NONLINEARITY–PHASE–GENERALIZATION THEORY: OOD BOUNDS FOR KANS AND MLPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent theory has sharpened OOD generalization, yet most analyses are weight-space and assume fixed activations, offering little guidance when nonlinearities are learned, as in Kolmogorov-Arnold Networks (KANs). We develop Nonlinearity–Phase–Generalization (NPG): a function-space framework that links per-nonlinearity smoothness (total variation of the derivative, TVD) to phase preservation (via global cross-bispectrum, GCB) and, in turn, to the source–target risk gap. The resulting bounds are finite-width, additive in depth, and architecture-aware, placing KANs and MLPs on common footing. NPG yields actionable rules: select smaller-TVD nonlinearities; for polynomial KANs, keep degree$\times$range small and decorrelate layers; within MLPs, Softplus reduces the model term vs ReLU under bounded inputs. Controlled PACS/VLCS studies follow these predictions (e.g., OOD gap: 21.85 vs 26.53 on PACS; 9.87 vs 12.67 on VLCS), and TVD/GCB operate as training diagnostics. By tying nonlinearity design to OOD error, NPG enables principled architecture choices and reproducible tuning under domain shift. Code is available at: **https://github.com/*****

1 INTRODUCTION

Out-of-distribution (OOD) generalization has advanced substantially, clarifying behavior under domain shift and informing robust training practices. However, most analyses reason in weight space and assume fixed nonlinearities, which leaves open how to assess architectures with learned nonlinearities, notably, Kolmogorov–Arnold Networks (KANs), which is shown to be competitive across modalities. As a result, we still lack a principled answer to a practical question: **when (if ever) should we expect KANs to generalize OOD better than MLPs?** Failing to answer this costs us reliable design principles for safety-critical deployment and blurs comparisons across architectures.

Why existing tools are insufficient. Fig 1(a) summarizes two dominant lines and their limitations for KANs. Model-agnostic approaches (e.g., H-divergence) cannot distinguish architectures, producing bounds that are too loose to guide design. Model-aware analyses (PAC-Bayes, NTK, norm/margin) assume fixed nonlinearities and/or infinite width; when nonlinearities are learned, the required constants become parameter-dependent and depth-wise multiplicative, yielding exploding or non-informative guarantees at realistic depth. More fundamentally, weight-space metrics are function-space blind: they do not connect how we design nonlinearities to the invariances that matter under domain shift.

Our premise and framework. We introduce Nonlinearity–Phase–Generalization (NPG) (Fig. 1(b)), which puts KANs and MLPs on the same footing via Fourier phase. The premise, supported by classical signal processing and deep-learning evidence, is that phase is relatively domain-stable and label-informative (formalized notions in Section 3.2). NPG connects nonlinearity-level properties to OOD generalization in two steps: *1. Nonlinearity $\rightarrow$ Phase.* It quantifies how each nonlinearity scrambles phase using a bispectral measure and shows that this disturbance is controlled additively in depth by a per-nonlinearity smoothness proxy: the total variation of the derivative (TVD). *2. Phase $\rightarrow$ Generalization.* It links phase preservation to a shrinking source-target risk gap, yielding nonlinearity-aware OOD bounds that are additive in depth, valid at finite and infinite width, and that operate in function space. Composed, these steps produce a predictive, depth-explicit bounds that link nonlinearity design to OOD error for both KANs and MLPs.

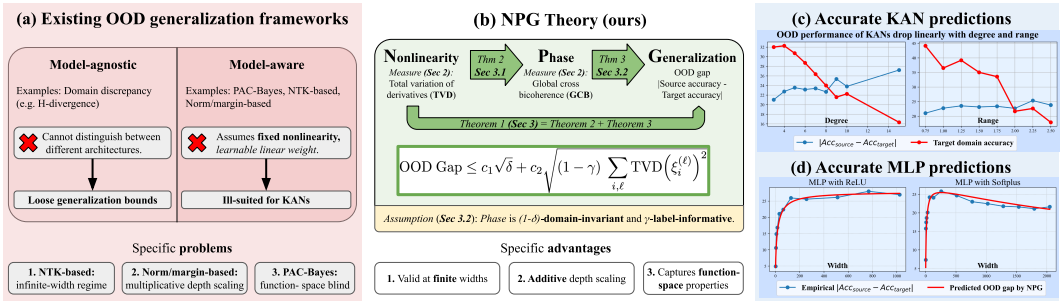


Figure 1: (a) Existing OOD theories: model-agnostic (loose) vs model-aware (fixed nonlinearity); neither certifies KANs. (b) NPG: unifying; width-robust; additive in depth and captures function-space properties. (c) KAN prediction accuracy: OOD accuracy drops linearly with polynomial degree and input range, as NPG predicts. (d) MLP prediction accuracy: NPG accurately predicts the scaling of MLP networks, and ranks $\mathrm{Softplus} > \mathrm{ReLU}$ for OOD generalization.

Predictive consequences. Specializing NPG to polynomial KANs vs ReLU MLPs, we show that, under normalized inputs and weight decay, polynomial KANs can achieve strictly smaller TVD than ReLU when the degree–range product is small, leading to tighter OOD guarantees. NPG predicts that OOD performance will degrade linearly with d and ρ, *cf.* Fig. 1(c), and within MLPs it predicts that $\mathrm{Softplus}$ should outperform ReLU under bounded inputs because its TVD is smaller, *cf.* Fig. 1(d).

The theory's predictions match behavior in controlled studies: On PACS and VLCS, polynomial KANs outperform ReLU baselines in leave-one-domain-out setting, and their accuracy drops linearly with $O(d\rho)$. Within MLPs, $\mathrm{Softplus} > \mathrm{ReLU}$ for OOD generalization when linear-layer norms are small; the same constants estimated on ReLU transfer to predict $\mathrm{Softplus}$ gains.

Our contributions: **1. Novel NPG theory.** A rigorous predictive theory that connects the total variation of learned nonlinearities to a phase measure, yielding depth-explicit OOD generalization bounds. **2. First theoretical comparison of KANs vs MLPs in OOD generalization.** Specialization to ReLU vs polynomial nonlinearities establishes when KANs admit strictly tighter OOD generalization guarantees. **3. Practical training diagnostic** per-layer TVD and a cross-bispectral score that correlate with OOD behavior and provide actionable guidance for nonlinearity design. **4. Empirical validation on standard DG settings.** KAN vs MLP (PACS/VLCS), degree/range sweeps for KANs and several common fixed nonlinearity choices within MLPs.

2 Preliminaries

Background. *Kolmogorov-Arnold Networks (Liu et al., 2024).* Unlike MLPs, which place fixed element-wise nonlinearities on nodes, KANs place *learnable, univariate* nonlinearities on edges. Formally, a KAN realizes a composition $\mathcal{N} = \Psi_L \circ \cdots \circ \Psi_1$ where each layer operator Ψ_i maps $\mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ via learnable univariate functions $\{\psi_{j,k}^{(i)}\}_{j \le d_i, k \le d_{i-1}}$ as $\left[\Psi_i(x)\right]_j = \sum_k \psi_{j,k}^{(i)}(x_k)$, in contrast to an MLP layer with a *fixed* nonlinearity σ and learnable *scalar* weights.

Fourier transform and phase. For a signal $x(t)$, its Fourier transform is $X(\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-i\omega t}\, dt$, which admits the polar form $X(\omega) = \mathcal{A}(\omega)\, e^{i\phi_X(\omega)}$, where $\phi_X(\omega)$ is the *Fourier phase*. For discrete signals, $x[n]$, the DFT is denoted $X[k]$; for a network output $z = \mathcal{N}(x)$ we write $z[n]$ and $Z[k]$.

Normalized cross-bispectrum is given by $\mathrm{cbs}_{xxy}(\omega_1, \omega_2) = \dfrac{U_X(\omega_1\omega_2)\, Y(\omega_1+\omega_2)}{\sqrt{P_X(\omega_1)\, P_X(\omega_2)\, P_Y(\omega_1+\omega_2)}}$ where $U_X(\omega_1, \omega_2) = (X(\omega_1)\, X(\omega_2) - \mathbb{E}[X(\omega_1)\, X(\omega_2)])$ and $P_X(\omega) = \mathbb{E}\left[|X(\omega)|^2\right]$ denotes the power spectrum. *Principal simplex* PS. Due to symmetries of cbs_{xxy}, information is redundantly represented over (ω_1, ω_2). The nonredundant region PS is called the *principal simplex*.

Input space and OOD domains $(\mathcal{X}, \mathcal{D})$. Input space $\mathcal{X}$ is the set from which inputs x are drawn. Family of OOD domains $\mathcal{D}$ is a collection of related data distributions over $\mathcal{X}$ that instantiate the OOD setting. Each $Q \in \mathcal{D}$ denotes a specific domain (e.g., photo, art, cartoon, sketch in PACS), i.e., a probability distribution on $\mathcal{X}$. The source and target distributions, P and Q, may

differ while both belonging to $\mathcal{D}$. *Hypothesis, loss, and risk.* Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ an arbitrary loss function. For a distribution R on $\mathcal{X} \times \mathcal{Y}$, the expected risk is $\mathcal{E}_R(h) := \mathbb{E}_{(x,y)\sim R}\big[\ell\big(h(x), y\big)\big]$. No specific structure on h, ℓ, or $\mathcal{E}_R$ is assumed unless stated otherwise. *Mutual information.* For random variables X and Y, the mutual information $I(X;Y) = H(Y) - H(Y \mid X)$ measures the reduction in uncertainty about Y after observing X.

