponential Activations

**Theorem E.14** (Two-layer polynomial–rational–exponential activations)**.** *Let $\sigma$ be as in Definition 5.8. For the two-layer MLP from the model setup,*

$$\mathrm{Ndim}(\mathcal{H}_\Theta) \le 2dp\Big(6\log(6dp) + 2\log\big(ep(dm+r+1)\big)\Big) = \widetilde{O}(dp).$$

*Proof of Theorem E.14.* Let $S = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ be Natarajan-shattered and set $N := n\binom{p}{2}$. For each example $i$ and hidden unit $j$, put $z_{j,i} := e^{k\langle w_j, x^{(i)}\rangle} > 0$. Since $c \geq 0$ and $\tau > 0$, the product

$$D_i(W) := \prod_{j=1}^{d} \big(cz_{j,i} + \tau\big) > 0.$$

Multiplying any pairwise margin $G_{i,y,y'} := s_y^\theta(x^{(i)}) - s_{y'}^\theta(x^{(i)})$ by $D_i(W)$ preserves its sign and yields

$$\widehat{G}_{i,y,y'}(W,V) = D_i(W)G_{i,y,y'}(W,V) = \sum_{j=1}^{d}(V_{yj} - V_{y'j})P(\langle w_j, x^{(i)}\rangle)(az_{j,i} + b)\prod_{\ell\neq j}(cz_{\ell,i} + \tau).$$

Introduce relaxed variables $u_{j,v} := e^{kw_{j,v}} \in (0,\infty)$. Then

$$z_{j,i} = e^{k\langle w_j, x^{(i)}\rangle} = \prod_{v=1}^{p} u_{j,v}^{x_v^{(i)}},$$

a monomial of total degree $m$ in $U_j := (u_{j,1}, \ldots, u_{j,p})$. Consequently, each summand in $\widehat{G}_{i,y,y'}$ is a product of: (i) a linear term in $V$; (ii) the degree-$r$ polynomial $P(\langle w_j, x^{(i)}\rangle)$ in $W_j$; (iii) a factor $(az_{j,i}+b)\prod_{\ell\neq j}(cz_{\ell,i}+\tau)$ of total degree $dm$ in the $U$-variables. Thus every $\widehat{G}_{i,y,y'}$ is a polynomial in

$$a := \big(V, (u_{j,v})_{j,v}, (w_{j,v})_{j,v}\big) \in \mathbb{R}^{3dp}$$

of degree at most $\rho := dm + r + 1$. Treating $a$ as the parameter vector, the reduction class $\mathcal{G}_\Theta$ is a 1-combination of $\operatorname{sgn}(\mathcal{F})$ with $\mathcal{F}$ being a family of polynomials of degree at most $\rho$ in $D = 3dp$ parameters. Applying Theorem E.8 with $k = 1$, $D = 3dp$, $N = n\binom{p}{2}$,

$$\Pi_{\mathcal{G}_\Theta}(N) \leq \max\Big\{2^D, 2\Big(\frac{2eN\rho}{D}\Big)^D\Big\} \leq (np(dm + r + 1))^{3dp}.$$

Combine with Lemma 5.5 and absorb the $\log n$ term via Lemma E.10 to obtain

$$n \leq 12dp\log(6dp/\log(2))/\log(2) + 6dp\log(p(dm + r + 1))/\log(2) = \widetilde{O}(dp)$$

Taking the supremum over shattered $S$ gives the claim. □

### E.1.4 Uniform Convergence Guarantees

Let $\mathcal{H}_\sigma \subseteq [p]^{\mathcal{X}}$ be a multiclass hypothesis class with Natarajan-dimension $\operatorname{Ndim}(\mathcal{H}_\sigma) < \infty$. Let $h \in \mathcal{H}$, denote by $\mathbb{P}_{(x,y)\in\mathcal{D}}(h(x) \neq y)$ the population 0–1 risk and by $\mathbb{P}_{(x,y)\in\mathcal{D}_{\text{train}}}(h(x) \neq y)$ the empirical 0–1 risk computed from an i.i.d. sample of size $n$.

**Theorem E.15** (Thm. 29.3 of (Shalev-Shwartz & Ben-David, 2014), Uniform convergence)**.** *There exists a universal constant $C > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{h\in\mathcal{H}_\sigma}\big|\mathbb{P}_{(x,y)\in\mathcal{D}}(h(x) \neq y) - \mathbb{P}_{(x,y)\in\mathcal{D}_{train}}(h(x) \neq y)\big| \;\leq\; C\sqrt{\frac{\operatorname{Ndim}(\mathcal{H}_\sigma)\log p + \log(1/\delta)}{n}}.$$

*Proof of Theorem 5.9.* By Theorem E.12, E.13, E.14, $\operatorname{Ndim}(\mathcal{H}_\sigma) = \widetilde{O}(dp)$. Substituting this into the multiclass uniform convergence bound (Theorem E.15) yields

$$\sup_{h\in\mathcal{H}_\sigma}\Big|\mathbb{P}_{(x,y)\sim\mathcal{D}}\big(h(x) \neq y\big) - \mathbb{P}_{(x,y)\sim\mathcal{D}_{\text{train}}}\big(h(x) \neq y\big)\Big| \;\leq\; C\sqrt{\frac{\operatorname{Ndim}(\mathcal{H}_\sigma)\log p + \log(1/\delta)}{n}} \;=\; \widetilde{O}\left(\sqrt{\frac{dp + \log(1/\delta)}{n}}\right),$$

where the $\log p$ factor is absorbed into $\widetilde{O}(\cdot)$. □

### E.2 Lower Bound of Natarajan-dimension

Let $\Theta$ denote the backbone parameter space determined by the architecture. The multiclass hypothesis class is

$$\mathcal{H} = \{ h_{\theta,V} \ : \ \theta \in \Theta,\ V \in \mathbb{R}^{p\times d} \}.$$

**Definition E.16** (Associated binary class)**.** We consider the binary (realizable) subclass of halfspaces in the representation:

$$\mathcal{M} \ = \ \Big\{ x \mapsto \mathbf{1}\big\{\langle v,\ f_\theta(x)\rangle \geq 0\big\} \ : \ (\theta, v) \in \Theta \times \mathbb{R}^d \Big\} \ \subseteq \ \{0,1\}^{\mathcal{X}}.$$

**Lemma E.17** (VC-dimension is bounded by the Natarajan-dimension)**.** *For the multiclass hypothesis class $\mathcal{H}$ with $p \geq 2$,*

$$\mathrm{VCdim}\big(\mathcal{M}\big) \ \leq \ \mathrm{Ndim}(\mathcal{H}).$$

*Proof.* Let $S \subseteq \mathcal{X}$ be a finite set that is VC-shattered by $\mathcal{M}$.

Fix $f_1, f_2 \in [p]^S$ by $f_1(x) \equiv 1$ and $f_2(x) \equiv 2$ for all $x \in S$ (possible since $p \geq 2$). Let $b : S \to \{1, 2\}$ be an arbitrary selector. Define the induced binary labeling

$$y_b(x) \ = \ \mathbf{1}\{b(x) = 1\} \ \in \ \{0,1\} \qquad (x \in S).$$

Since $S$ is VC-shattered by $\mathcal{M}$, there exist $\theta_b \in \Theta$ and $v_b \in \mathbb{R}^d$ such that

$$y_b(x) \ = \ \mathbf{1}\big\{\langle v_b, f_{\theta_b}(x)\rangle \geq 0\big\} \quad \text{for all } x \in S.$$

Construct $V_b \in \mathbb{R}^{p\times d}$ such that the first row is $v_b$, and all remaining rows are $0$. Then, for each $x \in S$,

$$h_{\theta_b,V_b}(x) \ = \ \begin{cases} 1, & \text{if } y_b(x) = 1 \ (b(x) = 1), \\ 2, & \text{if } y_b(x) = 0 \ (b(x) = 2), \end{cases}$$

i.e., $h_{\theta_b,V_b}(x) = f_{b(x)}(x)$ for all $x \in S$. Since $b$ was arbitrary, $S$ is Natarajan-shattered by $\mathcal{H}$ with witnesses $(f_1, f_2)$. Therefore $|S| \leq \mathrm{Ndim}(\mathcal{H})$. Taking the supremum over all such $S$ gives $\mathrm{VCdim}(\mathcal{M}) \leq \mathrm{Ndim}(\mathcal{H})$.

□

## F Construction of interpolation solutions

### F.1 Sine Activation $\sigma(z) = \sin z$

In this section, we provide interpolating solutions for both two-layer sine and ReLU MLPs.

#### F.1.1 A low width construction

*Proof of Theorem 4.1.* Activation is $\sigma(z) = \sin z$. Let $S = \sum_{i=1}^m s_i$ and $\phi = \frac{2\pi}{p}$.

**First Layer Weights** $W \in \mathbb{R}^{2\times p}$. For each input coordinate $r \in \{0, \ldots, p-1\}$,

$$W_{1,r} = \big(\phi\, r\big) \bmod 2\pi \in [-\pi, \pi), \qquad W_{2,r} = \big(\phi\, r + \tfrac{\pi}{2m}\big) \bmod 2\pi \in [-\pi, \pi).$$

Then the pre-activations are

$$(Wx)_1 = \phi S, \qquad (Wx)_2 = \phi S + \tfrac{\pi}{2},$$

so the hidden units are

$$\sigma\big((Wx)_1\big) = \sin(\phi S), \qquad \sigma\big((Wx)_2\big) = \cos(\phi S).$$

**Second Layer Weights** $V \in \mathbb{R}^{p\times 2}$. For each class $q \in \{0, \ldots, p-1\}$,

$$V_{q,1} = \sin(\phi q), \qquad V_{q,2} = \cos(\phi q).$$

**Verification** For the $q$-th output,

$$\begin{aligned} s_q^\theta(x) &= V_{q,1}\sin(\phi S) + V_{q,2}\cos(\phi S) \\ &= \sin(\phi q)\sin(\phi S) + \cos(\phi q)\cos(\phi S) \\ &= \cos\big(\phi(S-q)\big). \end{aligned}$$

Thus $h_\theta(x) = \arg\max_q s_q^\theta(x) = S \bmod p$. This construction achieves $100\%$ accuracy with width $d = 2$, margin $\gamma = \Omega(1/p^2)$, and satisfies $\|W\|_\infty \le \pi$, $\|V\|_\infty \le 1$. □

### F.1.2 A HIGH MARGIN CONSTRUCTION

**Theorem F.1** (High margin, $d_0 = 2p$)**.** *There exists a construction with hidden dimension $d_0 = 2p$ and sine activation computes $\sum_{i=1}^m s_i \bmod p$ for all $x = (s_1, \cdots, s_m) \in \mathcal{X}$, achieving margin $\gamma = p$ and $\|V\|_2 = \sqrt{p}$, $\|W\|_F \le \pi\sqrt{2}p$.*

*Proof.* Index hidden units by $h \in \{1, \ldots, 2p\}$ and group them as $(2k-1, 2k)$ for frequencies $k \in \{1, \ldots, p\}$. Let $\phi_k = \frac{2\pi k}{p}$.

**First Layer Weights** $W \in \mathbb{R}^{2p\times p}$. For each $k \in \{1, \ldots, p\}$ and $r \in \{0, \ldots, p-1\}$,

$$W_{2k-1,r} = \big(\phi_k r\big) \bmod 2\pi \in [-\pi, \pi), \qquad W_{2k,r} = \big(\phi_k r + \tfrac{\pi}{2m}\big) \bmod 2\pi \in [-\pi, \pi).$$

Then

$$(Wx)_{2k-1} = \phi_k S, \qquad (Wx)_{2k} = \phi_k S + \tfrac{\pi}{2},$$

and

$$\sigma\big((Wx)_{2k-1}\big) = \sin(\phi_k S), \qquad \sigma\big((Wx)_{2k}\big) = \cos(\phi_k S).$$

The first layer satisfies $\|W\|_\infty \le \pi$, hence $\|W\|_F \le \pi\sqrt{(2p)p} = \pi\sqrt{2}\,p$.

**Second Layer Weights** $V \in \mathbb{R}^{p\times 2p}$. For each $q \in \{0, \ldots, p-1\}$ and $k \in \{1, \ldots, p\}$,

$$V_{q,\,2k-1} = \sin(\phi_k q), \qquad V_{q,\,2k} = \cos(\phi_k q).$$

**Verification** For the $q$-th output,

$$\begin{aligned} s_q^\theta(x) &= \sum_{k=1}^{p}\Big[\sin(\phi_k q)\sin(\phi_k S) + \cos(\phi_k q)\cos(\phi_k S)\Big] \\ &= \sum_{k=1}^{p}\cos\big(\phi_k(S-q)\big) = \Re\left(\sum_{k=1}^{p} e^{\,i\frac{2\pi k}{p}(S-q)}\right) \\ &= \begin{cases} p, & S \equiv q \pmod p, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Hence $h_\theta(x) = S \bmod p$ with margin $\gamma = p$. The construction achieves $100\%$ accuracy with width $d = 2p$ and satisfies $\|W\|_\infty \le \pi$, $\|V\|_\infty \le 1$. □

**Lemma F.2** (Singular values of $V$)**.** *In the high-margin construction, all singular values of $V$ are exactly $\sqrt{p}$, so $\|V\|_2 = \sqrt{p}$.*

*Proof.* Compute $VV^\top$ entrywise. For $q, r \in \{0, \dots, p-1\}$,

$$\begin{aligned}(VV^\top)_{qr} &= \sum_{k=1}^{p} \Big( \sin(\phi_k q)\sin(\phi_k r) + \cos(\phi_k q)\cos(\phi_k r) \Big) \\ &= \sum_{k=1}^{p} \cos\big(\phi_k (q-r)\big) \qquad \text{(by } \cos(a-b) = \cos a \cos b + \sin a \sin b\text{)} \\ &= \Re\left( \sum_{k=1}^{p} e^{\, i \frac{2\pi k}{p} (q-r)} \right) \\ &= \begin{cases} 0, & \text{if } q - r \not\equiv 0 \mod p \\ p, & \text{if } q - r \equiv 0 \mod p \end{cases}\end{aligned}$$

Therefore all eigenvalues of $VV^\top$ are exactly $p$, so all singular values of $V$ are exactly $\sqrt{p}$. In particular, the spectral norm is $\|V\|_2 = \sqrt{p}$. □

### F.2 ReLU Activation $\sigma(z) = \mathrm{ReLU}(z)$

For a multi-index $a = (a_1, \dots, a_s) \in \mathbb{Z}^s_{\geq 0}$, denote $|a| := \sum_{i=1}^s a_i$.

