Scaling Laws and Pathologies of Single-Layer PINNs: Network Width and PDE Nonlinearity

Faris Chaudhry
Department of Computer Science
Imperial College London
`faris.chaudhry22@imperial.ac.uk`

Abstract

We establish empirical scaling laws for Single-Layer Physics-Informed Neural Networks on canonical nonlinear PDEs. We identify a dual optimization failure: (i) a baseline pathology, where the solution error fails to decrease with network width, even at fixed nonlinearity, falling short of theoretical approximation bounds, and (ii) a compounding pathology, where this failure is exacerbated by nonlinearity. We provide quantitative evidence that a simple separable power law is insufficient, and that the scaling behavior is governed by a more complex, non-separable relationship. This failure is consistent with the concept of spectral bias, where networks struggle to learn the high-frequency solution components that intensify with nonlinearity. We show that optimization, not approximation capacity, is the primary bottleneck, and propose a methodology to empirically measure these complex scaling effects.

1 Introduction

Physics-Informed Neural Networks (PINNs) [1] offer a compelling mesh-free paradigm for numerically solving partial differential equations (PDEs) by embedding the governing physics directly into the neural network's loss function. Inspired by the impact of scaling law studies in foundational language and vision models, [2] which have provided an empirical understanding of model performance, a similar quantitative framework for PINNs is critically underdeveloped. The relationship between model capacity, problem complexity, and solution accuracy remains largely unquantified.

At a theoretical level, the Universal Approximation Theorem (UAT) [3, 4] guarantees that even a Single-Layer Network (SLN) possesses the necessary expressive power to approximate continuous functions. Furthermore, theoretical bounds for function classes like Barron spaces [5] suggest error should decrease with network width N as $\mathcal{O}(N^{-1/2})$, which will correspond to a theoretical scaling exponent of $\alpha = 0.5$ in our experiments. These theorems guarantee existence, but not that gradient-based optimization will find such an approximation in practice. This creates a gap between theoretical expressivity and practical performance, where the primary bottleneck might shift from approximation capacity to optimization challenges in the complex, non-convex loss landscapes of PINNs. [6] A key mechanism behind this optimization challenge is spectral bias, the tendency for gradient-based optimization to fit low-frequency components of a function far more rapidly and robustly than high-frequency ones. [7, 8] As the nonlinearity and complexity of the PDE solution increase, requiring the network to learn these high-frequency features, spectral bias can lead to a failure of the training process.

This work investigates this theory-practice gap. We frame our investigation around the separable scaling law error $\approx A \cdot N^{-\alpha} \cdot \kappa^{\gamma}$, where κ will be our problem-dependent complexity parameter, α is the width scaling exponent, and γ is the nonlinearity scaling exponent. This is motivated by the theoretical laws derived from approximation theory. We use SLNs to isolate the effect of width and directly test the limits of the UAT's practical relevance. The goal is to test a two-part hypothesis: (i)

that practical SL-PINN training exhibits a baseline optimization pathology where $\alpha \neq 0.5$, and (ii) that this problem is compounded by a second pathology where the separable power law breaks down. We hypothesize that the scaling exponent α is itself a function of the nonlinearity κ, a non-separable relationship that points to a more fundamental optimization challenge. Our primary contribution is the systematic measurement and analysis of these exponents across a diverse suite of canonical PDEs representing different classes of nonlinear phenomena and PDE types: dispersive (Korteweg-de Vries), transcendental/hyperbolic (Sine-Gordon), and reactive/parabolic (Allen-Cahn).

2   Related Work

The challenge of training PINNs for complex physical systems is a topic of intense research. Early work by Krishnapriyan et al. [6] provided a foundational characterization of PINN failure modes, arguing that, for stiff or complex PDEs, the bottleneck is not network expressivity but rather a failure of optimization. Our work aims to quantify this failure with a precise scaling law. Other work has explored architectural solutions to handle problem complexity, such as domain decomposition methods like Finite-Basis PINNs, which scale PINNs to larger domains by avoiding single, monolithic networks. [9] Our study, in contrast, investigates the fundamental scaling properties of the monolithic PINN architecture itself.

This focus on network width reveals a distinction from mainstream deep learning. In many applications, wider networks are understood to facilitate optimization, for example by enabling stable large-batch training. [10] Our finding of "pathological scaling," where wider is worse, runs directly counter to this heuristic. This suggests that the PINN optimization landscape has fundamentally different properties, a notion supported by recent theoretical work. Wang et al. [7] identified gradient flow pathologies and spectral bias as mechanisms that hinder PINN training, especially for high-frequency or multi-scale problems. Further, while the Neural Tangent Kernel (NTK) [11] provides insight into infinitely-wide networks, Bonfanti et al. [12] have shown that standard NTK theory is misleading for nonlinear PDEs.

This optimization challenge is particularly acute for fluid dynamics simulations governed by the Navier-Stokes equations. Several studies have used the Reynolds number (Re), which quantifies the ratio of inertial to viscous forces, as a natural "hardness parameter." For instance, Reyes et al. [13] observed that PINN performance degraded significantly as Re increased in simulations of flow past a cylinder, identifying this as a key limitation. Similarly, Arzani et al. [14] noted that PINNs struggled to capture turbulent flow features at high Re. These works highlight a specific instance of the phenomenon we study, but they do not formalize it as a general scaling law. Our hardness parameter κ can be seen as a generalization of parameters like Re to other classes of nonlinear PDEs.