Measures. *Nonlinearity measure.* Total variation of derivative (TVD) for functions $f : \mathbb{R} \to \mathbb{R}$ is defined as $\int |df'|$ when f' has bounded variation. For convenience, with slight abuse of notation, we also use $\mathrm{TVD}(\sigma(\|A_{i:}\|_2 u + b_i)) = \|A_{i:}\|_2 \,\mathrm{TVD}(\sigma)$ for MLPs, and $\mathrm{TVD}(\psi_i) = \sqrt{\sum_j \mathrm{TVD}(\psi_{ij})^2}$ for KANs. *Phase measure.* Global cross bispectrum (GCB) is calculated by integrating cbs over the principal simplex: $\mathrm{GCB} = \int_{\mathsf{PS}} |\mathrm{cbs}_{xxy}(\omega_1, \omega_2)|^2 d\omega_1 d\omega_2$. $\mathrm{GCB}(h)$ implies $y = h(x)$.

3 Nonlinearity-Phase-Generalization Theory and Implications

The section opens with the main theoretical result (Theorem 1). **Sec. 3.1** establishes the nonlinearity-phase control (Theorem 2). **Sec. 3.2** introduces domain invariance and label informativity of phase (Defs. 1, 2), posits Assumption 1, and derives the phase-gap bound (Theorem 3). **Sec. 3.3** instantiates TVD for concrete nonlinearities: ReLU and learnable polynomials (Props. 1, 2, 3, 4), summarized in Lemma 1. **Sec. 3.4** composes these constants with Theorem 1 to produce the theoretical ReLU-MLP vs. polynomial-KAN comparison (Theorem 4) and the resulting design rule.

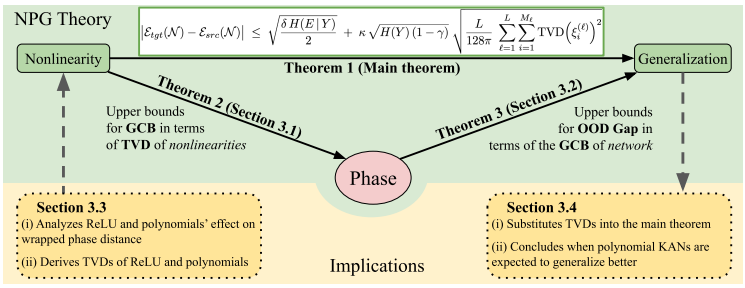


Figure 2: Method overview: nonlinearity → phase (Thm. 2: TVD controls GCB) → generalization (Thm. 3: GCB bounds OOD gap), composing to the main TVD bound (Thm. 1); Sections3.3–3.4 instantiate TVD for ReLU/polynomials and yield the KAN–MLP comparison.

Theorem 1 (**Main theorem**, TVD bound for $|\mathcal{E}_{src} - \mathcal{E}_{tgt}|$)**.** *Let $\mathcal{N}$ be a depth-L network, $\mathcal{N} = \xi^{(L)} \circ \xi^{(L-1)} \circ \cdots \circ \xi^{(1)}$ and M_ℓ denote the width of the layer $\xi^{(\ell)}$. Under Assumption 1 with κ as defined there, for environments E where Fourier phase has $(1-\delta)$-domain-invariance and γ-label-informativity,*

$$\big|\mathcal{E}_{tgt}(\mathcal{N}) - \mathcal{E}_{src}(\mathcal{N})\big| \;\le\; \sqrt{\frac{\delta\, H(E\,|\,Y)}{2}} \;+\; \kappa \sqrt{H(Y)\,(1-\gamma)} \sqrt{\frac{L}{128\pi} \sum_{\ell=1}^{L} \sum_{i=1}^{M_\ell} \mathrm{TVD}\Big(\xi_i^{(\ell)}\Big)^2} \quad (1)$$

where $\xi_i^{(\ell)}(x) = \sigma(\|A_{i:}x + b_i)$ for MLPs and $\xi_i^{(\ell)}(x) = \sum_j \psi_{ij}^{(\ell)}(x_j)$ for KANs. If layer bispectra are mutually orthogonal, replace $L/128\pi$ by $1/128\pi$.

Proof. Directly follows from Theorems 2 and 3.

This decomposes the OOD error into a domain term and a model term that depends *only* on the nonlinearities. **Consequences.** 1. A *unified* treatment of KANs and MLPs and is *function-space aware*, *finite-width valid*, and *depth-additive*. 2. To keep the model term fixed as depth L grows, it suffices to keep $\sum_{\ell,i} \mathrm{TVD}(\xi_i^{(\ell)})^2$ controlled, so that deeper nets remain OOD-stable when nonlinearities are smoothed at an appropriate rate. 3. Design guidance for fixed nonlinearities, e.g., $\mathrm{TVD}(\mathrm{ReLU}) = 1$ while $\mathrm{TVD}(\mathrm{Softplus}) = \tanh(\beta\rho/2) < 1$, so Softplus yields a strictly smaller model term than ReLU. 4. Because the model term is driven by TVD/GCB, it can be acted on: penalize TVD of learned nonlinearities; promote inter-layer decorrelation (orthogonal init/regularizers) to remove the L prefactor; monitor TVD/GCB during training to predict and prevent OOD degradation.

3.1 From Nonlinearities to Phase

Theorem 2 establishes the nonlinearity–phase link in a depth-explicit manner: GCB summarizes network-level phase scrambling, while TVD controls each layer's ability to create coupling. Weight–space proxies miss this nonlinear phase coupling that drives OOD failures. This formulation captures function-space properties, which are key to distinguishing between nonlinearities.

Theorem 2 (TVD bound for GCB)**.** *Let $\mathcal{N}$ be a depth-L network, $\mathcal{N} = \xi^{(L)} \circ \xi^{(L-1)} \circ \cdots \circ \xi^{(1)}$ and M_ℓ denote the width of the layer $\xi^{(\ell)}$. For i.i.d. and Gaussian inputs with zero-mean and unit-variance, the global cross-bicoherence of $\mathcal{N}$ is bounded by*

$$\mathrm{GCB}(\mathcal{N}) \leq \frac{L}{128\pi} \sum_{\ell=1}^{L} \sum_{i=1}^{M_\ell} \mathrm{TVD}\left(\xi_i^{(\ell)}\right)^2 \tag{2}$$

Specifically, the multiplying constant L drops when bispectra of layers are mutually orthogonal.

The detailed proof is presented in the Appendix C.1. The network's phase scrambling is controlled by a sum of squared layer TVDs. In the near-orthogonal case, the extra prefactor vanishes. **Consequences**: 1. choosing nonlinearities with smaller TVDs directly suppresses phase scrambling; 2. encouraging inter-layer decorrelation prevents worst-case growth; 3. TVD becomes an actionable diagnostic that can be used to keep GCB small.

3.2 From Phase to OOD Generalization

To make precise the idea that phase may improve OOD generalization, we separate two properties: $(1-\delta)$-domain invariance and γ-label-informativity. These definitions clarify the underlying assumptions of the field's hypothesis that emphasizing phase features improves OOD generalization.

Let E be a discrete domain index, X an input, and Y a label. A *phase feature map* is a measurable $\Phi : \mathcal{X} \to \mathbb{R}^m$ that outputs phase–based statistics of X (e.g., bicoherence, projections). All definitions depend only on the joint law of $(E, \Phi(X), Y)$ not on how Φ is implemented.

Definition 1 ($(1-\delta)$-domain-invariance)**.** *The (normalized) class-conditional domain leakage is*

$$\delta := \frac{I(E;\, \Phi(X) \,|\, Y)}{H(E \,|\, Y)} \in [0, 1]. \tag{3}$$

The phase features are said to have $(1-\delta)$-domain-invariance.

With this definition, $\delta = 0$ describes the perfect phase covariate shift scenario, $\Phi(X) \perp\!\!\!\perp E \mid Y$, that is when class-conditional phase features do not move across domains. Larger δ means more domain information leaks through $\Phi(X)$ even after conditioning on the label.

Definition 2 (γ-label-informativity)**.** *The (normalized) label information explained by phase is*

$$\gamma \;:=\; \frac{I(Y;\, \Phi(X))}{H(Y)} \in [0, 1]. \tag{4}$$

The phase features are said to have γ-label-informativity.

γ measures the fraction of label uncertainty removed by observing phase features. Specifically $\gamma = 1$ if and only if phase alone determines the label, $H(Y \,|\, \Phi(X)) = 0$.

When $\delta \approx 0$, the optimal phase-based Bayesian classifier $\mathcal{N}^\star$ is invariant across E. If moreover $\gamma \approx 1$, any residual OOD error is due to deviations of a learned predictor $\mathcal{N}$ from $\mathcal{N}^\star$ along phase features, summarized by GCB. This motivates the following assumption.

Assumption 1 (phase-disagreement control)**.** *Under $(\delta, \gamma) \approx (0, 1)$, there exists $\kappa > 0$ such that, for any target environment with the same label mechanism,*

$$\mathbb{P}(\mathcal{N}(\Phi(X)) \neq \mathcal{N}^\star(\Phi(X))) \;\leq\; \frac{\kappa^2}{2}\, \mathrm{GCB}(\mathcal{N}).$$

Theorem 3 converts these into an upper bound on the OOD error which tightens to a $\sqrt{\mathrm{GCB}}$ law.