**Lemma F.3** (Polynomial sign polarization)**.** *For $s \geq 1$,*

$$x_1 x_2 \cdots x_s = \frac{1}{s!\, 2^s} \sum_{\varepsilon \in \{\pm 1\}^s} \Big( \prod_{i=1}^{s} \varepsilon_i \Big) \Big( \sum_{i=1}^{s} \varepsilon_i x_i \Big)^s.$$

*Proof.* Multinomial expansion gives

$$\Big( \sum_{i=1}^{s} \varepsilon_i x_i \Big)^s = \sum_{|k|=s} \frac{s!}{k_1! \cdots k_s!} \prod_{i=1}^{s} (\varepsilon_i x_i)^{k_i}.$$

Multiplying by $\prod_{j=1}^s \varepsilon_j$ and summing over $\varepsilon$ gives

$$\sum_{\varepsilon \in \{\pm 1\}^s} \Big( \prod_{j=1}^{s} \varepsilon_j \Big) \Big( \sum_{i=1}^{s} \varepsilon_i x_i \Big)^s = \sum_{|k|=s} \frac{s!}{k_1! \cdots k_s!}\, x_1^{k_1} \cdots x_s^{k_s} \sum_{\varepsilon \in \{\pm 1\}^s} \prod_{i=1}^{s} \varepsilon_i^{k_i+1}.$$

Observe that

$$\sum_{\varepsilon \in \{\pm 1\}^s} \prod_{i=1}^{s} \varepsilon_i^{k_i+1} = \prod_{i=1}^{s} \big( (+1)^{k_i+1} + (-1)^{k_i+1} \big) = \begin{cases} 2^s, & \text{if each } k_i + 1 \text{ is even,} \\ 0, & \text{otherwise.} \end{cases}$$

Because $|k| = s$ and each $k_i \geq 1$ must be odd so that $\sum_{\varepsilon \in \{\pm 1\}^s} \prod_{i=1}^s \varepsilon_i^{k_i+1} \neq 0$, the only possibility is $k = (1, \dots, 1)$.

Therefore,

$$\sum_{\varepsilon \in \{\pm 1\}^s} \Big( \prod_{i=1}^{s} \varepsilon_i \Big) \Big( \sum_{i=1}^{s} \varepsilon_i x_i \Big)^s = s! 2^s x_1 x_2 \cdots x_s.$$

Dividing by $s!\, 2^s$ yields the stated identity.

□

**Lemma F.4** (Uniform ReLU–spline approximation of power functions)**.** *Let $s \geq 1$, and $\varepsilon > 0$. Partition $[-1, 1]$ uniformly with knots $z_k = -1 + \frac{2k}{N}$, $k = 0, 1, \dots, N$. Let $g$ be the linear spline that interpolates $f_s(z) = z^s$ at these knots. Then*

$$\|f_s - g\|_{L_\infty([-1,1])} \;\leq\; \frac{s(s-1)}{2N^2}.$$

*Moreover, $g$ admits an exact one-hidden-layer ReLU representation on $[-1,1]$ of the form*

$$\Phi_s(z) = \sum_{i=1}^{M} c_i \,\mathrm{ReLU}(a_i z - b_i),$$

*with at most $M \le N+1$ units and*

$$|a_i| \le 1, \qquad |b_i| \le 1, \qquad |c_i| \le \max\left\{ s + \tfrac{1}{2},\ \frac{2\,s(s-1)}{N} \right\}.$$

*Choosing*

$$N \ \ge\ \max\left\{1,\ \left\lceil \sqrt{\frac{s(s-1)}{2\varepsilon}} \right\rceil \right\}$$

*ensures $\|f_s - g\|_{L_\infty([-1,1])} \le \varepsilon$. Thus, the number of required ReLU units to achieve accuracy $\varepsilon$ is $M = O\Big(\frac{s}{\sqrt{\varepsilon}}\Big)$.*

*Proof.* The case $s = 1$ is trivial since $f_1(z) = z$ is linear and equals its linear spline interpolant.

For $s \ge 2$, $f_s \in C^2([-1,1])$ with $f_s''(z) = s(s-1)z^{\,s-2}$ and $\|f_s''\|_{L_\infty([-1,1])} = s(s-1)$. Fix $z \in [z_k, z_{k+1}]$ and define

$$\varphi(t) = f_s(t) - g(t) - \frac{f_s(z) - g(z)}{(z - z_k)(z - z_{k+1})}(t - z_k)(t - z_{k+1}).$$

Then $\varphi(z_k) = \varphi(z_{k+1}) = \varphi(z) = 0$ and, by Rolle's theorem, there exists $\xi_z \in (z_k, z_{k+1})$ such that

$$|f_s(z) - g(z)| = \left| \frac{f_s''(\xi_z)}{2}\,(z - z_k)(z_{k+1} - z) \right|.$$

Hence, with $h = \frac{2}{N}$,

$$\max_{z \in [z_k, z_{k+1}]} |f_s(z) - g(z)| \le \tfrac{1}{2}\,\|f_s''\|_{L_\infty}\,\frac{h^2}{4} = \frac{s(s-1)}{2N^2}.$$

Taking the maximum over $k$ yields the stated uniform bound.

For the ReLU representation, write $h = \frac{2}{N}$ and set the interval slopes

$$m_k = \frac{z_{k+1}^s - z_k^s}{h}, \quad k = 0, \dots, N-1, \qquad \gamma_j = m_j - m_{j-1} = \frac{z_{j+1}^s - 2z_j^s + z_{j-1}^s}{h}, \quad j = 1, \dots, N-1.$$

Then $g$ admits the exact expansion on $[-1,1]$:

$$g(z) = c_1\,\mathrm{ReLU}(z+1) + c_2\,\mathrm{ReLU}(1-z) + \sum_{j=1}^{N-1} \gamma_j\,\mathrm{ReLU}(z - z_j),$$

with

$$c_2 = \frac{f_s(-1)}{2} = \frac{(-1)^s}{2}, \qquad c_1 = m_0 + \frac{(-1)^s}{2},$$

and $(a_1, b_1) = (1, -1)$, $(a_2, b_2) = (-1, -1)$, $(a_{j+2}, b_{j+2}) = (1, z_j)$ for $j = 1, \dots, N-1$.

Since $|z_j| \le 1$, we have $|a_i| \le 1$ and $|b_i| \le 1$. By the mean value theorem, $|m_0| \le \|f_s'\|_{L_\infty} = s$, hence $|c_1| \le s + \frac{1}{2}$ and $|c_2| \le \frac{1}{2} \le s + \frac{1}{2}$.

Moreover, define $\Psi \in C^2[z_j - h, z_j + h]$ where

$$\Psi(t) = f_s(t) - \left( f_s(z_j) + \frac{f_s(z_j + h) - f_s(z_j - h)}{2h}\,(t - z_j) + \frac{f_s(z_j + h) - 2f_s(z_j) + f_s(z_j - h)}{2h^2}\,(t - z_j)^2 \right).$$

By Rolle's theorem,

$$\gamma_j = \frac{f_s(z_j + h) - 2f_s(z_j) + f_s(z_j - h)}{h} = h\,f_s''(\xi_j) \quad \text{for some } \xi_j \in (z_j - h, z_j + h),$$

so $|\gamma_j| \le h\,\|f_s''\|_{L_\infty} = \frac{2}{N}\,s(s-1)$. Counting two boundary hinges and $N-1$ interior hinges gives $M \le N + 1$ units.

□

**Lemma F.5** (Polarized Newton expansion for $f_{\cos}$ and $f_{\sin}$)**.** *Let $m \geq 1$. For angles $(\theta_1, \ldots, \theta_m)$, define*

$$C_k = \sum_{i=1}^{m} \cos(k\theta_i), \qquad S_k = \sum_{i=1}^{m} \sin(k\theta_i).$$

*Let*

$$\mathcal{K}_m := \Big\{ k = (k_1, \ldots, k_m) \in \mathbb{Z}_{\geq 0}^m : \sum_{j=1}^{m} j\, k_j = m \Big\}.$$

*We index*

$$\varepsilon = (\varepsilon_{1,1}, \ldots, \varepsilon_{1,k_1}, \varepsilon_{2,1}, \ldots, \varepsilon_{m,k_m}) \in \{\pm 1\}^{|k|}.$$

*For $p = (p_1, \ldots, p_m)$ we denote $p \leq k$ if $0 \leq p_j \leq k_j$. Then*

$$f_{\cos} = \cos\Big(\sum_{i=1}^{m} \theta_i\Big) = \sum_{k \in \mathcal{K}_m} \sum_{\substack{p \leq k \\ |p| - |k| \text{ even}}} \sum_{\varepsilon \in \{\pm 1\}^{|k|}} \alpha_{k,p,\varepsilon} G_{k,p,\varepsilon}^{|k|}$$

$$f_{\sin} = \sin\Big(\sum_{i=1}^{m} \theta_i\Big) = \sum_{k \in \mathcal{K}_m} \sum_{\substack{p \leq k \\ |p| - |k| \text{ odd}}} \sum_{\varepsilon \in \{\pm 1\}^{|k|}} \beta_{k,p,\varepsilon} G_{k,p,\varepsilon}^{|k|}$$

*Where*

$$G_{k,p,\varepsilon} = \sum_{j=1}^{m} \Big( \sum_{\ell=1}^{p_j} \varepsilon_{j,\ell} \Big) C_j + \sum_{j=1}^{m} \Big( \sum_{\ell=p_j+1}^{k_j} \varepsilon_{j,\ell} \Big) S_j$$

$$\alpha_{k,p,\varepsilon} = \frac{(-1)^{m - \sum k_j}}{\prod_{j=1}^{m} k_j! j^{k_j}} (-1)^{\frac{|k|-|p|}{2}} \left( \prod_{j=1}^{m} \binom{k_j}{p_j} \right) \frac{1}{|k|!\, 2^{|k|}} \left( \prod_{j=1}^{m} \prod_{\ell=1}^{k_j} \varepsilon_{j,\ell} \right)$$

$$\beta_{k,p,\varepsilon} = \frac{(-1)^{m - \sum k_j}}{\prod_{j=1}^{m} k_j! j^{k_j}} (-1)^{\frac{|k|-|p|-1}{2}} \left( \prod_{j=1}^{m} \binom{k_j}{p_j} \right) \frac{1}{|k|!\, 2^{|k|}} \left( \prod_{j=1}^{m} \prod_{\ell=1}^{k_j} \varepsilon_{j,\ell} \right)$$

*and thus*

$$|\alpha_{k,p,\varepsilon}| = |\beta_{k,p,\varepsilon}| = \frac{1}{\left( \prod_{j=1}^{m} j^{k_j} \right) \left( \prod_{j=1}^{m} p_j!(k_j - p_j)! \right) |k|! 2^{|k|}} \leq \frac{1}{2}$$

*Furthermore, for $N_{tot}(m)$, the total amount of triples $(k, p, \varepsilon)$ (with $k \in \mathcal{K}_m$, $p \leq k$, $\varepsilon \in \{\pm 1\}^{|k|}$),*

$$N_{tot}(m) = \sum_{k \in \mathcal{K}_m} 2^{|k|} \prod_{j=1}^{m} (k_j + 1) \in [m 2^m, 13 m 2^m].$$

*Proof.* Let $z_j = e^{i\theta_j}$ for $j = 1, \ldots, m$. The $k$-th power sum is $Z_k = \sum_{j=1}^{m} z_j^k = C_k + iS_k$. Let $e_m = \prod_{j=1}^{m} z_j = e^{i \sum_{j=1}^{m} \theta_j}$ be the $m$-th elementary symmetric polynomial in $z_1, \ldots, z_m$. The target functions are $f_{\cos} = \Re(e_m)$ and $f_{\sin} = \Im(e_m)$.

Newton's sum identities provide a formula expressing $e_m$ as a polynomial in the power sums $Z_1, \ldots, Z_m$:

$$e_m = P(Z_1, \ldots, Z_m) = \sum_{k \in \mathcal{K}_m} c_k \prod_{j=1}^{m} Z_j^{k_j}$$

where the coefficients $c_k$ are given by $c_k = \frac{(-1)^{m - \sum k_j}}{\prod_{j=1}^{m} k_j! j^{k_j}}$.