Theorem 3 (GCB bound for $|\mathcal{E}_{src} - \mathcal{E}_{tgt}|$)**.** *Let δ, γ, κ be as in Assumption 1. Then,*

$$\left|\mathcal{E}_{tgt}(\mathcal{N}) - \mathcal{E}_{src}(\mathcal{N})\right| \;\le\; \sqrt{\frac{\delta\, H(E\,|\,Y)}{2}} \;+\; \kappa\,\sqrt{\mathrm{GCB}(\mathcal{N})\, H(Y)\,(1-\gamma)}. \tag{5}$$

Proof idea. (i) Decompose error around the Bayes rule via the exact excess-risk identity and bound the disagreement term by Cauchy–Schwarz, controlling its factors by $H(Y \mid U)$ and the phase-disagreement control. (ii) Control the Bayes error shift between $\mathcal{E}_{src}$ and $\mathcal{E}_{tgt}$ by class-conditional total variation; upper bound it with Pinsker and relate the resulting KL sums to $I(E;U \mid Y) = \delta\, H(E \mid Y)$ using the nondegenerate environment prior. (iii) Symmetrize to obtain the absolute difference.

The bound splits the source→target gap into two parts: **Domain term.** $\sqrt{\delta\, H(E\,|\,Y)/2}$ measures class-conditional phase shift; it is 0 when $\delta = 0$. **Model term.** $\sqrt{\mathrm{GCB}(\mathcal{N})\, H(Y)\,(1-\gamma)}$ measures phase mismatch; it shrinks as phase becomes label-informative and as the network preserves phase (small GCB). **Consequences**: 1. When $(\delta, \gamma) \approx (0, 1)$, the gap scales as $\sqrt{\mathrm{GCB}}$; 2. If δ is large or γ small, phase-centric modeling will not close the OOD gap; improve invariance or label signal first.

3.3 Case Study: ReLU versus Learnable Polynomial Nonlinearities

This section quantifies how concrete nonlinearities distort phase and converts those distortions into measurable TVD constants that feed the main bound. The analysis isolates input-dependent effects and coefficient control (weight decay), producing nonlinearity-level design rules.

A nonlinearity alters phase via cross-frequency coupling; here, the magnitude of the induced phase shift is used for the main bounds. The next result makes this explicit for ReLU by tying distortion to the negative-mass ratio $X^-[k]/X[k]$.

Proposition 1 (ReLU's effect on phase)**.** *Decompose $x = x^+ + x^-$ where x^+ (x^-) denotes the positive (negative) parts of the signal. Then* $\mathrm{ReLU}(x) = x^+$*, yielding the phase distortion*

$$|\mathrm{wrap}(\phi_Z[k] - \phi_X[k])| \le \phi\left(1 - \frac{X^-[k]}{X[k]}\right), \tag{6}$$

and for small $\frac{X^-[k]}{X[k]}$, $\phi\left(1 - \frac{X^-[k]}{X[k]}\right) \approx \arcsin \frac{X^-[k]}{X[k]}$.

Fig. 6 (a) demonstrates the tightness of the bounds. Proposition 1 suggests that lower (arcsin) and upper (arccos) bounds for total phase error for ReLU are determined by input statistics. ReLU induces a unit jump in σ' and a phase error controlled only by input sign asymmetry; leaving little architectural control beyond input centering and clipping.

For learnable polynomials (denoted p_d), explicit coefficient control enables tunable phase-distortion.

Proposition 2 (Polynomials' effect on phase)**.** *Let p_d denote a polynomial nonlinearity of degree d. Then $z[n] = p_d(x[n])$ and $d[n] = \sum_{r\neq 1} a_r x^r[n] + (a_1 - 1)x[n]$. Let $C = |a_0| + |a_1 - 1| + \cdots + |a_d|$. Then the energy in $d[n]$ is $O(C^2)$. The phase distortion is*

$$\sum_{k=0}^{N-1} \left|\mathrm{wrap}\big(\phi_Z[k] - \phi_X[k]\big)\right| \le N^2 C. \tag{7}$$

For polynomials, phase distortion scales with the aggregate coefficient mass C, yielding $\mathrm{TVD}(p_d) \lesssim c\,C$ for some c. Because worst-case bounds can be pessimistic, a concentration envelope is required to capture typical behavior under i.i.d. bounded inputs.

Proposition 3 (Phase envelope for bounded iid inputs)**.** *Let $x[n]$ be i.i.d. zero-mean unit-variance random variables, and $|x[n]| \le 1$, and $C \le 1$. Then, for every length N,*

$$\sum_{k=0}^{N-1} \left|\mathrm{wrap}\big(\phi_Z[k] - \phi_X[k]\big)\right| \approx N\sqrt{N}\,\arcsin(C) \quad \textit{with probability } 1 - e^{-\Omega(N)}. \tag{8}$$

Under bounded i.i.d.inputs, the cumulative phase error concentrates near $N^{3/2}\arcsin(C)$. Empirical curves, Fig. 6 (b), track the predicted envelope, validating the use of TVD in the main OOD bound.

To translate design choices into TVD, coefficient growth must be controlled. Weight decay on polynomial coefficients provides that control, linking data range ρ and degree d to a small-C regime.

Proposition 4 (Polynomials learned with weight-decay)**.** *Assume the training loss* $\mathcal{L}$

$$\mathcal{L} = \frac{1}{2}\sum_{n=0}^{N-1}\left(p_d(x[n]) - t[n]\right)^2 + \lambda\sum_{r=0}^{d} a_r^2,$$

where $t[n]$ *is the target function and consider* $|x[n]| \le \rho \le 1$*. Then around the equilibrium point* $\frac{\partial \mathcal{L}}{\partial a_r} \approx 0$ *we have* $|a_r| = O(\rho^r)$*. The coefficient* C *in Proposition 2 is bounded by* $d\rho$*, yielding*

$$|\operatorname{wrap}(\phi_Z[k] - \phi_X[k])| = O(d\rho). \tag{9}$$

With $|a_r| = O(\rho^r)$, the aggregate coefficient mass satisfies $C = O(d\rho)$, hence $\mathrm{TVD}(p_d) \lesssim c_* \, d\rho$. Thus degree and input range act as a *smoothness budget*: reducing d or ρ tightens the TVD term and, via the main theorem, contracts the predicted OOD gap.

Lemma 1. *For* ReLU, TVD *is fixed and is given as* $\mathrm{TVD}(\mathrm{ReLU}) = 1$*. For polynomial* ϕ_{ij} *and inputs* $|x| \le \rho \le 1$*, by Proposition 4,* $|a_r| = O(\rho^r)$ *which implies* $\mathrm{TVD}\big(\phi_{ij}\big) \le c_* d\rho$*.*

3.4 NPG Implies Polynomial KANs (can) Generalize Better

The TVD constants from Lemma 1 can be substituted in Theorem 1 to produce a comparative OOD gap bound for ReLU-MLPs and weight-decayed polynomial KANs. at equal depth and widths. In the following, architectures are depth and width matches to isolate the effect of nonlinearity terms. The result is a decision rule in terms of degree-input scale product and inter-layer decorrelation.

Theorem 4 (Comparative TVD bound: ReLU vs.polynomial KAN)**.** *Let* $\mathcal{N}_{\mathrm{ReLU}}$ *and* $\mathcal{N}_{\mathrm{poly}}$ *be depth-*L *networks that differ only in their scalar nonlinearities. Assume Defs. 1, 2 and Assumption 1.. With* $\Delta_{\mathrm{dom}} := \sqrt{\delta\, H(E \mid Y)/2}$ *and* $S := \sum_{\ell=1}^{L} N_\ell$*, Theorem 1 and Lemma 1 give*

$$\big|\mathcal{E}_{tgt} - \mathcal{E}_{src}\big|(\mathcal{N}_{\mathrm{ReLU}}) \;\le\; \Delta_{\mathrm{dom}} \;+\; \kappa\sqrt{H(Y)(1-\gamma)}\,\sqrt{\frac{L}{128\pi}\,S},$$

$$\big|\mathcal{E}_{tgt} - \mathcal{E}_{src}\big|(\mathcal{N}_{\mathrm{poly}}) \;\le\; \Delta_{\mathrm{dom}} \;+\; \kappa\sqrt{H(Y)(1-\gamma)}\,\sqrt{\frac{L}{128\pi}\,S\,\left(c_* d\rho\right)^2}.$$

If layer bispectra are mutually orthogonal, replace L *by* 1 *in both bounds. In particular, under matched* S *the polynomial model term is a factor* $c_* d\rho$ *times the ReLU model term; it is strictly smaller whenever* $c_* d\rho < 1$*.*

Take-away. Polynomial KANs dominate when $d\rho$ is small and layer bispectra are near-orthogonal (the depth prefactor drops from L to 1); otherwise the advantage diminishes. Practically, decreasing d or ρ by a factor s reduces the model term by $\sim s$ and the predicted OOD gap by $\sim\sqrt{s}$; lack of decorrelation or large $d\rho$ erases the benefit, matching the theory's failure regime.

4 Results

The experiments use the **PACS** dataset (Li et al., 2017) and **VLCS** dataset (Torralba & Efros, 2011), the standard OOD benchmarks, each with four stylistic domains. All reported models share the same architecture, training strategies, and augmentations. Appendix A details the experiment settings.

4.1 When KANs Generalize Better

KAN vs. MLP. NPG predicts that once phase is reasonably invariant across sources, the model term is governed by nonlinearity smoothness (TVD). The data align with this. Polynomial KANs attain

Table 1: Leave-one-domain-out generalization accuracy (%). Columns report accuracy (%) on the **target domain** for models trained on the rest of the domains. Best are in **bold**, runner ups underlined.

Model	PACS						VLCS					
	Photo	Art	Cartoon	Sketch	Avg ↑ $\mathcal{E}_t$	Avg ↓ $\lvert\mathcal{E}_s-\mathcal{E}_t\rvert$	VOC	LabelMe	Caltech	SUN	Avg ↑ $\mathcal{E}_t$	Avg ↓ $\lvert\mathcal{E}_s-\mathcal{E}_t\rvert$
ReLU	39.5 ± 2.9	23.8 ± 0.5	19.7 ± 2.7	35.1 ± 0.8	29.50	26.53	40.0 ± 1.0	48.0 ± 0.6	66.0 ± 2.1	41.4 ± 0.7	48.84	12.67
LReLU	40.5 ± 1.2	23.9 ± 0.5	19.6 ± 1.7	34.8 ± 0.8	29.71	26.37	39.4 ± 1.0	48.2 ± 0.9	65.5 ± 0.7	41.7 ± 0.8	48.69	12.65
PReLU	38.7 ± 1.6	24.2 ± 0.7	20.5 ± 5.1	34.4 ± 1.3	29.41	26.63	38.9 ± 1.3	48.1 ± 0.8	65.0 ± 1.9	40.1 ± 0.8	48.00	12.60
GELU	39.2 ± 1.3	23.8 ± 0.5	21.2 ± 4.7	36.0 ± 0.9	30.04	26.77	39.4 ± 0.7	47.9 ± 0.6	66.2 ± 1.9	41.4 ± 0.8	48.71	12.67
Mish	38.9 ± 1.0	24.4 ± 0.7	19.3 ± 5.6	36.1 ± 0.9	29.68	27.15	39.4 ± 0.6	47.7 ± 0.8	65.7 ± 1.8	41.4 ± 0.9	48.54	12.76
Tanh	37.3 ± 1.7	21.8 ± 0.5	17.1 ± 3.9	28.8 ± 1.5	26.23	22.27	42.4 ± 0.4	46.6 ± 0.8	63.9 ± 1.1	42.8 ± 0.4	48.94	10.15
Softplus	40.5 ± 1.9	24.2 ± 0.3	21.9 ± 3.8	32.7 ± 0.5	29.80	23.93	42.0 ± 0.6	48.2 ± 0.3	68.4 ± 0.5	43.6 ± 0.5	**50.56**	10.99
KAN-poly	42.5 ± 0.7	24.2 ± 0.6	25.2 ± 0.8	33.2 ± 0.8	**31.29**	**21.85**	41.8 ± 0.8	48.6 ± 1.1	60.7 ± 1.2	45.0 ± 1.0	49.01	**9.87**
KAN-Bspl	35.9 ± 1.6	22.5 ± 0.6	22.2 ± 6.0	31.0 ± 1.4	27.90	29.34	39.6 ± 0.7	48.5 ± 1.2	64.8 ± 1.0	43.1 ± 1.5	49.00	10.76

the smallest OOD gaps on both datasets (PACS: 21.85 vs. ReLU 26.53; VLCS: 9.87 vs. ReLU 12.67) and the highest PACS target accuracy (31.29). Updated B-spline KAN results show dataset-dependent behavior: on VLCS–LODO, it reduces the gap to 10.76 (below ReLU's 12.67) and reaches the second-best target accuracy (49.00); on PACS–LODO, it lags (avg 27.90; gap 29.34). This pattern supports the theory's mechanism: the KAN label alone is insufficient—benefits accrue when the nonlinearity's TVD is controlled (degree/range for polynomials; spline knot density/coefficients for B-splines). The tighter standard deviations for polynomial KANs on PACS ($\approx \pm 0.6$–0.8) also match the expectation of lower phase scrambling.

Within–MLP patterns. Softplus generally improves over ReLU when multiple sources are available (LODO): it lifts VLCS target accuracy (50.56 vs. 48.84) and reduces the gap on both datasets (e.g., PACS 23.93 vs. 26.53), consistent with $\mathrm{TVD}(\mathrm{Softplus}) < 1$ on bounded inputs. In SDG, where a single source inflates the domain term, gains compress: Softplus is better on VLCS (gap 25.17 vs. 25.36) but not on PACS (32.04 vs. 29.36). Tanh often shows smaller gaps but at the cost of lower target accuracy (e.g., PACS avg 26.23), a saturation effect: shrinking phase disturbance by also shrinking label informativeness. The extended ReLU family (ReLU, LReLU, PReLU, GELU, Mish) clusters to similar target accuracies and OOD gaps across PACS/VLCS, exactly as NPG predicts when their effective $TVD \approx 1$; PReLU's slight drops are possibly due to additional in-domain fitting from the learned slope rather than a change in TVD.

Table 2 in Appendix B.1 reports *single domain generalization* results. All models admit large OOD gaps (PACS ≈29–34pp; VLCS ≈23–27pp), and the KAN's advantage largely disappears. This is consistent with NPG's decomposition: when class-conditional phase invariance is weak (large δ), the *domain term* dominates and smoothing the nonlinearity cannot, by itself, close the gap. Taken together with LODO results, this implies access to diverse sources (smaller δ) enables polynomial KANs to realize their TVD-driven advantage.

4.2 NPG - MLP Predictions

NPG explains Softplus' superior OOD performance over ReLU. Ji et al. (2024) reports better OOD generalization results when the ReLU nonlinearity is replaced by $\mathrm{Softplus}(x) = \ln(1 + \exp(x))$. NPG provides a natural explanation for their observation: the TVD of Softplus is lesser than or equal to that of ReLU, regardless of the domain of the integration (Fig. 3). Experiments on PACS dataset in leave-one-domain-out setting reported in Table 1 align with the predictions of NPG, where Softplus overfits less to the source domain and achieves better OOD accuracy than ReLU, resulting in a lower percentage drop in the accuracy. Although Tanh has a slightly improved percentage drop compared to ReLU, it is less preferable due to low absolute performance caused by the well-known optimization problems of Tanh. **NPG estimations are accurate.** Based on Barron's estimate on the model's training error for in-domain distribution with changing width, $\mathcal{E}_{source} = \mathcal{E}^* + C_{task}/\sqrt{width}$ we estimate $\mathcal{E}^*$ and C_{task} based on source accuracies. The architecture and data-dependent constants of our theory,

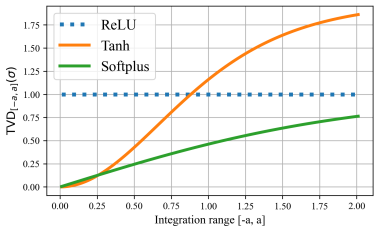
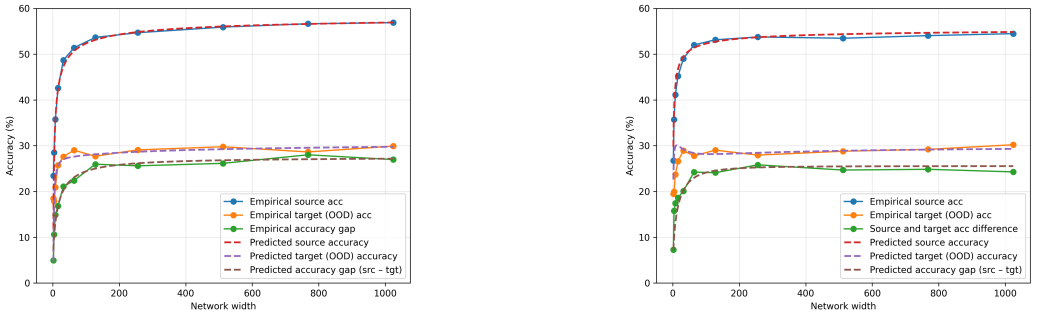


Figure 3: TVD(ReLU) is fixed, those of Tanh and Softplus' vary with the integration range. NPG predicts trained Softplus leads to better OOD performance than ReLU; while Tanh-networks depend on learned weights.

estimated on MLPs with ReLU nonlinearities. Since TVD of ReLU is fixed, it provides a clean setting for the estimation of these variables. Fig. 4a demonstrates that NPG can accurately capture OOD generalization scaling with width after finetuning the dataset-related constants.



(a) ReLU (fitting)

(b) Softplus (predicting)

Figure 4: NPG theory's predictive power is confirmed across different activation functions versus network width on the PACS dataset. **(a)** For ReLU networks, the NPG prediction (dashed curve) closely matches the empirical OOD error gap (markers) when parameters $(\gamma_{PACS}, C_{\Delta})$ are fitted. **(b)** NPG accurately predicts the performance for Softplus networks using the same parameters as the ReLU case without retuning, highlighting its robustness.

Fig.4b reports the results for Softplus nonlinearity using the same dataset and architecture, providing further support for accuracy of NPG predictions, and model-independence of the estimated constants.

4.3 Polynomial KANs: NPG Predictions Under Degree and Range Sweeps

NPG predicts that, fixing the architecture and training budget, the model term grows linearly with $\mathrm{TVD}(p_d) \propto d\rho$. Figure 5 tests this prediction on PACS (target = *Photo*). **Range (ρ) sweep.** As ρ increases, target accuracy drops while the OOD gap rises, both approximately linearly beyond $\rho \approx 1$. This matches the $\mathcal{O}(\rho)$ scaling implied by $\mathrm{TVD}(p_d)$. A small bump near $\rho = 2.25$ is within variance and does not alter the trend. **Degree (d) sweep.** Increasing d produces the same pattern: monotone decrease in target accuracy and monotone increase in gap, close to linear over $d \in [4, 15]$. This is the $\mathcal{O}(d)$ piece of the $d\rho$ law and gives a practical ceiling on degree for robust OOD behavior under a fixed range. **Take-away.** Controlling either knob reduces $\mathrm{TVD}(p_d)$ and contracts the gap; plotting accuracy/gap against the product $d\rho$ (not shown) collapses the curves, providing a single design rule: *keep $d\rho$ small*. This explains why polynomial KAN outperforms ReLU in Sec.3.3: the polynomial nonlinearity can be tuned into a low-TVD regime where the model term is smaller.