Binomial expansion yields

$$\prod_{j=1}^{m} Z_j^{k_j} = \prod_{j=1}^{m} (C_j + iS_j)^{k_j} = \sum_{p \leq k} \prod_{j=1}^{m} \left( \binom{k_j}{p_j} C_j^{p_j} (iS_j)^{k_j - p_j} \right) = \sum_{p \leq k} i^{|k|-|p|} \left( \prod_{j=1}^{m} \binom{k_j}{p_j} \right) \left( \prod_{j=1}^{m} C_j^{p_j} S_j^{k_j - p_j} \right)$$

For each pair of $(p, k)$, where $p \le k$ and $k \in \mathcal{K}_m$, let

$$s := \sum_{j=1}^{m} k_j = |k|, \qquad x_{j,\ell} := \begin{cases} C_j, & 1 \le \ell \le p_j, \\ S_j, & p_j < \ell \le k_j. \end{cases}$$

List the $s$ variables as $(x_{1,1}, \dots, x_{1,k_1}, x_{2,1}, \dots, x_{m,k_m})$. Applying Lemma F.3 to $x_1 \cdots x_s = \prod_{j=1}^{m} C_j^{p_j} S_j^{k_j - p_j}$ gives

$$\prod_{j=1}^{m} C_j^{p_j} S_j^{k_j - p_j} = \frac{1}{|k|!\, 2^{|k|}} \sum_{(\varepsilon_{j,\ell}) \in \{\pm 1\}^{|k|}} \left( \prod_{j=1}^{m} \prod_{\ell=1}^{k_j} \varepsilon_{j,\ell} \right) \left( \sum_{j=1}^{m} \Big( \sum_{\ell=1}^{p_j} \varepsilon_{j,\ell} \Big) C_j + \sum_{j=1}^{m} \Big( \sum_{\ell=p_j+1}^{k_j} \varepsilon_{j,\ell} \Big) S_j \right)^{|k|}$$

Therefore,

$$\begin{aligned} e_m &= \sum_{k \in \mathcal{K}_m} c_k \prod_{j=1}^{m} Z_j^{k_j} \\ &= \sum_{k \in \mathcal{K}_m} c_k \sum_{p \le k} i^{|k|-|p|} \left( \prod_{j=1}^{m} \binom{k_j}{p_j} \right) \left( \prod_{j=1}^{m} C_j^{p_j} S_j^{k_j - p_j} \right) \\ &= \sum_{k \in \mathcal{K}_m} c_k \sum_{p \le k} i^{|k|-|p|} \left( \prod_{j=1}^{m} \binom{k_j}{p_j} \right) \frac{1}{|k|!\, 2^{|k|}} \sum_{(\varepsilon_{j,\ell}) \in \{\pm 1\}^{|k|}} \left( \prod_{j=1}^{m} \prod_{\ell=1}^{k_j} \varepsilon_{j,\ell} \right) \left( \sum_{j=1}^{m} \Big( \sum_{\ell=1}^{p_j} \varepsilon_{j,\ell} \Big) C_j + \sum_{j=1}^{m} \Big( \sum_{\ell=p_j+1}^{k_j} \varepsilon_{j,\ell} \Big) S_j \right)^{|k|} \end{aligned}$$

Separate Real and Imaginary part yields the polarized Newton expansion for $f_{\cos}$ and $f_{\sin}$.

For $j \ge 1$,

$$\sum_{k_j \ge 0} (k_j + 1)\, 2^{k_j}\, t^{j k_j} = \sum_{r \ge 0} (r+1)(2t^j)^r = \frac{1}{(1 - 2t^j)^2}.$$

Multiplying over $j$ gives the ordinary generating function

$$F(t) := \sum_{m \ge 0} N_{\text{tot}}(m) t^m = \prod_{j \ge 1} \frac{1}{(1 - 2t^j)^2} = \frac{1}{(1-2t)^2} \cdot H(t),$$

where

$$H(t) := \prod_{j \ge 2} (1 - 2t^j)^{-2} = \sum_{r \ge 0} h_r t^r, \qquad h_r \ge 0.$$

Since $(1-2t)^{-2} = \sum_{n \ge 0} (n+1) 2^n t^n$, the Cauchy product gives

$$N_{\text{tot}}(m) = \sum_{r=0}^{m} h_r\, (m - r + 1)\, 2^{\,m-r} \le (m+1) 2^m \sum_{r=0}^{\infty} h_r\, 2^{-r} = (m+1) 2^m\, H\big(\tfrac{1}{2}\big)\,.$$

Here $H(\frac{1}{2}) = \prod_{j \ge 2} (1 - 2^{1-j})^{-2} = \prod_{r \ge 1} (1 - 2^{-r})^{-2} < \infty$ is a finite absolute constant.

By Bernoulli's inequality, for all $x_i \in [0, 1]$,

$$(1 - x_1)(1 - x_2) \cdots (1 - x_s) \ge 1 - (x_1 + x_2 + \cdots + x_s).$$

Now observe that $\prod_{r \ge 1} (1 - 2^{-r}) = \frac{3}{8} \prod_{r \ge 3} (1 - 2^{-r})$, we have

$$\prod_{r \ge 1} (1 - 2^{-r}) = \frac{3}{8} \prod_{r \ge 3} (1 - 2^{-r}) \ge \frac{3}{8} (1 - \sum_{r=3}^{\infty} 2^{-r}) = \frac{9}{32}$$

Therefore,

$$H(\tfrac{1}{2}) \le \frac{1}{(9/32)^2} = \frac{1024}{81} \le 13$$

Thus

$$N_{\text{tot}}(m) \;\le\; H\big(\tfrac{1}{2}\big)\,(m+1)\,2^m \;\le\; 13m2^m.$$

On the other hand, taking just the term $k = (m, 0, 0, \dots) \in \mathcal{K}_m$ yields

$$N_{\text{tot}}(m) \;\ge\; 2^{\,|k|}\prod_j (k_j+1) = 2^m(m+1) \ge m2^m,$$

so $m2^m \le N_{\text{tot}}(m) \le 13m2^m$.

□

We are finally able to provide interpolations for ReLU networks, whose embedding weights echoes with "Pizza" algorithm in (Zhong et al., 2023).

**Theorem F.6** (ReLU construction)**.** *Fix integers $m \ge 1$ and $p \ge 2$. On*

$$\mathcal{X} = \{x \in \{0, 1, \dots, m\}^p : \; \|x\|_1 = m\},$$

*let the target be $y(x) \equiv (\sum_{i=1}^m s_i) \bmod p$ for $x = \sum_{i=1}^m e_{s_i}$. For any $\tau \in (0, \frac{1}{4}]$, there exists a two-layer ReLU network $s^\theta(x) = V\,\sigma(Wx) \in \mathbb{R}^p$ such that, for all $x \in \mathcal{X}$,*

$$h_\theta(x) = \arg\max_{q\in[p]} s^\theta_q(x) = y(x), \qquad s^\theta_{y(x)}(x) - \max_{q\ne y(x)} s^\theta_q(x) \;\ge\; (1-4\tau)\,p.$$

*Moreover, the width $d$ is bounded by*

$$d \;\le\; 13pm2^m\Big(m\sqrt{\tfrac{em}{\tau}}(1+2em)^{\frac{m-1}{2}} + 2\Big), \tag{2}$$

*and the weights satisfy the bounds*

$$\|W\|_\infty \le \frac{2}{m}, \qquad \|W\|_F \le \frac{2}{m}p\sqrt{13m2^m\Big(m\sqrt{\tfrac{em}{\tau}}(1+2em)^{\frac{m-1}{2}} + 2\Big)}, \qquad \|V\|_\infty \le \frac{(m+\frac{1}{2})m^{2m}}{m!\,2^m}. \tag{3}$$

*In addition, the second layer enjoys the spectral-norm bound*

$$\|V\|_2 \;\le\; \sqrt{p}\sqrt{13m2^m\Big(m\sqrt{\tfrac{em}{\tau}}(1+2em)^{\frac{m-1}{2}} + 2\Big)}\cdot\frac{\sqrt{2}(m+\frac{1}{2})m^{2m}}{m!\,2^m}. \tag{4}$$

*Proof.* For $t \in \mathbb{Z}$, define $c^{\langle t\rangle}, s^{\langle t\rangle} \in \mathbb{R}^p$ by $c_r^{\langle t\rangle} = \cos(2\pi tr/p)$ and $s_r^{\langle t\rangle} = \sin(2\pi tr/p)$ for $r = 0, \dots, p-1$. For $x = \sum_{i=1}^m e_{s_i}$ and any $j \ge 1$,

$$\sum_{i=1}^m \cos\big(j\tfrac{2\pi\nu}{p}s_i\big) = \langle c^{\langle \nu j\rangle}, x\rangle, \qquad \sum_{i=1}^m \sin\big(j\tfrac{2\pi\nu}{p}s_i\big) = \langle s^{\langle \nu j\rangle}, x\rangle, \quad \nu \in \{0, \dots, p-1\}.$$

Fix $\nu \in \{0, \dots, p-1\}$ and apply Lemma F.5 to $\theta_i^{(\nu)} = 2\pi\nu s_i/p$. The multi-indices are $\kappa = (\kappa_1, \dots, \kappa_m) \in \mathcal{K}_m$ and $\pi = (\pi_1, \dots, \pi_m) \le \kappa$, and write $r := |\kappa| = \sum_j \kappa_j$. Define

$$u^{(\nu)}_{\kappa,\pi,\varepsilon} = \sum_{j=1}^m\Big(\sum_{\ell=1}^{\pi_j}\varepsilon_{j,\ell}\Big)c^{\langle \nu j\rangle} + \sum_{j=1}^m\Big(\sum_{\ell=\pi_j+1}^{\kappa_j}\varepsilon_{j,\ell}\Big)s^{\langle \nu j\rangle}.$$

Then

$$C_\nu(x) = \sum_{\substack{\kappa,\pi,\varepsilon\\ |\pi|-|\kappa| \text{ even}}} \alpha_{\kappa,\pi,\varepsilon}\,\langle u^{(\nu)}_{\kappa,\pi,\varepsilon}, x\rangle^r, \quad S_\nu(x) = \sum_{\substack{\kappa,\pi,\varepsilon\\ |\pi|-|\kappa| \text{ odd}}} \beta_{\kappa,\pi,\varepsilon}\,\langle u^{(\nu)}_{\kappa,\pi,\varepsilon}, x\rangle^r,$$

with

$$|\alpha_{\kappa,\pi,\varepsilon}| = |\beta_{\kappa,\pi,\varepsilon}| = \frac{1}{\left(\prod_{j=1}^{m} j^{\kappa_j}\right)\left(\prod_{j=1}^{m} \pi_j!\,(\kappa_j-\pi_j)!\right) r!\,2^r} \;\le\; \frac{1}{r!\,2^r}.$$

As $|\langle c^{\langle \nu j\rangle}, x\rangle|, |\langle s^{\langle \nu j\rangle}, x\rangle| \le m$, we have $|\langle u^{(\nu)}_{\kappa,\pi,\varepsilon}, x\rangle| \le mr$ and thus $z^{(\nu)}_{\kappa,\pi,\varepsilon}(x) := \langle u^{(\nu)}_{\kappa,\pi,\varepsilon}, x\rangle/(mr) \in [-1,1]$.

By Lemma F.4, for each $r \in \{1,\dots,m\}$ and $\delta > 0$ there exists

$$\Phi_r(z) = \sum_{i=1}^{M_r} c_{r,i}\,\mathrm{ReLU}(a_{r,i}z - b_{r,i}), \qquad |a_{r,i}|, |b_{r,i}| \le 1,$$

such that $\sup_{|z|\le 1} |z^r - \Phi_r(z)| \le \delta$ and

$$M_r \;\le\; \frac{m}{\sqrt{2\delta}} + 2, \qquad |c_{r,i}| \;\le\; r + \tfrac{1}{2}. \tag{5}$$

In Equation 5, summing $|\alpha|$ (or $|\beta|$) over $\varepsilon \in \{\pm 1\}^r$ and summing over $\pi \le \kappa$ factorizes:

$$\sum_{\substack{\pi\le\kappa \\ \varepsilon\in\{\pm1\}^r}} |\alpha_{\kappa,\pi,\varepsilon}| = \frac{1}{r!} \cdot \frac{2^r}{\prod_{j=1}^{m} j^{\kappa_j}\,\kappa_j!}.$$

Summing over $\kappa \in \mathcal{K}_m$ with $|\kappa| = r$ and using the classical cycle-index identity

$$\sum_{\substack{\kappa\in\mathcal{K}_m \\ |\kappa|=r}} \frac{1}{\prod_{j=1}^{m} j^{\kappa_j}\,\kappa_j!} \;=\; \frac{1}{m!}\begin{bmatrix} m \\ r \end{bmatrix},$$

where $\left[\begin{smallmatrix} m \\ r \end{smallmatrix}\right]$ are the unsigned Stirling numbers of the first kind.

Now we have

$$\sum_{\substack{\kappa,\varepsilon \\ \pi\le\kappa}} |\alpha_{\kappa,\pi,\varepsilon}| \;=\; \sum_{r=1}^{m} \frac{2^r}{r!\,m!}\begin{bmatrix} m \\ r \end{bmatrix}, \qquad \sum_{\substack{\kappa,\varepsilon \\ \pi\le\kappa}} |\beta_{\kappa,\pi,\varepsilon}| \;=\; \sum_{r=1}^{m} \frac{2^r}{r!\,m!}\begin{bmatrix} m \\ r \end{bmatrix}.$$

As each power is approximated within $\delta$ and $|\langle u, x\rangle| \le mr$, the uniform error is bounded by

$$\sum_{r=1}^{m} \frac{2^r (mr)^r}{r!\,m!}\begin{bmatrix} m \\ r \end{bmatrix} \cdot \delta.$$

We choose

$$\delta \;:=\; \frac{\tau}{\Lambda_m}, \qquad \Lambda_m \;:=\; \sum_{r=1}^{m} \frac{2^r (mr)^r}{r!\,m!}\begin{bmatrix} m \\ r \end{bmatrix},$$

which ensures $\max\{|C_\nu - \widehat{C}_\nu|, |S_\nu - \widehat{S}_\nu|\} \le \tau$ uniformly on $\mathcal{X}$ for all $\nu$.