5 Related Work

Phase–centric analysis offers a complementary view of generalization by focusing on cross–frequency structure rather than weight–space surrogates. The NPG perspective connects three strands: empirical evidence that phase carries task and domain information; classical generalization analyses that are phase–agnostic; and the recently trending exploration of learnable nonlinearities (including KANs).

Phase in vision and learning. A growing body of experiments suggests that Fourier phase is label–predictive and relatively stable across domains: amplitude–phase swaps and phase–only training alter model behavior while preserving task performance, and phase–stability augmentations reduce the OOD gap (Chen et al., 2021; Xu et al., 2021; Li et al., 2024; Vaish et al., 2024; Xu et al., 2021; Hu et al., 2023; Lee et al., 2024; Xu et al., 2023; Lei et al., 2023). Signal processing long ago linked phase to perceptual identity in images and speech (Oppenheim & Lim, 1981; Gegenfurtner et al., 2003). Bispectral tools (Nikias & Mendel, 1993) provide a way to quantify nonlinear phase interactions, yet these ideas have not been integrated into generalization guarantees.

Generalization bounds (phase–agnostic). Margin/norm–based PAC analyses control risk via products of spectral/Frobenius norms (Bartlett et al., 2017), with path–norm refinements (Neyshabur

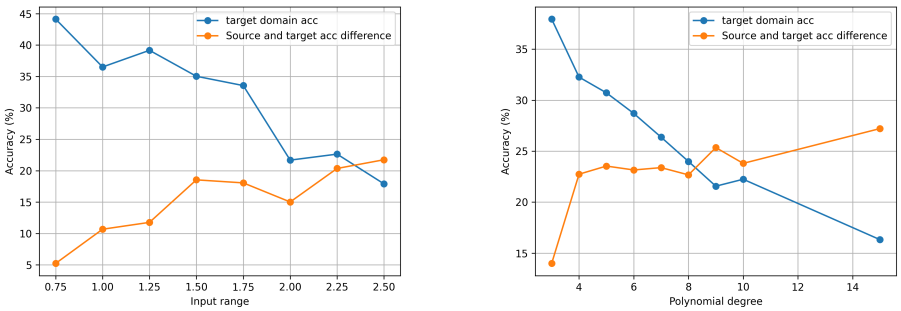


Figure 5: Polynomial KANs under input range (left) and degree (right) sweeps on PACS (*Photo* as target). Target accuracy (blue) declines and the OOD gap (orange) grows approximately *linearly* with the knob, as NPG predicts via $\mathrm{TVD}(p'_d) \propto d\rho$. The linear trend yields a single actionable rule: maintain a small $d\rho$ to preserve OOD generalization.

et al., 2015), PAC–Bayesian and sharpness–aware approaches (Dziugaite & Roy, 2017; Foret et al., 2020), and kernel/NTK or mean–field limits (Jacot et al., 2018; Mei et al., 2018). These treat the nonlinearity as fixed and emphasize linear-weight capacity, often via linearization regimes. In KANs, expressivity is carried by *learnable* nonlinearities, whose derivatives can vary with training; existing bounds thus offer limited control and remain blind to phase structure. For KANs specifically, (Zhang & Zhou, 2024) gives in–domain guarantees via RKHS–rank arguments; however phase interactions and OOD behavior are not addressed.

Learnable nonlinearities and KANs. While non–polynomial activations enable universal approximation (Cybenko, 1989), recent work revisits polynomial forms (Dubey et al., 2022; Liu et al., 2021) and replaces fixed activations with learned splines or bases (Goodfellow et al., 2013; Agostinelli et al., 2014; Tavakoli et al., 2021; Liu et al., 2024; Chen et al.; Bozorgasl & Chen, 2024; Li, 2024). Nonlinearity choice shapes capacity and optimization: depth and VC–dimension scale under ReLU but not under linear/step units (Cybenko, 1989; Bartlett et al., 2019); dynamical analyses identify activation–specific "edge of chaos" effects (Hayou et al., 2019); spectral studies show ReLU's low-to-high bias that smoother B-splines mitigate (Hong et al., 2022). Prior work, however, remains largely weight–space or empirical. In contrast, NPG uses TVD of the learned nonlinearity and GCB to obtain depth-explicit, finite-width bounds that link nonlinearity design to OOD generalization, while remaining compatible with both MLPs and KANs.

6 Conclusions

NPG links nonlinearity design to OOD behavior by tying a layer's total variation of the derivative (TVD) to phase scrambling (GCB) and, in turn, to a depth-additive source–target gap. This function-space view unifies KANs and MLPs, yields actionable guidance (choose smaller-TVD nonlinearities; encourage inter-layer decorrelation; monitor/penalize TVD), and explains when polynomial KANs can outperform ReLU-MLPs—namely when the degree–range product is small and phase is reasonably domain-stable and label-informative. The theory's predictions matched controlled studies on PACS/VLCS and within-MLP nonlinearity swaps, indicating practical utility for architecture selection and training diagnostics. Together, these results move beyond architecture-name heuristics toward measurable design rules for OOD generalization.

Scope. NPG is developed for feed-forward MLP and KAN networks. Theoretical comparisons and conducted experiments focuses on polynomial KANs and ReLU-networks. The empirical support for NPG tests the hyperparameters internal to polynomial KANs and a range of fixed nonlinearities. **Limitations.** Assumption 1 rigorously formalizes empirical observations from a wide range of applications, including medical, audio and image processing. Nonetheless, we acknowledge the assumption is not universally valid. **Future work.** Elaborate investigation of TVD regularization as an OOD generalization strategy for a wider range of architectures (e.g., CNN, Transformer), and estimation of the dataset constants, δ and γ, beyond curve fitting are promising future directions.

REFERENCES

Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Zavareh Bozorgasl and Hao Chen. Wav-kan: Wavelet kolmogorov-arnold networks. *arXiv preprint arXiv:2405.12832*, 2024.

David R Brillinger. An introduction to polyspectra. In *Selected Works of David Brillinger*, pp. 149–172. Springer, 2011.

Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 458–467, 2021.

Wei Chen, Qingfeng Xia, and JiaHui Sun. Legendre-kan: High accuracy ka network based on legendre polynomials.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Abhimanyu Dubey, Filip Radenovic, and Dhruv Mahajan. Scalable interpretability via polynomials. *Advances in neural information processing systems*, 35:36748–36761, 2022.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Karl R Gegenfurtner, Doris I Braun, and Felix A Wichmann. The importance of phase information for recognizing natural images. *Journal of Vision*, 3(9):519–519, 2003.

Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pp. 1319–1327. PMLR, 2013.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International conference on machine learning*, pp. 2672–2680. PMLR, 2019.

Qingguo Hong, Jonathan W Siegel, Qinyang Tan, and Jinchao Xu. On the activation function dependence of the spectral bias of neural networks. *arXiv preprint arXiv:2208.04924*, 2022.

Chengming Hu, Yeqian Du, Rui Wang, Hao Chen, and Congcong Zhu. Phase matching for out-of-distribution generalization. *arXiv preprint arXiv:2307.12622*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Svante Janson. *Gaussian hilbert spaces*. Number 129. Cambridge university press, 1997.

Yingrui Ji, Yao Zhu, Zhigang Li, Jiansheng Chen, Yunlong Kong, and Jingbo Chen. Advancing out-of-distribution detection through data purification and dynamic activation function design. *arXiv preprint arXiv:2403.03412*, 2024.

Ingyun Lee, Wooju Lee, and Hyun Myung. Domain generalization with vital phase augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2892–2900, 2024.

Chengxi Lei, Satwinder Singh, Feng Hou, Xiaoyun Jia, and Ruili Wang. Phaseperturbation: Speech data augmentation via phase perturbation for automatic speech recognition. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pp. 1–6, 2023.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Fengpeng Li, Kemou Li, Haiwei Wu, Jinyu Tian, and Jiantao Zhou. Dat: Improving adversarial robustness via generative amplitude mix-up in frequency domain. *arXiv preprint arXiv:2410.12307*, 2024.

Ziyao Li. Kolmogorov-arnold networks are radial basis function networks. 2024.

Li-Ping Liu, Ruiyuan Gu, and Xiaozhe Hu. Ladder polynomial neural networks. *arXiv preprint arXiv:2106.13834*, 2021.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

Chrysostomos L Nikias and Jerry M Mendel. Signal processing with higher-order spectra. *IEEE Signal processing magazine*, 10(3):10–37, 1993.

Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, 2012.

Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69 (5):529–541, 1981.

Mohammadamin Tavakoli, Forest Agostinelli, and Pierre Baldi. Splash: Learnable activation functions for improving accuracy and adversarial robustness. *Neural Networks*, 140:1–12, 2021.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Puru Vaish, Shunxin Wang, and Nicola Strisciuglio. Fourier-basis functions to bridge augmentation gap: Rethinking frequency augmentation in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17763–17772, 2024.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Qinwei Xu, Ruipeng Zhang, Ziqing Fan, Yanfeng Wang, Yi-Yan Wu, and Ya Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023.

Xianyang Zhang and Huijuan Zhou. Generalization bounds and model complexity for kolmogorov-arnold networks. *arXiv preprint arXiv:2410.08026*, 2024.

A Experiment Settings

All experiments use the **PACS** dataset (Li et al., 2017) and **VLCS** dataset (Torralba & Efros, 2011), the standard OOD benchmarks, each with four stylistic domains. All networks, MLP and KAN, have two hidden layers with 512 and 256 neurons, and trained with SGD optimizer ($\texttt{lr} = 3 \times 10^{-3}$ and $\texttt{wd} = 6 \times 10^{-4}$) for 100 epochs using CosineAnnealing learning rate scheduler. For KAN variants of both architectures we utilized B-spline Liu et al. (2024) and polynomial (Legendre) basis Chen et al.. Models are trained with minimal augmentations: padding (4 px), and random horizontal flips (probability 0.5). Each reported result is averaged over 10 runs per target/source domain.

B Additional Empirical Results

B.1 Single Domain Generalization

Table 2 per-domain target accuracy and OOD gaps for all activations (ReLU, LReLU, PReLU, GELU, Mish, Tanh, Softplus) and KAN variants on PACS/VLCS. Protocols, budgets, and tuning are identical to LODO experiments.

Table 2: Single-domain generalization accuracy (%). Columns report accuracy (%) per **source domain** when models are tested on the rest of the domains. Best are in **bold**, runner ups underlined.

Model	PACS						VLCS					
	Photo	Art	Cartoon	Sketch	Avg ↑ $\mathcal{E}_t$	Avg ↓ $\lvert\mathcal{E}_s-\mathcal{E}_t\rvert$	VOC	LabelMe	Caltech	SUN	Avg ↑ $\mathcal{E}_t$	Avg ↓ $\lvert\mathcal{E}_s-\mathcal{E}_t\rvert$
ReLU	23.3 ± 0.8	29.2 ± 0.6	15.4 ± 1.2	28.7 ± 0.8	24.15	<u>29.36</u>	52.6 ± 1.0	41.2 ± 1.2	26.7 ± 1.5	41.8 ± 2.5	40.59	25.36
LReLU	22.8 ± 1.9	29.6 ± 0.8	14.4 ± 1.1	28.6 ± 1.3	23.85	30.53	52.2 ± 1.4	41.2 ± 1.4	26.7 ± 0.3	42.1 ± 2.6	40.54	27.05
PReLU	22.9 ± 1.3	28.1 ± 1.3	15.7 ± 0.4	28.8 ± 0.7	23.87	33.74	51.6 ± 2.0	41.5 ± 1.5	25.7 ± 0.7	42.5 ± 3.5	40.31	27.43
GELU	24.1 ± 0.8	29.2 ± 0.9	15.5 ± 0.5	27.9 ± 1.2	<u>24.17</u>	32.66	52.0 ± 1.6	40.5 ± 1.4	25.9 ± 0.4	43.0 ± 1.8	40.36	27.27
Mish	24.5 ± 1.2	28.8 ± 0.7	15.4 ± 0.5	28.8 ± 1.1	**24.36**	33.52	52.1 ± 1.3	41.3 ± 1.5	25.5 ± 0.8	43.5 ± 1.2	40.61	26.82
Tanh	22.1 ± 0.4	29.8 ± 0.5	16.1 ± 0.4	26.4 ± 0.5	23.59	**28.41**	52.2 ± 0.7	44.5 ± 0.4	28.6 ± 0.9	42.7 ± 0.7	**41.99**	**23.46**
Softplus	19.3 ± 1.3	29.3 ± 0.6	15.2 ± 0.3	30.1 ± 1.1	23.46	32.04	53.0 ± 0.5	44.6 ± 1.1	26.2 ± 1.0	43.0 ± 2.6	<u>41.70</u>	25.17
KAN-poly	17.6 ± 0.8	30.2 ± 0.6	18.0 ± 0.7	27.9 ± 0.7	23.41	32.39	50.7 ± 0.9	40.8 ± 2.8	28.7 ± 1.5	43.6 ± 3.2	40.96	<u>24.19</u>
KAN-Bspl	23.0 ± 2.0	27.0 ± 1.3	17.0 ± 0.5	27.9 ± 0.8	23.74	35.65	52.2 ± 1.9	41.0 ± 2.9	24.4 ± 1.1	45.9 ± 2.9	40.88	26.32

B.2 Tightness of the Bounds and Estimates

Figure 6 benchmarks the theory against synthetic data and confirms that the closed-form bounds in Propositions 1&3 are numerically *tight* rather than merely asymptotic. Signals of length 256 were drawn from three common distributions: white Gaussian, uniform, and $1/f$ ("pink") noise. These are then fed through either a single ReLU or a degree-20 polynomial nonlinearity with coefficients sampled iid. For each trial the *exact* phase error $\sum_k |\mathrm{wrap}(\phi_Z[k] - \phi_X[k])|$ is measured and compared to the analytical upper bounds predicted by the propositions.

Key observations. *i)* ReLU (Fig.6a). Across all input types the empirical phase error falls between the upper and lower bounds, validating that the negative-part energy governs ReLU's spectral distortion. *ii)* Polynomial (Fig.6b. As predicted by Proposition 3, the phase error scales like $N^{3/2}\arcsin(C)$. Varying the regularization strength changes C and shifts the curve almost exactly along the analytical slope, showing that the coefficient norm is a reliable control knob.

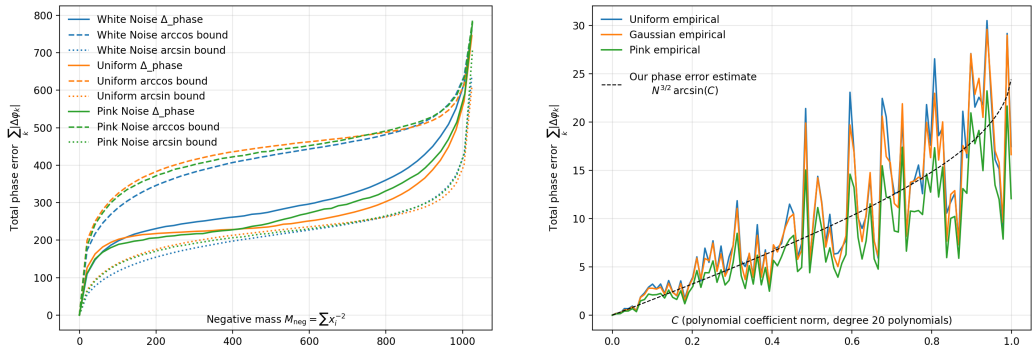
Take-away. The close match demonstrates that (i) the constants in the theory are not over-conservative, and (ii) the total-variation–based metric TVD is a practical surrogate for real phase behavior. This tightness justifies using the bounds later to interpret KAN generalization performance.

C Detailed Proofs

C.1 Theorem 2

Proof. Fix ω_1, ω_2 and set $\omega_3 = \omega_1 + \omega_2$. Write $U = X(\omega_1)X(\omega_2)$ and $\tilde{U} = U - \mathbb{E}[U]$.

Equalization is a contraction and preserves the normalized cbs magnitude. By construction Y is replaced by $\tilde{Y} = H\,Y$ with $\mathbb{E}|\tilde{Y}(\omega)|^2 \equiv 1$ and $\|H\|_\infty \le 1$. Linearity yields $(h * y)^{(2)} = h * y^{(2)}$.



(a) (b)

Figure 6: (a) For ReLU (Proposition 1), input statistics determine the phase error; (b) for polynomial nonlinearities (Proposition 3), **polynomial coefficients-which are controllable-are the deciding factors**. The empirical verification as *tight* bounds/estimates on random inputs (uniform, Gaussian, or pink noise) and for polynomials, randomly sampled coefficients (degree $= 20$).

Parseval identity gives

$$\sum_{i=1}^{M} \mathbb{E}\,|(h * y^{(2)})_i|^2 \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\omega)|^2\, \mathbb{E}\,|Y^{(2)}(\omega)|^2\, d\omega \;\le\; \sum_{i=1}^{M} \mathbb{E}\,|y_i^{(2)}|^2.$$

The normalized cbs magnitude is invariant since both numerator and denominator acquire the common factor $H(\omega_3)$, (Brillinger, 2011). Hence it is enough to work after equalization, and we drop tildes on Y.

Only the second chaos contributes. Each y_i is a measurable function of a Gaussian vector and admits a Wiener chaos expansion. Orthogonality of distinct chaoses implies

$$\mathbb{E}\big[\tilde{U}\, Y(\omega_3)^*\big] \;=\; \mathbb{E}\big[\tilde{U}\, \langle y^{(2)}, f_{\omega_3}\rangle^*\big],$$

where f_{ω_3} is the unit norm DFT atom and $y^{(2)}$ is the second chaos projection of y, see for example (Nourdin & Peccati, 2012; Janson, 1997).

Two Cauchy Schwarz estimates. Since $\|f_{\omega_3}\|_2 = 1$,

$$\big|\mathbb{E}[\tilde{U}\, Y^*]\big| = \Big|\sum_{i=1}^{M} f_{\omega_3}[i]^*\, \mathbb{E}[\tilde{U}\, y_i^{(2)}]\Big| \;\le\; \Big(\sum_{i=1}^{M} \big|\mathbb{E}[\tilde{U}\, y_i^{(2)}]\big|^2\Big)^{1/2}.$$

By Cauchy Schwarz in probability,

$$\big|\mathbb{E}[\tilde{U}\, y_i^{(2)}]\big|^2 \;\le\; \mathbb{E}|\tilde{U}|^2\, \mathbb{E}|y_i^{(2)}|^2.$$

Therefore,

$$|b_{xxy}(\omega_1,\omega_2)|^2 = \frac{\big|\mathbb{E}[\tilde{U}\, Y^*]\big|^2}{\mathbb{E}|\tilde{U}|^2\, \mathbb{E}|Y|^2} \;\le\; \sum_{i=1}^{M} \mathbb{E}|y_i^{(2)}|^2,$$

and here $\mathbb{E}|Y|^2 \equiv 1$ by equalization. **Controlling the second chaos by** TVD**.** Let $Z \sim \mathcal{N}(0,1)$ and let $\mathrm{He}_2(z) = z^2 - 1$. For $g \in \mathcal{G}$, the Hermite coefficient at order two is

$$a_2 = \frac{\mathbb{E}[g(Z)\mathrm{He}_2(Z)]}{2} = \frac{\mathbb{E}[g''(Z)]}{2},$$

by Stein integration by parts (Nourdin & Peccati, 2012). Since $\mathrm{Var}(\mathrm{He}_2(Z)) = 2$, the second chaos variance is

$$\mathrm{Var}\big(g^{(2)}(Z)\big) \;=\; a_2^2\, \mathrm{Var}(\mathrm{He}_2(Z)) \;=\; \frac{(\mathbb{E}[g''(Z)])^2}{2}.$$