We now prove by induction on $m$ that

$$\left[\begin{smallmatrix} m \\ r \end{smallmatrix}\right] \;\le\; \binom{m-1}{r-1} m!, \qquad 1 \le r \le m. \tag{6}$$

For $m = 1$, both sides equal 1. Assume equation 6 holds for $m-1$. Using the recurrence $\left[\begin{smallmatrix} m \\ r \end{smallmatrix}\right] = \left[\begin{smallmatrix} m-1 \\ r-1 \end{smallmatrix}\right] + (m-1)\left[\begin{smallmatrix} m-1 \\ r \end{smallmatrix}\right]$,

$$\begin{aligned}
\begin{bmatrix} m \\ r \end{bmatrix} &\le \binom{m-2}{r-2}(m-1)! + (m-1)\binom{m-2}{r-1}(m-1)! \\
&= (m-1)!\Big[\binom{m-2}{r-2} + (m-1)\binom{m-2}{r-1}\Big] \\
&\le m\,(m-1)!\,\binom{m-1}{r-1} = \binom{m-1}{r-1} m!,
\end{aligned}$$

since $\binom{m-1}{r-1} = \binom{m-2}{r-2} + \binom{m-2}{r-1}$. This proves equation 6.

Using equation 6 and Stirling's lower bound $r! \geq (r/e)^r$, we have

$$\Lambda_m \;\leq\; \sum_{r=1}^{m} \frac{2^r (mr)^r}{r!} \binom{m-1}{r-1} \;\leq\; \sum_{r=1}^{m} (2em)^r \binom{m-1}{r-1} \;=\; (2em) \sum_{t=0}^{m-1} \binom{m-1}{t} (2em)^t \;=\; (2em)\,(1{+}2em)^{m-1}.$$

Hence

$$\frac{1}{\sqrt{2\delta}} = \sqrt{\frac{\Lambda_m}{2\tau}} \;\leq\; \sqrt{\frac{em}{\tau}} (1 + 2em)^{\frac{m-1}{2}}. \tag{7}$$

For $x \in \mathcal{X}$, $\langle \mathbf{1}, x \rangle = m$. So

$$\mathrm{ReLU}\big(a_{r,i} z^{(\nu)}_{\kappa,\pi,\varepsilon}(x) - b_{r,i}\big) = \sigma\left(\left\langle \frac{a_{r,i}}{mr}\, u^{(\nu)}_{\kappa,\pi,\varepsilon} - \frac{b_{r,i}}{m}\mathbf{1},\; x \right\rangle\right),$$

Each spline unit is a single ReLU of a linear form. Explicitly, $W \in \mathbb{R}^{d \times p}$ has rows $W_{j,:} = \frac{a_{r,i}}{mr}\, u^{(\nu)}_{\kappa,\pi,\varepsilon} - \frac{b_{r,i}}{m}\mathbf{1}$ for $j = (\nu, \kappa, \pi, \varepsilon, i)$ with $r = |\kappa|$ and $u^{(\nu)}_{\kappa,\pi,\varepsilon} = \sum_{t=1}^{m} \big(\sum_{\ell=1}^{\pi_t} \varepsilon_{t,\ell}\big) c^{\langle \nu t \rangle} + \sum_{t=1}^{m} \big(\sum_{\ell=\pi_t+1}^{\kappa_t} \varepsilon_{t,\ell}\big) s^{\langle \nu t \rangle}$. Since $\|u^{(\nu)}_{\kappa,\pi,\varepsilon}\|_\infty \leq r$ and $|a_{r,i}|, |b_{r,i}| \leq 1$, each coordinate obeys $|W_{j,t}| \leq \frac{|a_{r,i}|}{mr}\, r + \frac{|b_{r,i}|}{m} \leq \frac{1}{m} + \frac{1}{m} = \frac{2}{m}$, hence $\|W\|_\infty \leq \frac{2}{m}$.

For class $q \in \{0, \ldots, p-1\}$ and hidden index $(\nu, \kappa, \pi, \varepsilon, i)$ set

$$V_{q,(\nu,\kappa,\pi,\varepsilon,i)} = \Big[\cos\big(\tfrac{2\pi\nu}{p} q\big)\, \alpha_{\kappa,\pi,\varepsilon} + \sin\big(\tfrac{2\pi\nu}{p} q\big)\, \beta_{\kappa,\pi,\varepsilon}\Big]\, (mr)^r\, c_{r,i},$$

so that $s^\theta_q(x) = \sum_{\nu=0}^{p-1} \big[\cos(\frac{2\pi\nu}{p} q)\, \widehat{C}_\nu(x) + \sin(\frac{2\pi\nu}{p} q)\, \widehat{S}_\nu(x)\big]$. Let $q^\star \equiv (\sum_i s_i) \bmod p$. Discrete Fourier orthogonality gives $s^\star_q(x) = \sum_{\nu=0}^{p-1} \cos(\frac{2\pi\nu}{p}(\sum_i s_i - q)) = \mathbf{1}_{\{q=q^\star\}} p$. Since each mode is within $\tau$, we have $\max_q |s^\theta_q(x) - s^\star_q(x)| \leq 2p\tau$ and thus the claimed margin $(1 - 4\tau)p$.

For each fixed $\nu \in \{0, \ldots, p-1\}$, by Lemma F.5, there are $N_{\mathrm{tot}}(m)$ triples $(\kappa, \pi, \varepsilon)$, each contributes at most $M_r$ units, with $M_r$ bounded in equation 5. Hence for each $\nu$, the width is at most $N_{\mathrm{tot}}(m)\big(\frac{m}{\sqrt{2\delta}} + 2\big)$.

Summing over $\nu = 0, 1, \ldots, p-1$ and using equation 7,

$$d \;\leq\; p\, N_{\mathrm{tot}}(m) \Big(\frac{m}{\sqrt{2\delta}} + 2\Big) \;\leq\; p\, N_{\mathrm{tot}}(m) \Big(m \sqrt{\tfrac{em}{\tau}}\, (1 + 2em)^{\frac{m-1}{2}} + 2\Big),$$

and the bound $N_{\mathrm{tot}}(m) \leq 13 m 2^m$ gives equation 2.

Thus, $\|W\|_F \leq \|W\|_\infty \sqrt{dp} \leq \frac{2}{m} p \sqrt{13 m 2^m \Big(m \sqrt{\frac{em}{\tau}} (1 + 2em)^{\frac{m-1}{2}} + 2\Big)}$.

Finally, using $|\alpha|, |\beta| \leq 1/(r!\, 2^r)$ and equation 5,

$$|V_{q,(\nu,\kappa,\pi,\varepsilon,i)}| \;\leq\; |c_{r,i}|\, (mr)^r \cdot \frac{1}{r!\, 2^r} \;\leq\; \frac{(r + \frac{1}{2})(mr)^r}{r!\, 2^r},$$

so taking the maximum over all hidden indices yields equation 3.

For the spectral norm, denote the matrix

$$T = \begin{pmatrix} c^{\langle 0 \rangle} & c^{\langle 1 \rangle} & s^{\langle 1 \rangle} & \cdots & c^{\langle p-1 \rangle} & s^{\langle p-1 \rangle} \end{pmatrix}$$

So

$$TT^\top = c^{(0)} c^{(0)\top} + \sum_{\nu=1}^{p-1} \Big(c^{(\nu)} c^{(\nu)\top} + s^{(\nu)} s^{(\nu)\top}\Big) = pI_p, \text{ and thus } \|T\|_2 = \sqrt{p}.$$

Index the hidden units by $j = (\nu, \kappa, \pi, \varepsilon, i)$, with $r = |\kappa|$. For that unit, the corresponding column of $V$ was

$$V_{:,j} \;=\; \big[\alpha_{\kappa,\pi,\varepsilon}(mr)^r c_{r,i}\big]\, c^{\langle\nu\rangle} \;+\; \big[\beta_{\kappa,\pi,\varepsilon}(mr)^r c_{r,i}\big]\, s^{\langle\nu\rangle}.$$

Hence $V_{:,j}$ is a linear combination of the two columns of $S_\nu$.

Define $B \in \mathbb{R}^{(2p-1)\times d}$, for each column $j = (\nu, \kappa, \pi, \varepsilon, i)$,

$$B_{k,j} \;=\; \begin{cases} \alpha_{\kappa,\pi,\varepsilon}(mr)^r c_{r,i}, & k = 0 \text{ and } \nu = 0, \\ \alpha_{\kappa,\pi,\varepsilon}(mr)^r c_{r,i}, & k = 2\nu \text{ with } \nu \in \{1,\dots,p-1\}, \\ \beta_{\kappa,\pi,\varepsilon}(mr)^r c_{r,i}, & k = 2\nu - 1 \text{ with } \nu \in \{1,\dots,p-1\}, \\ 0, & \text{otherwise.} \end{cases}$$

One has $V = TB$, and each column $b_j$ of $B$ has support in at most two rows (one when $\nu = 0$). Thus,

$$\|b_j\|_2 = \sqrt{\alpha^2_{\kappa,\pi,\varepsilon} + \beta^2_{\kappa,\pi,\varepsilon}} \;\cdot\; |c_{r,i}|\,(mr)^r \;\le\; \frac{\sqrt{2}\,(r+\frac{1}{2})(mr)^r}{r!\,2^r} \;\le\; \frac{\sqrt{2}\,(m+\frac{1}{2})\,m^{2m}}{m!\,2^m}.$$

Let $n_\nu$ be the number of hidden units at frequency $\nu$. From the construction,

$$n_\nu \;\le\; N_{\mathrm{tot}}(m)\Big(\frac{m}{\sqrt{2\delta}} + 2\Big) \;\le\; 13\, m\, 2^m \left(m\sqrt{\tfrac{em}{\tau}}\,(1+2em)^{\frac{m-1}{2}} + 2\right).$$

Since $BB^\top$ is block diagonal across frequencies, $\|B\|_2 = \max_\nu \|B_\nu\|_2 \le \max_\nu \sqrt{n_\nu} \cdot \frac{\sqrt{2}(m+\frac{1}{2})m^{2m}}{m!\,2^m}$. Therefore

$$\|V\|_2 \;\le\; \|T\|_2\,\|B\|_2 \;\le\; \sqrt{p}\,\sqrt{\max_\nu n_\nu} \;\cdot\; \frac{\sqrt{2}\,(m+\frac{1}{2})\,m^{2m}}{m!\,2^m},$$

which gives equation 4.

□

**Corollary F.7** (Explicit two-layer ReLU construction for $m = 2$)**.** *Fix $p \ge 2$. Define the input set*

$$\mathcal{X} = \big\{x \in \{0,1,2\}^p :\; \|x\|_1 = 2\big\}.$$

*There exists a two-layer ReLU network $s^\theta(x) = V\,\sigma(Wx) \in \mathbb{R}^p$ of width $d = 36p$ such that, for all $x \in \mathcal{X}$,*

$$h_\theta(x) = \arg\max_{q\in[p]} s^\theta_q(x) = \Big(\sum_{i=1}^{2} s_i\Big) \bmod p, \qquad s^\theta_{y(x)}(x) - \max_{q\neq y(x)} s^\theta_q(x) \;\ge\; \frac{25}{49}p + \frac{20}{49}.$$

*Moreover, the weights satisfy*

$$\|W\|_\infty \le 1, \qquad \|V\|_\infty \le \tfrac{34}{7}, \qquad \|V\|_2 \;\le\; 11\sqrt{p}.$$

*Proof.* For $\nu \in \{0,\dots,p-1\}$ let $c^{\langle\nu\rangle}, s^{\langle\nu\rangle} \in \mathbb{R}^p$ be defined by $c^{\langle\nu\rangle}_r = \cos(2\pi\nu r/p)$ and $s^{\langle\nu\rangle}_r = \sin(2\pi\nu r/p)$. For inputs $x \in \mathcal{X}$, write

$$C_k = \langle c^{\langle k\nu\rangle}, x\rangle, \qquad S_k = \langle s^{\langle k\nu\rangle}, x\rangle \qquad (k = 1, 2).$$

From Lemma F.5, for any $\theta_1, \theta_2 \in \mathbb{R}$,

$$\cos(\theta_1+\theta_2) = \frac{1}{2}(C_1^2 - S_1^2 - C_2) = 2(\frac{1}{2}C_1)^2 - 2(\frac{1}{2}S_1)^2 - \frac{1}{2}C_2$$

$$\sin(\theta_1+\theta_2) = C_1S_1 - \frac{1}{2}S_2 = 4\left(\left(\frac{C_1+S_1}{4}\right)^2 - \left(\frac{C_1-S_1}{4}\right)^2\right) - \frac{1}{2}S_2$$

For $\|x\|_1 = 2$, we have $\frac{C_1}{2}, \frac{S_1}{2}, \frac{C_1\pm S_1}{4} \in [-1, 1]$.

Let $\Phi_2$ be the piecewise-linear interpolant of $z^2$ on the uniform grid $z_k = -1 + \frac{2k}{7}$, $k = 0, \dots, 7$.

Using Lemma F.4 with $s = 2$, $N = 7$, $\|\Phi_2 - z^2\|_{L_\infty([-1,1])} \le 1/49$, and $\Phi_2(z) = \sum_{i=1}^{8} c_i \,\mathrm{ReLU}(a_i z - b_i)$, where

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $(a_i, b_i)$ | $(1, -1)$ | $(-1, -1)$ | $(1, -\frac{5}{7})$ | $(1, -\frac{3}{7})$ | $(1, -\frac{1}{7})$ | $(1, \frac{1}{7})$ | $(1, \frac{3}{7})$ | $(1, \frac{5}{7})$ |
| $c_i$ | $-\frac{17}{14}$ | $\frac{1}{2}$ | $\frac{4}{7}$ | $\frac{4}{7}$ | $\frac{4}{7}$ | $\frac{4}{7}$ | $\frac{4}{7}$ | $\frac{4}{7}$ |

We now construct a two-layer $\mathrm{ReLU}$ MLP with total width $d = 36p$.

First layer. For $r \in \{0, \dots, p-1\}$ and $i = 1, \dots, 8$ define

$$
\begin{aligned}
w_r^{(\nu,1,i)} &= \tfrac{a_i}{2}\, c_r^{\langle\nu\rangle} - \tfrac{b_i}{2}, & w_r^{(\nu,2,i)} &= \tfrac{a_i}{2}\, s_r^{\langle\nu\rangle} - \tfrac{b_i}{2}, \\
w_r^{(\nu,3,i)} &= \tfrac{a_i}{4}\big(c_r^{\langle\nu\rangle} + s_r^{\langle\nu\rangle}\big) - \tfrac{b_i}{2}, & w_r^{(\nu,4,i)} &= \tfrac{a_i}{4}\big(c_r^{\langle\nu\rangle} - s_r^{\langle\nu\rangle}\big) - \tfrac{b_i}{2}, \\
w_r^{(\nu,C_2^{\pm})} &= \pm\tfrac{1}{2}\, c_r^{\langle 2\nu\rangle}, & w_r^{(\nu,S_2^{\pm})} &= \pm\tfrac{1}{2}\, s_r^{\langle 2\nu\rangle}.
\end{aligned}
$$

Then $\sigma(\langle w^{(\nu,1,i)}, x\rangle) = \mathrm{ReLU}(a_i C_1/2 - b_i)$, etc. Since $|a_i| \le 1$, $|b_i| \le 1$, and $|c_r^{\langle\nu\rangle}|, |s_r^{\langle\nu\rangle}| \le 1$, we have $\|W\|_\infty \le 1$.

Second layer. For $q \in \{0, \dots, p-1\}$, $\nu \in \{0, \dots, p-1\}$ set

$$
\begin{aligned}
V_{q,(\nu,1,i)} &= +2c_i\, \cos(2\pi\nu q/p), & V_{q,(\nu,2,i)} &= -2c_i\, \cos(2\pi\nu q/p), \\
V_{q,(\nu,3,i)} &= +4c_i\, \sin(2\pi\nu q/p), & V_{q,(\nu,4,i)} &= -4c_i\, \sin(2\pi\nu q/p),
\end{aligned}
\qquad i = 1, \dots, 8,
$$

and

$$
V_{q,(\nu,C_2^{\pm})} = \mp\, \cos(2\pi\nu q/p), \qquad V_{q,(\nu,S_2^{\pm})} = \mp\, \sin(2\pi\nu q/p).
$$

We have $\|V\|_\infty \le \max\{|4c_i|, 1\} = \frac{34}{7}$.

Let $T = [\, c^{\langle 0\rangle}\ c^{\langle 1\rangle}\ s^{\langle 1\rangle}\ \cdots\ c^{\langle p-1\rangle}\ s^{\langle p-1\rangle}\,]$ and write $V = TB$. Then

$$
TT^\top = c^{\langle 0\rangle} c^{\langle 0\rangle\top} + \sum_{\nu=1}^{p-1} \Big( c^{\langle\nu\rangle} c^{\langle\nu\rangle\top} + s^{\langle\nu\rangle} s^{\langle\nu\rangle\top} \Big) = p\, I_p,
$$

so $\|T\|_2 = \sqrt{p}$.

Each hidden unit loads a single row in $B$, hence $BB^\top$ is diagonal. The largest row norm equals $\sqrt{2\sum_{i=1}^{8}(4c_i)^2 + 2} = \frac{\sqrt{5874}}{7}$, so

$$
\|V\|_2 \le \|T\|_2\, \|B\|_2 \le 11\sqrt{p}.
$$

Finally, define

$$
\widehat{C}_\nu(x) = 2\,\Phi_2\big(\tfrac{C_1}{2}\big) - 2\,\Phi_2\big(\tfrac{S_1}{2}\big) - \tfrac{1}{2}C_2, \quad \widehat{S}_\nu(x) = 4\,\Phi_2\big(\tfrac{C_1+S_1}{4}\big) - 4\,\Phi_2\big(\tfrac{C_1-S_1}{4}\big) - \tfrac{1}{2}S_2,
$$

and logits $s_q^\theta(x) = \sum_{\nu=0}^{p-1} \big[\cos(2\pi\nu q/p)\, \widehat{C}_\nu(x) + \sin(2\pi\nu q/p)\, \widehat{S}_\nu(x)\big]$. Since $\|\Phi_2 - z^2\|_\infty \le 1/49$ and $\nu = 0$ contributes a class-independent offset, for $\nu \ge 1$,

$$
|\widehat{C}_\nu - C_\nu| \le 4/49 \quad \text{and} \quad |\widehat{S}_\nu - S_\nu| \le 8/49.
$$

Therefore,

$$
\max_q |s_q^\theta(x) - s_q^\star(x)| \le \tfrac{12}{49}(p-1) + \frac{2}{49},
$$

where $s_q^\star(x) = \sum_{\nu=0}^{p-1} \cos\big(2\pi\nu(\sum_i s_i - q)/p\big)$ satisfies $s_{y(x)}^\star(x) = p$ and $s_q^\star(x) = 0$ if $q \ne y(x)$. The margin follows:

$$
s_{y(x)}^\theta(x) - \max_{q \ne y(x)} s_q^\theta(x) \;\ge\; p - 2\left(\tfrac{12}{49}(p-1) + \frac{2}{49}\right) = \frac{25}{49}p + \frac{20}{49}.
$$

□

# G NECESSARY WIDTH FOR MODULAR ADDITION WITH RELU

By a one-dimensional counting-path argument, we show that any ReLU MLP that exactly implements modular addition requires width $\Omega(m/p)$.

*Proof of Theorem 4.2.* Consider the one-dimensional path of count vectors

$$x^{(s)} \;=\; (m-s)e_1 + se_2, \qquad s \in \{0,1,\dots,m\}.$$

Hence the correct label along the path is

$$\ell(s) \;\equiv\; \big((m-s)\cdot 1 + s\cdot 2\big) \pmod p \;=\; (m+s) \pmod p,$$

Write $y \in \mathbb{R}^d$ for the first column of $W$ and $z \in \mathbb{R}^d$ for the difference between the second and first columns, i.e., $y_k = W_{k,1}$ and $z_k = W_{k,2} - W_{k,1}$ for $k \in \{1,\dots,d\}$. Along the path,

$$a_k(s) \;:=\; [Wx^{(s)}]_k \;=\; (m-s)W_{k,1} + sW_{k,2} \;=\; m\,y_k + s\,z_k,$$

so $h_k(s) := \sigma(a_k(s))$ is piecewise-affine in $s$ with a single potential breakpoint at $s_k := -m\,y_k/z_k$ when $z_k \neq 0$. Consequently, for each class $r \in [p]$ the score

$$s_r^\theta\big(x^{(s)}\big) \;=\; \sum_{k=1}^{d} v_{r,k}\,h_k(s)$$

is a univariate piecewise-affine function whose breakpoints lie in the shared set

$$\mathcal{B} \;=\; \{\, s_k :\; z_k \neq 0,\; s_k = -m\,y_k/z_k \,\}, \qquad |\mathcal{B}| \le d.$$

Define the adjacent-class margin

$$g_r(s) \;:=\; s_r^\theta\big(x^{(s)}\big) \;-\; s_{r\oplus 1}^\theta\big(x^{(s)}\big), \quad r \in [p],$$

where $r \oplus 1 = r+1$ if $r < p$ and $r \oplus 1 = 1$ if $r = p$. Each $g_r$ is continuous, piecewise-affine with breakpoints in $\mathcal{B}$, hence has at most $|\mathcal{B}| \le d$ breakpoints and can change sign at most $d+1$ times on $\mathbb{R}$.

Exact realization of modular addition implies that, for every $s \in \{0,\dots,m-1\}$,

$$g_{\ell(s)}(s) \;>\; 0 \quad \text{and} \quad g_{\ell(s)}(s+1) \;<\; 0,$$

because the winner at $x^{(s)}$ is $\ell(s)$ and at $x^{(s+1)}$ is its successor $\ell(s) \oplus 1 = \ell(s+1)$.

By continuity, $g_{\ell(s)}$ has a zero crossing in $(s, s+1)$. The $m$ disjoint intervals $(0,1),\dots,(m-1,m)$ therefore contain at least $m$ zero crossings in total, each attributed to one of the $p$ functions $\{g_r\}_{r\in[p]}$.

Since for each $r \in [p]$, $g_r$ changes sign at most $d+1$ times, we have

$$m \le p(d+1)$$

which rearranges to $d \ge m/p - 1$. □

# H MARGIN BOUNDS VIA $\ell_\infty$ VECTOR CONTRACTION

## H.1 MARGIN SURROGATES AND EMPIRICAL $\gamma$-MARGIN ERROR.

Given scores $s \in \mathbb{R}^p$ for an example with label $y \in [p]$, the *sample margin error*

$$\phi_y(s) = \max_{k\neq y}(s_k - s_y)$$

The $\gamma$-*ramp loss*

$$\psi_\gamma(u) = \min\{1,\ \max\{0,\ 1 + u/\gamma\}\} \in [0,1].$$

The map $u \mapsto \psi_\gamma(u)$ is $1/\gamma$-Lipschitz on $\mathbb{R}$, and $\phi_y$ is 2-Lipschitz w.r.t. $\|\cdot\|_\infty$ (changing any coordinate of $s$ by at most $\varepsilon$ changes $\phi_y$ by at most $2\varepsilon$), hence

$$g_y := \psi_\gamma \circ \phi_y \quad \text{is} \quad \tfrac{2}{\gamma}\text{-Lipschitz w.r.t. } \|\cdot\|_\infty, \qquad g_y \in [0,1].$$

**Definition H.1** (Empirical Margin Error)**.** For a score function $s^\theta$ and sample $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, the *empirical $\gamma$-margin error* is

$$\widehat{\mathcal{R}}_\gamma(s^\theta; S) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big[s^\theta\big(x^{(i)}\big)_{y^{(i)}} \le \gamma + \max_{j \neq y^{(i)}} s^\theta\big(x^{(i)}\big)_j\Big].$$

For an interpolating solution, it suffices to take $\gamma = \gamma_\theta(S)$, the minimum sample margin, in which case $\widehat{\mathcal{R}}_\gamma(s^\theta; S) = 0$.

**Definition H.2** (Empirical Rademacher complexity)**.** Let $S = \big\{z_i = (x^{(i)}, y^{(i)})\big\}_{i=1}^n$ be fixed, and let $\mathcal{G} \subset [0,1]^{\mathcal{Z}}$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ be i.i.d. Rademacher variables ($\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = 1/2$). The empirical Rademacher complexity of $\mathcal{G}$ on $S$ is

$$\mathfrak{R}_S(\mathcal{G}) \;=\; \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \epsilon_i\, g(z_i)\right].$$

**Theorem H.3** (Thm. 3.3 of (Mohri et al., 2018))**.** *Let $\mathcal{D}$ be the true distribution, $\mathcal{G} \subset [0,1]^{\mathcal{Z}}$ and let $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$. With probability at least $1-\delta$ over $S$, the following holds simultaneously for all $g \in \mathcal{G}$:*

$$\mathbb{E}_{z\sim\mathcal{D}}[g(z)] \;\le\; \frac{1}{n}\sum_{i=1}^{n} g(z_i) \;+\; 2\,\mathfrak{R}_S(\mathcal{G}) \;+\; 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

*where $\mathfrak{R}_S(\mathcal{G})$ is the empirical Rademacher complexity of $\mathcal{G}$ on $S$.*

Apply Theorem H.3 with $\mathcal{G} = \mathcal{F}_\gamma := \{(x,y) \mapsto \psi_\gamma \circ \phi_y(f(x)) : f \in \mathcal{F}\}$, and note $\mathbf{1}\{\arg\max_i f_i(x) \neq y\} \le \psi_\gamma \circ \phi_y\big(f(x)\big)$. That yields the following corollary:

**Corollary H.4** (Rademacher complexity and Multiclassification)**.**

$$\mathbb{P}_{(x,y)\in\mathcal{D}}\big[f(x) \neq y\big] \le \widehat{\mathcal{R}}_\gamma(f) + 2\,\mathfrak{R}_S(\mathcal{F}_\gamma) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{8}$$

Let $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ be the training sample generated from the true distribution and write

$$Q_2(S) \;=\; \Big(\tfrac{1}{n}\sum_{i=1}^{n} \|x^{(i)}\|_2^2\Big)^{1/2}.$$

### H.2 Margin Bounds for sine MLP

**Definition H.5** (Covering Number for sets)**.** Let $(X, d)$ be a metric space, $F \subseteq X$ a non-empty subset, and $r > 0$. The covering number of $F$, denoted $\mathcal{N}(F, d, r)$, is

$$\mathcal{N}(F,d,r) = \min\left\{k \in \mathbb{N} \mid \exists\{x_1,\ldots,x_k\} \subseteq X \text{ such that } F \subseteq \bigcup_{i=1}^{k} B_d(x_i, r)\right\},$$

where $B_d(x,r) = \{y \in X \mid d(x,y) \le r\}$ is the closed ball of radius $r$ centered at $x$.