Because $\sup_u \varphi(u) = (2\pi)^{-1/2}$ for the standard Gaussian density φ,

$$|\mathbb{E}[g''(Z)]| = \Big| \int_{\mathbb{R}} \varphi(u)\, dg''(u) \Big| \;\le\; \|\varphi\|_\infty \int_{\mathbb{R}} |dg''(u)| \;=\; (2\pi)^{-1/2}\, \mathrm{TVD}(g).$$

Hence

$$\mathrm{Var}\big(g^{(2)}(Z)\big) \;\le\; \frac{1}{4\pi}\, \mathrm{TVD}(g)^2.$$

Apply this with

$$\text{MLP:}\quad g(u) = \zeta_i(u) = \sigma(\|A_{i:}\|_2\, u + b_i) \quad\Rightarrow\quad \mathrm{TVD}(g) = \|A_{i:}\|_2 \mathrm{TVD}(\sigma),$$

and with

$$\text{KAN:}\quad g(u) = \phi_{ij}(u), \quad y_i = \sum_j \phi_{ij}(x_j), \quad \mathrm{Var}(y_i^{(2)}) = \sum_j \mathrm{Var}(\phi_{ij}(Z)^{(2)}),$$

using independence across j. Summing over i gives

$$\sum_{i=1}^{M} \mathbb{E}|y_i^{(2)}|^2 \;\le\; \frac{1}{4\pi} \sum_{i=1}^{M} \mathrm{TVD}(\zeta_i)^2. \tag{10}$$

Integrate on the principal simplex. The area of PS is $\pi^2/4$, so

$$\iint_{\mathsf{PS}} \frac{d\omega_1\, d\omega_2}{(2\pi)^2} = \frac{1}{16}.$$

Combining the pointwise bound from Step 3 with equation 10,

$$\iint_{\mathsf{PS}} |b_{xxy}|^2\, \frac{d\omega_1\, d\omega_2}{(2\pi)^2} \;\le\; \frac{1}{16} \cdot \frac{1}{4\pi} \sum_{i=1}^{M} \mathrm{TVD}(\zeta_i)^2.$$

Finally, by the definition of S,

$$S \;\le\; \frac{1}{2} \cdot \frac{1}{16} \cdot \frac{1}{4\pi} \sum_{i=1}^{M} \mathrm{TVD}(\zeta_i)^2 \;=\; \frac{1}{128\,\pi} \sum_{i=1}^{M} \mathrm{TVD}(\zeta_i)^2.$$

The network aggregation follows directly from taking the cross-terms involving GCB into account.

□ □

C.2 Proposition 1

Proposition 1 (Extended to amplitude effect) Let the real-valued length–N signal $x[n]$ be decomposed as $x[n] = x^+[n] + x^-[n]$ with $x^+[n] = \max\{x[n], 0\}$ and $x^-[n] = \min\{x[n], 0\} \le 0$. Define the ReLU output $z[n] = \mathrm{ReLU}(x[n]) = x^+[n]$, and let

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i2\pi kn/N}, \qquad Z[k] = \sum_{n=0}^{N-1} z[n] e^{-i2\pi kn/N}$$

be the discrete Fourier transforms with phases $\phi_X[k] = \arg X[k]$ and $\phi_Z[k] = \arg Z[k]$. Then

(i) *Amplitude distortion.*

$$\sum_{k=0}^{N-1} \big(|Z[k]| - |X[k]|\big)^2 \;\le\; N \sum_{n=0}^{N-1} \big(x^-[n]\big)^2 \;=\; O\Big(\sum x^-[n]^2\Big).$$

(ii) *Phase distortion.* Writing $X^-[k] := \mathcal{F}\{x^-\}[k]$ one has

$$|\phi_Z[k] - \phi_X[k]| \;\le\; \arccos\Big(1 - \tfrac{|X^-[k]|}{|X[k]|}\Big),$$

and for the small-ratio regime $|X^-[k]|/|X[k]| \ll 1$ the bound linearizes to $|\phi_Z[k] - \phi_X[k]| \approx \arcsin\big(|X^-[k]|/|X[k]|\big)$.

Proof. **(i) Amplitude part.** For any complex numbers a, b, $\big||a| - |b|\big| \le |a - b|$ (triangle inequality on the complex plane). Hence

$$\big(|Z[k]| - |X[k]|\big)^2 \;\le\; |Z[k] - X[k]|^2.$$

Because $Z = X - X^-$ in the frequency domain, $Z[k] - X[k] = -X^-[k]$. Summing and invoking Parseval's identity,

$$\sum_k |Z[k] - X[k]|^2 = \sum_k |X^-[k]|^2 = N \sum_n |x^-[n]|^2.$$

Combining the two displays proves the amplitude bound.

(ii) Phase part. Fix k and denote $A := Z[k]$, $B := X^-[k]$ so that $X[k] = A + B$. Let $\theta := |\phi_Z[k] - \phi_X[k]|$ be the angle between vectors A and $A + B$ in the complex plane. Elementary geometry gives the exact relation $\sin\theta = \frac{|B|\sin\beta}{|A+B|}$, where β is the (unknown) angle between A and B. The worst-case (largest) θ arises when $\beta = \pi/2$, whence $\sin\theta \le |B|/|A + B| = |X^-[k]|/|X[k]|$. Because $\theta \in [0, \pi/2]$, $\theta \le \arcsin\big(|X^-[k]|/|X[k]|\big) \le \arccos\big(1 - |X^-[k]|/|X[k]|\big)$, where the final inequality uses $\arcsin u \le \arccos(1 - u)$ for $u \in [0, 1]$. For small ratios $u = |X^-|/|X| \ll 1$ we have $\arcsin u \approx u$ and $\arccos(1 - u) \approx \sqrt{2u}$; keeping the tighter linear approximation yields the advertised $\theta \approx \arcsin(|X^-[k]|/|X[k]|)$. □

Interpretation. The proposition (i) shows that ReLU's *amplitude* distortion is governed solely by the energy of the negative input part. (ii) shows that the *phase* distortion per frequency bin is controlled by the *relative* negative-energy contribution; for spectra where the negative component is small, the phase shift scales like $\arcsin$ with that ratio, while in the worst case it is capped by $\arccos$.

C.3 Proposition 2

Proposition 2 (Extended to include amplitude) Let $p_d(u) = \sum_{r=0}^{d} a_r u^r$ with real coefficients and write

$$z[n] = p_d\big(x[n]\big), \qquad d[n] := z[n] - x[n] = \sum_{r \ne 1} a_r x^r[n] + (a_1 - 1)x[n].$$

Define the coefficient slack $C := |a_0| + |a_1 - 1| + \sum_{r=2}^{d} |a_r|$. Assume the input is range-normalized: $|x[n]| \le 1 \;\; \forall n$. Then

(i) *Energy of the deviation.*

$$\sum_{n=0}^{N-1} d^2[n] \;\le\; N\,C^2.$$

(ii) *Amplitude distortion.* For the DFTs $X[k], Z[k]$ of x, z

$$\sum_{k=0}^{N-1} \big(|Z[k]| - |X[k]|\big)^2 \;\le\; N^2 C^2.$$

(iii) *Aggregate phase distortion.* With $\mathrm{wrap}(\cdot)$ denoting principal-value phase wrapping,

$$\sum_{k=0}^{N-1} \big|\mathrm{wrap}\big(\phi_Z[k] - \phi_X[k]\big)\big| \;\le\; N^2 C.$$

Proof. **Bounding the point-wise deviation.** Because $|x[n]| \le 1$,

$$|d[n]| \;\le\; |a_0| + |a_1 - 1|\,|x[n]| + \sum_{r=2}^{d} |a_r|\,|x[n]|^r \;\le\; C \quad \to \quad d^2[n] \le C^2.$$

Summing over n proves (i).

Amplitude distortion. Let $D[k]$ be the DFT of $d[n]$. Triangle inequality on the complex plane gives $||Z[k]| - |X[k]|| \le |D[k]|$. Thus

$$\sum_k \big(|Z[k]| - |X[k]|\big)^2 \;\le\; \sum_k |D[k]|^2 \overset{\text{Parseval}}{=} N \sum_n d^2[n] \;\overset{(i)}{\le}\; N^2 C^2,$$

which is statement (ii).

Phase distortion. For each frequency bin write $X[k] = A_k$ and $D[k] = B_k$, so $Z[k] = A_k + B_k$. The wrapped phase difference is $\theta_k := \arg\big(1 + B_k/A_k\big)$. Using $|\arg(1+u)| \le |u|$ for any complex u,

$$|\theta_k| \;\le\; \frac{|B_k|}{|A_k|}.$$

Because $|x[n]| \le 1$, the trivial bound $|B_k| \le \sum_n |d[n]| \le NC$ holds, while $|A_k| \ge 0$. Discarding the denominator gives the crude but universal estimate $|\theta_k| \le NC$. Adding over k yields $\sum_k |\theta_k| \le N^2 C$, which is claim (iii). □

Interpretation. Proposition 2 shows that if the coefficient deviation from the identity map is small ($C \ll 1$) and inputs are range-bounded, both the amplitude and phase distortions introduced by a degree-d polynomial activation remain tightly controlled, scaling at most quadratically and linearly in C, respectively.