**Definition H.6** (Empirical $L_2$ covering number of a function class)**.** Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of real-valued functions and let $x_{1:n} = (x_1, \ldots, x_n) \in \mathcal{X}^n$. Define the empirical $L_2$ metric

$$d_{2,x_{1:n}}(f,g) \;:=\; \Big(\frac{1}{n}\sum_{i=1}^{n}\big(f(x_i) - g(x_i)\big)^2\Big)^{1/2}.$$

For $\varepsilon > 0$, the empirical $L_2$ covering number of $\mathcal{F}$ at scale $\varepsilon$ with respect to the sample $x_{1:n}$ is

$$\mathcal{N}_2(\varepsilon, \mathcal{F}, x_{1:n}) \;:=\; \min\Big\{k \in \mathbb{N} \,:\, \exists\, f_1,\ldots,f_k \text{ such that } \mathcal{F} \subseteq \bigcup_{j=1}^{k} B_{d_{2,x_{1:n}}}(f_j, \varepsilon)\Big\},$$

where $B_{d_{2,x_{1:n}}}(f,\varepsilon) = \{g :\, d_{2,x_{1:n}}(f,g) \le \varepsilon\}$.

**Lemma H.7** (Covering the box $[-\pi,\pi)^p$ by Euclidean balls)**.** *Fix $p \in \mathbb{N}$ and $r > 0$. Then*

$$\mathcal{N}\big([-\pi,\pi)^p, \|\cdot\|_2, r\big) \;\le\; \Big\lceil \frac{\pi\sqrt{p}}{r} \Big\rceil^p.$$

*Proof.* Covering numbers are translation invariant: for any $a \in \mathbb{R}^p$, $\mathcal{N}(F, \|\cdot\|_2, r) = \mathcal{N}(F + a, \|\cdot\|_2, r)$. Hence it suffices to cover $[0, 2\pi)^p$.

Set the grid step $h := 2r/\sqrt{p}$ and the number of points per dimension $m := \lceil 2\pi/h \rceil = \lceil \pi\sqrt{p}/r \rceil$. Along each coordinate, place grid points with a half-step offset from the origin:

$$G_1 := \big\{(j + \tfrac{1}{2})h \,:\, j = 0, 1, \ldots, m-1\big\},$$

so $|G_1| = m$. Let the full grid be the Cartesian product $G := G_1^p$; then $|G| = m^p$.

Given any point $x \in [0, 2\pi)^p$, choose $g \in G$ by rounding each coordinate of $x$ to the nearest point in $G_1$ (breaking ties arbitrarily). By construction, the distance from any coordinate $x_i$ to its corresponding grid point $g_i$ is at most half the grid step, so $\|x - g\|_\infty \le h/2 = r/\sqrt{p}$. We have

$$\|x - g\|_2 \le \sqrt{p}\,\|x - g\|_\infty \le \sqrt{p} \cdot \frac{r}{\sqrt{p}} = r.$$

Therefore, the set of closed $\ell_2$-balls $\{B_2(g, r) : g \in G\}$ covers the box $[0, 2\pi)^p$, and

$$\mathcal{N}\big([0, 2\pi)^p, \|\cdot\|_2, r\big) \le |G| = m^p = \left( \Big\lceil \frac{\pi\sqrt{p}}{r} \Big\rceil \right)^p.$$

□

**Lemma H.8** (Standard Dudley entropy integral )**.** *Assume that all $\mathcal{F}_{x_{1:n}} \subset \mathbb{R}^n$. Let $\mathfrak{R}_n(\mathcal{F})$ be the empirical Rademacher number of $\mathcal{F}$ on $x_{1:n}$. We have:*

$$\mathfrak{R}_n(\mathcal{F}) \le \inf_{\alpha \ge 0} \left( 4\alpha + 12 \int_\alpha^\infty \sqrt{\frac{\log N_2(\epsilon, \mathcal{F}, x_{1:n})}{n}} d\epsilon \right)$$

**Theorem H.9** (Width-independent multiclass margin bound for the sine MLP)**.** *Consider the two-layer sine network with parameters $\theta = (W, V) \in \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d}$, where the output matrix satisfies $\|V\|_\infty \le S_1$. Then for any $\gamma > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of the training sample $S$, the following holds simultaneously for all such $\theta$:*

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}}\big[h_\theta(X) \ne Y\big] \;\le\; \widehat{\mathcal{R}}_\gamma\big(s^\theta\big) \;+\; \widetilde{O}\Big(\frac{S_1}{\gamma} \cdot \frac{p}{\sqrt{n}}\Big) \;+\; \widetilde{O}\Big(\frac{1}{\sqrt{n}}\Big).$$

*Proof.* Because inputs are bag of words ($x \in \{0, 1, \ldots, m\}^p$ with $\|x\|_1 = m$), shifting any element of $W$ by $2\pi k$ ($k \in \mathbb{Z}$) does not change $s^\theta(x) = V \sin(Wx)$. Hence without loss of generality, each element of $W$ may be reduced to modulo $2\pi$ to $[-\pi, \pi)$ with no effect on the model output. This periodic reduction is the core argument in the sine analysis.

Notice that $g_y := \psi_\gamma \circ \phi_y$ is $\frac{2}{\gamma}$-Lipschitz w.r.t. $\|\cdot\|_\infty$ and $g_y \in [0, 1]$. Applying Theorem H.3 with $\mathcal{G} = \mathcal{F}_\gamma := (x, y) \mapsto g_y(s^\theta(x)) : \theta$ and recalling $\mathbf{1}_{\arg\max f \ne y} \le g_y$, we obtain

$$\mathbb{P}\big[h_\theta(X) \ne Y\big] \;\le\; \widehat{\mathcal{R}}_\gamma(s^\theta) + 2\mathfrak{R}_S(\mathcal{F}_\gamma) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{9}$$

**$\ell_\infty$ vector contraction**. Let $\mathcal{S} := \{s^\theta : \theta = (W, V),\ \|V\|_\infty \le S_1,\ W \in [-\pi, \pi)^{d\times p}\}$ and denote the coordinate classes

$$\mathcal{S}|_j \;:=\; \big\{\, x \mapsto v_j^\top \sin(Wx) :\ \|v_j\|_1 \le S_1,\ W \in [-\pi, \pi)^{d\times p} \big\}.$$

For fixed $S = (x^{(1)}, \ldots, x^{(n)})$ and the Lipschitz maps $\varphi_i \equiv g_{y^{(i)}}$ (each $\frac{2}{\gamma}$-Lipschitz w.r.t. $\|\cdot\|_\infty$), the $\ell_\infty$ vector contraction inequality (Thm. 1 of (Foster & Rakhlin, 2019)) gives

$$\mathfrak{R}_S(\mathcal{F}_\gamma) \;\le\; C\frac{2}{\gamma}\sqrt{p}\max_{j\in[p]} \mathfrak{R}_S(\mathcal{S}|_j) \log^{\frac{3}{2}+\delta_0}\Big(\frac{\beta}{\max_j \mathfrak{R}_S(\mathcal{S}|_j)}\Big), \tag{10}$$

for any fixed $\delta_0 > 0$ and some $C = C(\delta_0)$. Since $\sin(Wx) \in [-1,1]^d$ and $\|v_j\|_1 \le S_1$, we have

$$\|s^\theta(x)\|_\infty \le S_1, \text{ and thus } \beta \;\le\; 1 + S_1. \tag{11}$$

**Coordinate reduction via $\ell_1$-$\ell_\infty$ duality**. For any fixed $S = (x^{(1)}, \dots, x^{(n)})$ and $j \in [p]$,

$$\begin{aligned}
n\mathfrak{R}_S(\mathcal{S}|_j) &= \mathbb{E}_\epsilon \sup_{\substack{\|v_j\|_1 \le S_1 \\ W \in \mathbb{R}^{d\times p}}} \sum_{i=1}^n \epsilon_i v_j^\top \sin(Wx^{(i)}) \\
&= \mathbb{E}_\epsilon \sup_{\substack{\|v_j\|_1 \le S_1 \\ W \in \mathbb{R}^{d\times p}}} v_j^\top \Big( \sum_{i=1}^n \epsilon_i \sin(Wx^{(i)}) \Big) \\
&\le S_1 \mathbb{E}_\epsilon \sup_{W \in \mathbb{R}^{d\times p}} \Big\| \sum_{i=1}^n \epsilon_i \sin(Wx^{(i)}) \Big\|_\infty \\
&= S_1 \mathbb{E}_\epsilon \sup_{w \in \mathbb{R}^p} \Big| \sum_{i=1}^n \epsilon_i \sin(w^\top x^{(i)}) \Big| \\
&= S_1 \mathbb{E}_\epsilon \sup_{w \in \mathbb{R}^p} \sum_{i=1}^n \epsilon_i \sin(w^\top x^{(i)}) \\
&= S_1 \mathbb{E}_\epsilon \sup_{w \in [-\pi,\pi)^p} \sum_{i=1}^n \epsilon_i \sin(w^\top x^{(i)}) \\
&= S_1 n \mathfrak{R}_S(\mathcal{F}_{\sin}).
\end{aligned} \tag{12}$$

Here we used $\sup_{\|a\|_1 \le S_1} \langle a, b\rangle = S_1 \|b\|_\infty$, and denoted the single-sine family

$$\mathcal{F}_{\sin} := \big\{ x \mapsto \sin(w^\top x) : w \in [-\pi, \pi)^p \big\}.$$

**Rademacher complexity of the single-sine family.**

Endow $\mathcal{F}_{\sin}$ with the empirical $L_2$ metric

$$d(w, w')^2 \;:=\; \frac{1}{n} \sum_{i=1}^n \big( \sin(w^\top x^{(i)}) - \sin(w'^\top x^{(i)}) \big)^2.$$

Notice that $d(w, w') \le 2$ for all $w, w'$, so for any $\varepsilon \in (2, \infty)$, $\mathcal{N}_2(\varepsilon, \mathcal{F}_{\sin}, x_{1:n}) = 1$.

For any $i$,

$$\big| \sin(w^\top x^{(i)}) - \sin(w'^\top x^{(i)}) \big| \;\le\; |(w - w')^\top x^{(i)}| \;\le\; \|w - w'\|_2 \, \|x^{(i)}\|_2 \le m \, \|w - w'\|_2$$

so if $\|w - w'\|_2 \le \varepsilon/m$ then $d(w, w') \le \varepsilon$. Consequently, for any $\varepsilon \in (0, 2]$,

$$\mathcal{N}_2(\varepsilon, \mathcal{F}_{\sin}, x_{1:n}) \;\le\; \mathcal{N}\big( [-\pi, \pi)^p, \|\cdot\|_2, \varepsilon/m \big) \;\le\; \left\lceil \frac{\pi m \sqrt{p}}{\varepsilon} \right\rceil^p, \tag{13}$$

where we used Lemma H.7.

Applying the standard Dudley entropy integral with any $\alpha \in (0, 1]$ yields

$$\mathfrak{R}_S(\mathcal{F}_{\sin}) \le 4\alpha \;+\; 12 \int_\alpha^2 \sqrt{\frac{\log \mathcal{N}_2(\varepsilon, \mathcal{F}_{\sin}, x_{1:n})}{n}} \; d\varepsilon \tag{14}$$

Let $C := \pi m \sqrt{p} > 2$. Then $\left\lceil \frac{\pi m \sqrt{p}}{\varepsilon} \right\rceil \le \frac{\pi m \sqrt{p}}{\varepsilon} + 1 \le \frac{2\pi m \sqrt{p}}{\varepsilon}$, for all $\varepsilon \in (0, 2]$. Thus

$$\log \mathcal{N}_2(\varepsilon, \mathcal{F}_{\sin}, x_{1:n}) \;\le\; \log\Big( \left\lceil \frac{\pi m \sqrt{p}}{\varepsilon} \right\rceil^p \Big) \;\le\; p \log\left( \frac{2\pi m \sqrt{p}}{\varepsilon} \right)$$

Hence for any $\alpha \in (0, 1]$,

$$\int_\alpha^2 \sqrt{\frac{\log \mathcal{N}_2(\varepsilon, \mathcal{F}_{\sin}, x_{1:n})}{n}}\, d\varepsilon \;\le\; \int_\alpha^2 \sqrt{\frac{p}{n}\log\left(\frac{2\pi m\sqrt{p}}{\varepsilon}\right)} d\varepsilon \le (2-\alpha)\sqrt{\frac{p}{n}\log\left(\frac{2\pi m\sqrt{p}}{\alpha}\right)}$$

Plugging this into equation 14 gives

$$\mathfrak{R}_S(\mathcal{F}_{\sin}) \;\le\; 4\alpha \;+\; 12(2-\alpha)\sqrt{\frac{p}{n}\log\left(\frac{2\pi m\sqrt{p}}{\alpha}\right)}$$

Choosing $\alpha = \frac{1}{\pi mn\sqrt{p}} \in (0, 1]$. Then

$$\log\Big(\frac{2\pi m\sqrt{p}}{\alpha}\Big) = \log\big(2\pi m\sqrt{p}\cdot \pi mn\sqrt{p}\big) = \log(2\pi^2 m^2 pn),$$

So

$$\mathfrak{R}_S(\mathcal{F}_{\sin}) \;\le\; \frac{4}{\pi mn\sqrt{p}} \;+\; 24\sqrt{\frac{p}{n}}\sqrt{\log\left(2\pi^2 m^2 pn\right)} = \widetilde{O}\left(\sqrt{\frac{p}{n}}\right). \tag{15}$$

Combining equation 12 and equation 15 we obtain, for every $S$,

$$\max_{j\in[p]} \mathfrak{R}_S(\mathcal{S}|_j) \;\le\; S_1 \mathfrak{R}_S(\mathcal{F}_{\sin}) = \widetilde{O}\left(S_1\sqrt{\frac{p}{n}}\right). \tag{16}$$

Fix $\delta_0 = \frac{1}{2}$, substituting equation 16 into equation 9 yields

$$\mathbb{P}\big[h_\theta(X) \neq Y\big] \;\le\; \widehat{\mathcal{R}}_\gamma(s^\theta) \;+\; \widetilde{O}\Big(\frac{S_1}{\gamma}\cdot\frac{p}{\sqrt{n}}\Big) \;+\; \widetilde{O}\Big(\frac{1}{\sqrt{n}}\Big).$$

□

### H.3 Margin bounds for ReLU MLP

**Lemma H.10.** *Let $Z \in \mathbb{R}^{p\times n}$ be the data matrix whose $i$-th column is $z_i = \sum_{k=1}^m e_{s_{i,k}} \in \{0, 1, \ldots, m\}^p$. Let $N_{j\ell} = \sum_{i=1}^n \mathbf{1}_{s_{i,j}=s_{i,\ell}}$. Then $\|Z\|_F^2 = \sum_{j=1}^m \sum_{\ell=1}^m N_{j\ell}$.*

*Proof.* Write $Z = \sum_{j=1}^m Z_j$ where $Z_j := \big(e_{s_{1,j}}, \ldots, e_{s_{n,j}}\big) \in \mathbb{R}^{p\times n}$. Then

$$\|Z\|_F^2 = \Big\langle \sum_{j=1}^m Z_j, \sum_{\ell=1}^m Z_\ell \Big\rangle_F = \sum_{j=1}^m\sum_{\ell=1}^m \mathrm{tr}(Z_j^\top Z_\ell).$$

For $r, c$,

$$(Z_j^\top Z_\ell)_{rc} = \sum_{s=1}^p (Z_j)_{sr}(Z_\ell)_{sc} = (e_{s_{r,j}})^\top e_{s_{c,\ell}},$$

so $(Z_j^\top Z_\ell)_{ii} = (e_{s_{i,j}})^\top e_{s_{i,\ell}} = \mathbf{1}_{s_{i,j}=s_{i,\ell}}$. Hence

$$\mathrm{tr}(Z_j^\top Z_\ell) = \sum_{i=1}^n \mathbf{1}\{s_{i,j} = s_{i,\ell}\} = N_{j\ell},$$

and substituting yields $\|Z\|_F^2 = \sum_{j=1}^m \sum_{\ell=1}^m N_{j\ell}$. □

**Lemma H.11** (Hoeffding bound)**.** *Assume that for each $i \in [n]$, the symbols $(s_{i,1}, \ldots, s_{i,m})$ are i.i.d. uniform on $[p] := \{1, \ldots, p\}$, and that they are independent across $i$. Let $z_i = \sum_{k=1}^m e_{s_{i,k}} \in \mathbb{R}^p$, $Z = (z_1, \ldots, z_n) \in \mathbb{R}^{p\times n}$, and $x^{(i)} := z_i$. Then for any $\delta' \in (0, 1)$, with probability at least $1 - \delta'$,*

$$\sum_{i=1}^n \|x^{(i)}\|_2^2 \;\le\; nm\Big(1 + \frac{m-1}{p}\Big) + m(m-1)\sqrt{\frac{n\log(1/\delta')}{2}},$$

*and therefore*

$$Q_2(S) \;:=\; \Big(\tfrac{1}{n}\sum_{i=1}^{n}\|x^{(i)}\|_2^2\Big)^{1/2} \;\le\; \overline{Q}_2(m,p,n,\delta') := \left[m\Big(1+\frac{m-1}{p}\Big)+m(m-1)\sqrt{\frac{\log(1/\delta')}{2n}}\right]^{1/2}.$$

*Proof.* For a fixed $i$, define

$$Y_i \;:=\; \sum_{j,\ell=1}^{m} \mathbf{1}\{s_{i,j}=s_{i,\ell}\}.$$

Note that $z_i=\sum_{k=1}^{m} e_{s_{i,k}}$ has coordinates $z_i(c)=\sum_{k=1}^{m}\mathbf{1}\{s_{i,k}=c\}$, hence

$$\|z_i\|_2^2=\sum_{c=1}^{p} z_i(c)^2=\sum_{c=1}^{p}\Big(\sum_{j=1}^{m}\mathbf{1}\{s_{i,j}=c\}\Big)\Big(\sum_{\ell=1}^{m}\mathbf{1}\{s_{i,\ell}=c\}\Big)=\sum_{j,\ell=1}^{m}\mathbf{1}\{s_{i,j}=s_{i,\ell}\}=Y_i.$$

Therefore $\sum_{i=1}^{n}\|x^{(i)}\|_2^2=\sum_{i=1}^{n}\|z_i\|_2^2=\sum_{i=1}^{n}Y_i$. Observe that

$$\mathbb{E}[Y_i]=\sum_{j=1}^{m}\mathbb{E}\,\mathbf{1}\{s_{i,j}=s_{i,j}\}+\sum_{\substack{j,\ell=1\\ j\neq\ell}}^{m}\mathbb{E}\,\mathbf{1}\{s_{i,j}=s_{i,\ell}\}=m+m(m-1)\cdot\mathbb{P}(s_{i,1}=s_{i,2}).$$

Since $s_{i,1},s_{i,2}$ are independent uniform on $[p]$, $\mathbb{P}(s_{i,1}=s_{i,2})=1/p$, hence

$$\mathbb{E}[Y_i]=m\Big(1+\frac{m-1}{p}\Big),\qquad \mathbb{E}\Big[\sum_{i=1}^{n}Y_i\Big]=n\,m\Big(1+\frac{m-1}{p}\Big).$$

Also notice that $m\le Y_i\le m^2$ and $(Y_i)_{i=1}^{n}$ are independent, let $S_n:=\sum_{i=1}^{n}Y_i$. Hoeffding's inequality for independent $Y_i\in[a_i,b_i]$ gives

$$\mathbb{P}(S_n-\mathbb{E}S_n\ge t)\;\le\;\exp\!\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)=\exp\!\left(-\frac{2t^2}{n\,(m^2-m)^2}\right).$$

Set the right-hand side to $\delta'$ and solve for $t$ to get

$$t\;=\;(m^2-m)\sqrt{\frac{n\log(1/\delta')}{2}}\;=\;m(m-1)\sqrt{\frac{n\log(1/\delta')}{2}}.$$

Therefore, with probability at least $1-\delta'$,

$$\sum_{i=1}^{n}\|x^{(i)}\|_2^2=\sum_{i=1}^{n}Y_i\le n\,m\Big(1+\frac{m-1}{p}\Big)\;+\;m(m-1)\sqrt{\frac{n\log(1/\delta')}{2}}.$$

Dividing by $n$ and taking square roots yields the stated bound on $Q_2(S)$. □

We now state and prove the width-independent multiclass margin bound for homogeneous activation. The main idea is to use $\ell_\infty$ contraction to reduce the problem to the real output, and then utilize a technical lemma from (Golowich et al., 2017). The core part of the proof is almost identical, and is included only for completeness.

**Lemma H.12** (Lemma 1 of (Golowich et al., 2017))**.** *Let $\sigma$ be a 1-Lipschitz, positive-homogeneous activation function which is applied element-wise (such as the ReLU). Then for any class of vector-valued functions $\mathcal{F}$, and any convex and monotonically increasing function $g:\mathbb{R}\to[0,\infty)$,*

$$\mathbb{E}_\epsilon\sup_{f\in\mathcal{F},W:\|W\|_F\le R} g\left(\left\|\sum_{i=1}^{m}\epsilon_i\sigma(Wf(x_i))\right\|_2\right)\le 2\cdot\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}} g\left(R\cdot\left\|\sum_{i=1}^{m}\epsilon_i f(x_i)\right\|_2\right).$$

**Theorem H.13** (Width-independent multiclass margin bound for homogeneous activation)**.** *Assume $p>m$ and $n>m^2$, $n\ge 17$, and $\sigma$ is a 1-Lipschitz, positive-homogeneous activation function. For any $\gamma>0$ and $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of the training sample $S$, the following holds simultaneously for all $\theta=(W,V)$ with $\|V\|_2\le S_2$ and $\|W\|_F\le B$,*

$$\mathbb{P}_{(X,Y)\in\mathcal{D}}\big[h_\theta(X)\neq Y\big]\;\le\;\widehat{\mathcal{R}}_\gamma(s^\theta)\;+\;\widetilde{O}\!\left(\frac{S_2B}{\gamma}\sqrt{\frac{p\,m}{n}}\right)\;+\;\widetilde{O}\!\left(\frac{1}{\sqrt{n}}\right).$$

*Here $\widetilde{O}(\cdot)$ hides factors polylogarithmic in $n$ and $\delta^{-1}$.*

*Proof of Theorem H.13.* The multiclass margin satisfies $|\phi_y(s) - \phi_y(s')| \le 2\|s - s'\|_\infty$ for all $s, s'$, hence $g_y := \psi_\gamma \circ \phi_y$ is $\frac{2}{\gamma}$-Lipschitz w.r.t. $\|\cdot\|_\infty$ and $|g_y| \le 1$.

**$\ell_\infty$-vector contraction**. For a vector class $\mathcal{S} \subset \{x \mapsto s(x) \in \mathbb{R}^p\}$ and $L$-Lipschitz maps $\{\varphi_i\}_{i=1}^n$ w.r.t. $\|\cdot\|_\infty$, a standard $\ell_\infty$ vector contraction inequality (see, e.g., Thm. 1 in (Foster & Rakhlin, 2019)) implies that for the fixed sample $S = (x^{(1)}, \dots, x^{(n)})$,

$$\mathfrak{R}_S(\varphi \circ \mathcal{S}) := \frac{1}{n}\mathbb{E}_\varepsilon\Big[\sup_{s\in\mathcal{S}} \sum_{i=1}^n \varepsilon_i\, \varphi_i(s(x^{(i)}))\Big] \;\le\; C\, L\, \sqrt{p}\, \max_{j\in[p]} \mathfrak{R}_S(\mathcal{S}|_j)\, \log^{\frac{3}{2}+\delta_0}\Big(\frac{\beta}{\max_j \mathfrak{R}_S(\mathcal{S}|_j)}\Big), \tag{17}$$

for any fixed $\delta_0 > 0$, with $C = C_{\delta_0} < \infty$. Here

$$\mathfrak{R}_S(\mathcal{S}|_j) \;:=\; \frac{1}{n}\,\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{S}} \sum_{i=1}^n \varepsilon_i\, s_j(x^{(i)})\right], \qquad \beta \;\ge\; \sup_\theta \max_i \big\{|\varphi_i(s^\theta(x^{(i)}))|,\; \|s^\theta(x^{(i)})\|_\infty\big\}.$$

Let $\mathcal{S} = \{s^\theta :\; \|V\|_2 \le S_2,\; \|W\|_F \le B\}$ and $\mathcal{S}|_j = \{x \mapsto v_j^\top \sigma(Wx) :\; \|V\|_2 \le S_2,\; \|W\|_F \le B\}$, where $v_j \in \mathbb{R}^d$ is the $j$-th row of $V$. Fix $\lambda > 0$, to be chosen later. For any fixed $x_{1:n}$, the Rademacher complexity can be upper bounded as

$$\begin{aligned}
n\,\mathfrak{R}_S(\mathcal{S}|_j) &= \mathbb{E}_\epsilon \sup_{\substack{\|V\|_2\le S_2\\ \|W\|_F\le B}} \sum_{i=1}^n \epsilon_i\, v_j^\top \sigma\Big(Wx^{(i)}\Big) \\
&\le \mathbb{E}_\epsilon \sup_{\substack{\|v_j\|_2\le S_2\\ \|W\|_F\le B}} \sum_{i=1}^n \epsilon_i\, v_j^\top \sigma\Big(Wx^{(i)}\Big) \qquad \text{(Cauchy–Schwarz)} \\
&\le \frac{1}{\lambda}\log \mathbb{E}_\epsilon \sup_{\substack{\|v_j\|_2\le S_2\\ \|W\|_F\le B}} \exp\left(\lambda \sum_{i=1}^n \epsilon_i\, v_j^\top \sigma\Big(Wx^{(i)}\Big)\right) \\
&\le \frac{1}{\lambda}\log \mathbb{E}_\epsilon \sup_{\substack{\|v_j\|_2\le S_2\\ \|W\|_F\le B}} \exp\left(\|v_j\|_2 \cdot \lambda \left\|\sum_{i=1}^n \epsilon_i\, \sigma\Big(Wx^{(i)}\Big)\right\|_2\right) \\
&\le \frac{1}{\lambda}\log \mathbb{E}_\epsilon \sup_{\|W\|_F\le B} \exp\left(S_2 \cdot \lambda \left\|\sum_{i=1}^n \epsilon_i\, \sigma\Big(Wx^{(i)}\Big)\right\|_2\right).
\end{aligned}$$