C.4 Proposition 3

Proposition 3 (Phase envelope for bounded iid inputs) Let $x[n]$ ($0 \le n < N$) be i.i.d. real random variables with $\mathbb{E}[x[n]] = 0$, $\mathbb{E}[x[n]^2] = 1$, $|x[n]| \le 1$. Let $z[n] = p_d\big(x[n]\big)$ be the output of a degree-d polynomial non-linearity whose coefficient slack $C := |a_0| + |a_1 - 1| + \cdots + |a_d|$ obeys $0 < C \le 1$. Denote by $X[k]$ and $Z[k]$ the (unnormalized) DFTs of x and z and write $\phi_X[k] = \arg X[k]$, $\phi_Z[k] = \arg Z[k]$. Then, for every $N \ge 1$,

$$\sum_{k=0}^{N-1} \big|\mathrm{wrap}\big(\phi_Z[k] - \phi_X[k]\big)\big| = O\big(N^{3/2} \arcsin C\big) \quad \text{with probability } 1 - e^{-\Omega(N)}. \tag{11}$$

Proof. Throughout, $c_1, c_2, \ldots$ denote positive numerical constants independent of N.

Let $d[n] := z[n] - x[n]$. From Proposition 2 we have $|d[n]| \le C$. Define $D[k] = \sum_n d[n] e^{-i2\pi kn/N}$. Hoeffding's inequality for the sum of N bounded, independent C-subgaussian variables yields, for any $t > 0$,

$$\Pr\big(|D[k]| \ge t\big) \;\le\; 2\exp\big(-t^2/(2NC^2)\big). \tag{12}$$

Setting $t = c_1 C\sqrt{N \log N}$ and union-bounding over $k = 0, \ldots, N-1$ gives the event

$$|D[k]| \;\le\; c_1 C\sqrt{N \log N} \quad \forall k \tag{13}$$

with probability at least $1 - N^{-3}$.

Write $X[k] = U_k + iV_k$ with $U_k := \sum_n x[n]\cos(2\pi kn/N)$ and $V_k := -\sum_n x[n]\sin(2\pi kn/N)$. Both U_k, V_k are sums of independent ± 1-bounded, zero-mean variables of variance $\approx N/2$. Applying Hoeffding again,

$$\Pr\big(|X[k]| \le \tfrac{1}{2}\sqrt{N}\big) \;\le\; e^{-c_2 N}. \tag{14}$$

A union bound shows that, with probability at least $1 - e^{-c_2 N}$,

$$|X[k]| \;\ge\; \tfrac{1}{2}\sqrt{N} \quad \forall k.$$

For each k write $X[k] = A_k$ and $Z[k] = A_k + B_k$ with $B_k := D[k]$. Let $\theta_k := \big|\mathrm{wrap}\big(\phi_Z[k] - \phi_X[k]\big)\big| = \big|\arg(1 + B_k/A_k)\big|$. Because $|\arg(1+u)| \le |u|$, $\theta_k \;\le\; \frac{|B_k|}{|A_k|}$. On the high-probability event Eq.13–Eq.14,

$$\theta_k \;\le\; \frac{c_1 C\sqrt{N \log N}}{\frac{1}{2}\sqrt{N}} \;=\; 2c_1\sqrt{\log N}\, C.$$

Since $C \le 1$, $\arcsin C \asymp C$ and $\theta_k = O\big(\sqrt{\log N}\, \arcsin C\big)$.

Add the bound for all k:

$$\sum_{k=0}^{N-1} \theta_k \;\le\; 2c_1 N \sqrt{\log N}\, C \;=\; O\big(N^{3/2}\, \arcsin C\big),$$

because $\log N \le N$ for $N \ge 2$. The same union bound that delivered Eq.13–Eq.14 shows the complement event has probability $e^{-\Omega(N)}$, completing the proof. □

C.5 Proposition 4

Proposition (Polynomials learned with weight-decay) Let $p_d(u) = \sum_{r=0}^{d} a_r u^r$ and consider the regularized training objective

$$\mathcal{L}_a(x) \;=\; \tfrac{1}{2} \sum_{n=0}^{N-1} \big(p_d(x[n]) - t[n]\big)^2 \;+\; \lambda \sum_{r=0}^{d} a_r^2,$$

where the inputs satisfy $|x[n]| \le \rho \le 1$. Write $z[n] = p_d(x[n])$ and $X[k], Z[k]$ for the DFTs of x, z with phases $\phi_X[k], \phi_Z[k]$. At a stationary point $\nabla_a \mathcal{L} \approx 0$ we have

$$|a_r| \;=\; O(\rho^r), \qquad r = 0, \dots, d,$$

so that the coefficient slack $C := |a_0| + |a_1 - 1| + \sum_{r=2}^{d} |a_r|$ from Proposition 2 obeys

$$C \;\le\; d\rho.$$

Consequently

$$\sum_k \big(|Z[k]| - |X[k]|\big)^2 = O\big(d^2\rho^2\big), \qquad |\phi_Z[k] - \phi_X[k]| = O\big(d\rho\big) \quad \forall k.$$

Proof. The gradient component w.r.t. a_r is

$$\partial_{a_r} \mathcal{L} = \sum_n (p_d(x[n]) - t[n])\, x[n]^r + 2\lambda a_r.$$

At a local minimum this derivative is (close to) zero, giving

$$|a_r| \;=\; \frac{1}{2\lambda} \Big| \sum_n \big(t[n] - p_d(x[n])\big)\, x[n]^r \Big|. \tag{15}$$

Denote the *training error* by $\epsilon := \max_n |t[n] - p_d(x[n])|$. Because $|x[n]| \le \rho$, $|x[n]|^r \le \rho^r$. Hence the right-hand side of equation 15 is bounded by $(N\epsilon\,\rho^r)/(2\lambda) = O(\rho^r)$, since N and ϵ/λ are constants during training. Thus $|a_r| = O(\rho^r)$.

Split the sum:

$$C = |a_0| + |a_1 - 1| + \sum_{r=2}^{d} |a_r|.$$

For $r \ge 2$ we have $\rho^r \le \rho^2 \le \rho$, so each term $|a_r| = O(\rho)$. There are at most d such terms, whence $C \le d\rho + O(\rho) = O(d\rho)$; for concreteness we write $C \le d\rho$.

Proposition 2 gave the generic bounds

$$\sum_k \big(|Z[k]| - |X[k]|\big)^2 \le N^2 C^2, \qquad \big|\phi_Z[k] - \phi_X[k]\big| \le NC.$$

Most empirical works scale the DFT by $1/N$, in which case those bounds lose an N factor each. Adopting that normalization (so our spectra match the usual "power spectrum" convention) we obtain

$$\sum_k \big(|Z[k]| - |X[k]|\big)^2 \;\le\; C^2 N^0 = O(d^2\rho^2), \quad |\phi_Z[k] - \phi_X[k]| \;\le\; C = O(d\rho),$$

as claimed. □

Interpretation. L^2 regularization steers the learned polynomial towards the identity map by forcing each coefficient to decay geometrically with exponent ρ. The resulting slack parameter C therefore shrinks *linearly* with both the input scale ρ and the degree d, and the amplitude and phase distortions inherit this favorable dependence.

C.6 LEMMA 1

Lemma 1 (Total–variation constants for the two activations) For ReLU, TVD is fixed and is given as $\mathrm{TVD}(\mathrm{ReLU}) = 1$. For polynomial ϕ_{ij} and inputs $|x| \le \rho \le 1$, by Proposition 4, $|a_r| = O(\rho^r)$ which implies $\mathrm{TVD}\big(\phi_{ij}\big) \le c_* d\rho$.

Proof. **(a) ReLU.** The first derivative of ReLU is the Heaviside step function, and the second derivative is the Dirac delta function. $\int_a^b \delta(x)dx = 1$ over any interval $[a, b]$ that includes 0.

(b) Polynomial. Let $\sigma_{\mathrm{poly}}(u) = \sum_{r=0}^{d} a_r u^r$ with $|a_r| \le K\rho^r$ for a constant $K = O(1)$ from Proposition 4. Then

$$\sigma'_{\mathrm{poly}}(u) = \sum_{r=1}^{d} r a_r u^{r-1}, \qquad \sigma''_{\mathrm{poly}}(u) = \sum_{r=2}^{d} r(r-1) a_r u^{r-2}.$$

Because the inputs satisfy $|u| \le \rho \le 1$, one may bound $|u|^{r-2} \le \rho^{r-2}$. Hence

$$|\sigma''_{\mathrm{poly}}(u)| \;\le\; \sum_{r=2}^{d} r(r-1)|a_r|\rho^{r-2} \;\le\; K\sum_{r=2}^{d} r(r-1)\rho^{r-2}\rho^{r} \;=\; K\sum_{r=2}^{d} r(r-1)\rho^{2r-2}.$$

Because $\rho \le 1$, $\rho^{2r-2} \le \rho$ for all $r \ge 2$, so the right-hand side is bounded by $K\rho\sum_{r=2}^{d} r(r-1) \le K\rho d(d-1) = O(d^2\rho)$. Since $\mathrm{TVD}(\sigma) = \int_x |\sigma''(u)|\, du$ and the input support has length at most 2ρ, we obtain

$$\mathrm{TVD}\big(\sigma_{\mathrm{poly}}\big) \le 2\rho \max_{|u|\le\rho} |\sigma''_{\mathrm{poly}}(u)| \le c_* d\rho.$$

□

D LLM USAGE

In this paper, LLMs are used for organizing the tables, polishing the writing, and keeping track of notational consistency.