Applying Lemma H.12 with the given 1-Lipschitz, positive-homogeneous $\sigma$, $\mathcal{F} = \{f :\; f(x) = x\}$ (identity class), and $g(t) = \exp(S_2\lambda t)$, we obtain

$$\frac{1}{\lambda}\log \mathbb{E}_\epsilon \sup_{\|W\|_F\le B} \exp\left(S_2 \cdot \lambda \left\|\sum_{i=1}^n \epsilon_i \sigma\Big(Wx^{(i)}\Big)\right\|_2\right) \;\le\; \frac{1}{\lambda}\log\left(2\,\mathbb{E}_\epsilon \exp\left(S_2 \cdot \lambda B \left\|\sum_{i=1}^n \epsilon_i x^{(i)}\right\|_2\right)\right).$$

Denote $M = S_2 B$, and define the random variable (as a function of $\epsilon = (\epsilon_1, \dots, \epsilon_n)$):

$$Z \;=\; M \cdot \left\|\sum_{i=1}^n \epsilon_i x^{(i)}\right\|_2.$$

Then

$$\frac{1}{\lambda}\log\left(2\,\mathbb{E}_\epsilon \exp(\lambda Z)\right) \;=\; \frac{\log 2}{\lambda} + \frac{1}{\lambda}\log\left(\mathbb{E}_\epsilon \exp\left(\lambda(Z - \mathbb{E}Z)\right)\right) + \mathbb{E}Z.$$

By Jensen's inequality,

$$\mathbb{E}Z \;\le\; M\sqrt{\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^n \epsilon_i x^{(i)}\right\|_2^2\right]} = M\sqrt{\mathbb{E}_\epsilon\left[\sum_{i,i'=1}^m \epsilon_i\epsilon_{i'} x_i^\top x_{i'}\right]} \;=\; M\sqrt{\sum_{i=1}^n \|x^{(i)}\|_2^2}.$$

Moreover, $Z$ satisfies a bounded-difference condition

$$Z(\epsilon_1, \dots, \epsilon_i, \dots, \epsilon_n) - Z(\epsilon_1, \dots, -\epsilon_i, \dots, \epsilon_n) \le 2M\|x^{(i)}\|_2,$$

and hence is sub-Gaussian with variance factor $v = M^2 \sum_{i=1}^n \|x^{(i)}\|_2^2$, yielding

$$\frac{1}{\lambda} \log \left( \mathbb{E}_\epsilon \exp \lambda (Z - \mathbb{E}Z) \right) \leq \frac{\lambda M^2}{2} \sum_{i=1}^n \|x^{(i)}\|_2^2.$$

Choosing $\lambda = \frac{\sqrt{2 \log 2}}{M \sqrt{\sum_{i=1}^n \|x^{(i)}\|_2^2}}$ gives

$$\frac{1}{\lambda} \log \left( 2 \cdot \mathbb{E}_\epsilon \exp(\lambda Z) \right) \;\leq\; M \left( \sqrt{2 \log 2} + 1 \right) \sqrt{\sum_{i=1}^n \|x^{(i)}\|_2^2}.$$

Therefore,

$$\mathfrak{R}_S(\mathcal{S}|_j) \;\leq\; S_2 B \left( \sqrt{2 \log 2} + 1 \right) \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n \|x^{(i)}\|_2^2}\,. \tag{18}$$

**Controlling $\max_j \mathfrak{R}_S(\mathcal{S}|_j)$ and the log term.** Define the "good" subset

$$\mathcal{X}^n_{\text{good}}(\delta') := \Big\{ x_{1:n} \in \mathcal{X}^n : \; \tfrac{1}{n} \sum_{i=1}^n \|x^{(i)}\|_2^2 \leq \overline{Q}_2(m, p, n, \delta')^2 \Big\}.$$

By Lemma H.11, with probability $\geq 1 - \delta'$ the realized sample satisfies $x_{1:n} \in \mathcal{X}^n_{\text{good}}(\delta')$. On this event, equation 18 yields

$$0 \leq \max_{j \in [p]} \mathfrak{R}_S(\mathcal{S}|_j) \;\leq\; S_2 B \left( \sqrt{2 \log 2} + 1 \right) \frac{1}{\sqrt{n}} \overline{Q}_2(m, p, n, \delta'). \tag{19}$$

Furthermore, for any $\theta$ and $x$, $\|s^\theta(x)\|_\infty \leq \|V\|_2 \, \|\sigma(Wx)\|_2 \leq S_2 B \|x\|_2$, and since here $x \in \{0, 1, \ldots, m\}^p$ with $\|x\|_1 = m$, we have $\|x\|_2 \leq m$. Thus we may take the simple, deterministic bound

$$\beta \;\leq\; 1 + S_2 B \, m.$$

To upper bound the logarithm in equation 17 more conveniently, also define

$$b \;:=\; 1 + S_2 B \sqrt{n} \, \overline{Q}_2(m, p, n, \delta')\,,$$

so that $\beta \leq b$ and hence $\log(\beta/t) \leq \log(b/t)$ for all $t > 0$.

Applying equation 17 with $L = 2/\gamma$ and using equation 19, we obtain on the event of Lemma H.11

$$\begin{aligned} \mathfrak{R}_S(\mathcal{F}_\gamma) &\leq C \, \frac{2}{\gamma} \sqrt{p} \, \max_{j \in [p]} \mathfrak{R}_S(\mathcal{S}|_j) \, \log^{\frac{3}{2} + \delta_0} \Big( \frac{\beta}{\max_j \mathfrak{R}_S(\mathcal{S}|_j)} \Big) \\ &\leq C \, \frac{2}{\gamma} \sqrt{p} \, \max_{j \in [p]} \mathfrak{R}_S(\mathcal{S}|_j) \, \log^{\frac{3}{2} + \delta_0} \Big( \frac{b}{\max_j \mathfrak{R}_S(\mathcal{S}|_j)} \Big). \end{aligned}$$

Let

$$h(t) \;=\; t \, \log^a \Big( \frac{b}{t} \Big), \qquad a := \tfrac{3}{2} + \delta_0 > \tfrac{3}{2}\,.$$

Substituting $\delta_0 = 0.5$ gives $a = 2$. From equation 19, with $t := \max_j \mathfrak{R}_S(\mathcal{S}|_j)$ we have

$$t \;\leq\; \frac{\sqrt{2 \log 2} + 1}{\sqrt{n}} \, S_2 B \, \overline{Q}_2(m, p, n, \delta') \;=\; \frac{\sqrt{2 \log 2} + 1}{n} \, (b - 1) \;\leq\; \frac{\sqrt{2 \log 2} + 1}{n} \, b.$$

Since $n \geq 17 \geq e^2 (\sqrt{2 \log 2} + 1)$, we have $t \leq b \, e^{-2}$; on $[0, \, b e^{-2}]$ the function $h$ is increasing, hence

$$h(t) \;\leq\; h\Big( \tfrac{\sqrt{2 \log 2} + 1}{n} \, b \Big) \;=\; \frac{\sqrt{2 \log 2} + 1}{n} \, b \, \log^2 \Big( \frac{b}{b(\sqrt{2 \log 2} + 1)/n} \Big) \;=\; \frac{\sqrt{2 \log 2} + 1}{n} \, b \, \log^2 \Big( \frac{n}{\sqrt{2 \log 2} + 1} \Big).$$

Therefore, for some absolute $C' > 0$,

$$\mathfrak{R}_S(\mathcal{F}_\gamma) \;\leq\; C' \, \frac{1}{\gamma} \sqrt{\frac{p}{n}} \, S_2 B \, \overline{Q}_2(m, p, n, \delta') \, \log^2 \Big( \frac{n}{\sqrt{2 \log 2} + 1} \Big). \tag{20}$$

**Final bound.** By Lemma H.11, with probability at least $1-\delta'$,

$$\overline{Q}_2(m,p,n,\delta')^2 = m\left(1+\frac{m-1}{p}\right) + m(m-1)\sqrt{\frac{\log(1/\delta')}{2n}} \;\le\; 2m + m\sqrt{\log(1/\delta')},$$

where we used $p > m$ and $n > m^2$. Hence $\overline{Q}_2(m,p,n,\delta') = \widetilde{O}(\sqrt{m})$. Since $\frac{1}{n}\sum_{i=1}^n \psi_\gamma(\phi_{y^{(i)}}(s^\theta(x^{(i)}))) \le \widehat{\mathcal{R}}_\gamma(s^\theta)$, combining equation 8 and equation 20, and taking a union bound with the choice $\delta' = \delta/2$ while applying Corollary H.4 with confidence parameter $\delta/2$, yields the stated result with overall probability at least $1-\delta$. □

*Remark* H.14 (Data-dependent specialization). The bound is width-independent and depends on the sample only through $Q_2(S)$. In our setup, $x \in \{0,1,\dots,m\}^p$ with $\|x\|_1 = m$; thus $\|x\|_2 \le m$, so $\beta \le 1 + S_2 B m$ deterministically. We further used distributional assumptions on $s_{1:m}$ (e.g., i.i.d. uniform over $[p]$) only to obtain sharper high-probability bounds on $Q_2(S)$.

We are now able to prove theorems in Section6:

*Proof of Theorem 6.2.* The proof consists of showing all networks with small training error and small normalized margin generalize, and at least one such network exist.

In Theorem H.9, set $\gamma = \gamma_\theta(\mathcal{D}_{\text{train}})$, then the empirical $\gamma$-margin error is

$$\widehat{\mathcal{R}}_\gamma\big(s^\theta\big) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left[f(x_i)_{y_i} \le \gamma + \max_{j\neq y_i} f(x_i)_j\right] = 0.$$

Notice that $\overline{\gamma}_\theta = \frac{\gamma_\theta(\mathcal{D}_{\text{train}})}{\|V\|_{1,\infty}}$, by Theorem H.9,

$$\mathbb{P}_{(X,Y)\in\mathcal{D}}\big[h_\theta(X)\neq Y\big] \;\le\; \widetilde{O}\left(\frac{1}{\overline{\gamma}_\theta}\, p\sqrt{\frac{1}{n}}\right) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right) \le \widetilde{O}\left(p\sqrt{\frac{1}{n}}\right) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right) = \widetilde{O}\left(p\sqrt{\frac{1}{n}}\right).$$

When $2p \le d$, Section F.1.2 gives a network whose normalized margin is

$$\overline{\gamma}_\theta \;=\; \frac{\gamma_\theta(\mathcal{D}_{\text{train}})}{\|V\|_{1,\infty}} \;\ge\; \frac{p}{2p} \;=\; \frac{1}{2} = \Omega(1).$$

□

*Proof of Theorem 6.3.* In Theorem H.13, set $\gamma = \gamma_\theta(\mathcal{D}_{\text{train}})$. Then the empirical $\gamma$-margin error is zero,

$$\widehat{\mathcal{R}}_\gamma\big(s^\theta\big) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left[f(x_i)_{y_i} \le \gamma + \max_{j\neq y_i} f(x_i)_j\right] = 0,$$

and Theorem H.13 gives

$$\mathbb{P}_{(X,Y)\in\mathcal{D}}\big[h_\theta(X)\neq Y\big] \;\le\; \widetilde{O}\left(\frac{1}{\overline{\gamma}_\theta}\sqrt{\frac{p\,m}{n}}\right) \;+\; \widetilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

Apply Theorem F.6 with $\tau = 0.1$, which yields a margin $\gamma(x) \ge 0.6p$ on $\mathcal{X}$ and width $d \le p\,C_m$, where

$$C_m \;=\; 13\,m\,2^m\left(m\sqrt{10em}\,(1+2em)^{\frac{m-1}{2}} + 2\right).$$

Using $(1+2em)^{(m-1)/2} \le (2em)^{(m-1)/2} e^{1/(4e)}$, we obtain

$$C_m \;\le\; 26\sqrt{5}\,e^{\frac{1}{4e}}\; m^{\frac{m}{2}+2}\,(\sqrt{8e})^m \;\le\; 64\,m^{\frac{m}{2}+2}\,(4.67)^m.$$

Thus the width condition in the statement $d \ge 64\,p\,m^{\frac{m}{2}+2}(4.67)^m$ is sufficient for $d \ge p\,C_m$.

From equation 3–equation 4 in Theorem F.6,

$$\|W\|_F \;\le\; \frac{2}{m}p\sqrt{C_m}, \qquad \|V\|_2 \;\le\; \sqrt{2p}\,\sqrt{C_m}\,\frac{(m+\frac{1}{2})m^{2m}}{m!\,2^m}.$$

Write

$$K_m \;:=\; C_m \,\frac{(m+\frac{1}{2})m^{2m}}{m!\,2^m}.$$

Using Stirling's lower bound $m! \geq \sqrt{2\pi m}\,(m/e)^m$ and the same $(1+2em)$ bound as above gives the clean upper bound

$$K_m \;\leq\; \frac{39\sqrt{5}\,e^{1/(4e)}}{\sqrt{4\pi e}}\; m^{1.5m+2.5}\,(\sqrt{2}\,e^{3/2})^m \;\leq\; 17\,m^{1.5m+2.5}\,(6.34)^m.$$

Consequently,

$$\|V\|_2\|W\|_F \;\leq\; 2\sqrt{2}\,p\sqrt{p}\,K_m,$$

and

$$\overline{\gamma}_\theta \;=\; \frac{\gamma_\theta(\mathcal{D}_{\text{train}})}{\|V\|_2\|W\|_F} \;\geq\; \frac{0.6\,p}{2\sqrt{2}\,p\sqrt{p}\,K_m} \;=\; \frac{0.3}{\sqrt{2}}\cdot\frac{1}{K_m\sqrt{p}} = \Omega\left(\frac{1}{\sqrt{p}}\cdot\frac{1}{m^{1.5m+2.5}\,(6.34)^m}\right).$$

□