3   Methodology

The total loss, a weighted sum of the mean squared residuals for the PDE, boundary conditions (BCs), and initial conditions (ICs), is minimized:

$$\mathcal{L}_{\text{total}} = w_{\text{pde}}\mathcal{L}_{\text{pde}} + w_{\text{bc}}\mathcal{L}_{\text{bc}} + w_{\text{ic}}\mathcal{L}_{\text{ic}}. \tag{1}$$

For all models, we use equal weights ($w = 1$) and train for 25,000 epochs using the Adam optimizer with a learning rate of 10^{-3}. We use Adam as it is broadly representative of first-order methods commonly used in the wider deep learning community, allowing us to focus specifically on the architectural effects of width.

3.1   PDE Suite and Hardness Parameter

We analyze a suite of canonical PDEs with scalar outputs in one spatial dimension (x) and an optional time dimension (t). For each nonlinear PDE, we define a tunable hardness parameter κ that controls the strength of the nonlinear effects. [1]

- **Poisson Eq.:** $-u_{xx} = \sin(\pi x)$ on $x \in [0, 1]$ with Dirichlet BCs. This serves to validate our framework against known theoretical scaling rates for PINNs. Here, there is no dependence on κ.

---

[1] https://github.com/farischaudhry/pinn-width-vs-nonlinearity

- **KdV Eq. (Dispersive):** $u_t + \kappa u u_x + u_{xxx} = 0$. The hardness $\kappa = A$ is the soliton amplitude, which dictates its speed and sharpness.

- **Sine-Gordon Eq. (Hyperbolic/Transcendental):** $u_{tt} - u_{xx} + \kappa \sin(u) = 0$. The hardness κ scales the strength of the nonlinear potential term. The IC includes a constraint on $u_t(0, x)$.

- **Allen-Cahn Eq. (Reactive/Parabolic):** $u_t - D u_{xx} + (u^3 - u) = 0$. We define the hardness $\kappa = 1/D$, where smaller diffusion D leads to sharper, harder-to-resolve interfaces.

In all three nonlinear cases, this hardness parameter κ is designed to control the strength of nonlinear effects, which manifests as an increase in the high-frequency components of the ground truth solution (e.g., sharper solitons or interfaces), thereby creating a direct challenge to the network's inherent spectral bias.

3.2 Experimental Sweep and Evaluation

For each PDE, we perform a systematic sweep over:

- **Network Widths** $N \in \{16, 32, 64, 128, 256, 512, 1024\}$.
- **Hardness Parameter** κ: 7 logarithmically-spaced values for each nonlinear PDE.
- **Activations**: We test both `tanh` and `ReLU` activations.
- **Random Seeds**: Each configuration is run with 5 random seeds for statistical robustness.

Collocation points are sampled using a Sobol sequence for the Poisson equation and uniform random sampling for the nonlinear PDEs. Error is measured as the mean relative L_2 error, $\|\hat{u} - u_{\text{true}}\|_2 / \|u_{\text{true}}\|_2$, evaluated on a fine test grid against high-fidelity analytical or numerical ground truth solutions.

To test our two-part hypothesis, we perform two analyses. First, we fit the univariate scaling error $\approx AN^{-\alpha}$ at each fixed κ value to analyze the function $\alpha(\kappa)$ (results in Tab. 2). Second, we fit and compare several multivariate scaling laws, including the simple separable model error $\approx AN^{-\alpha}\kappa^{\gamma}$ (Tab. 1) and a more complex, non-separable interaction model (see Appendix C), to quantify the coupling between width and nonlinearity.

4 Results and Discussion

Our experiments reveal a significant and often counterintuitive relationship between network width, PDE nonlinearity, and final solution error. We find that the practical performance of SLN-PINNs is consistently dominated by optimization challenges, which prevent the realization of theoretical approximation benefits and can lead to pathological scaling behavior where wider networks perform worse.

4.1 Evidence of Width Scaling Pathology on a Linear Benchmark

We first test our framework on the linear Poisson equation, whose solution with a smooth forcing term is analytic. For such highly smooth functions, classical approximation theory suggests that neural networks can achieve faster convergence rates than the general Barron rate for less regular functions. [15, 16] These theoretical results suggest an error decay of $\mathcal{O}(N^{-k/d})$, where k relates to the function's smoothness, which for our one-dimensional ($d = 1$) problem, could imply a width scaling exponent $\alpha > 0.5$.

The results are given in Fig. 1. `Tanh`-activated networks successfully converge to low-error solutions ($\approx 10^{-3}$), though with high variance across seeds and no consistent scaling trend ($\alpha \approx 0.06$). In contrast, `ReLU`-activated networks catastrophically fail to learn; the error remains high (≈ 1.0) regardless of network width, yielding an effective scaling exponent of $\alpha \approx 0.01$. This is a classic example of spectral bias. While a `ReLU` network can theoretically approximate the solution, its second derivative (penalized by the u_{xx} term) is a sparse set of Dirac delta functions. [8] This makes it particularly ill-suited for representing the smooth, continuous derivatives required by the PDE loss [7]. The C^{∞} `tanh` activation, being smoother, suffers less severely but is still affected by a baseline failure.
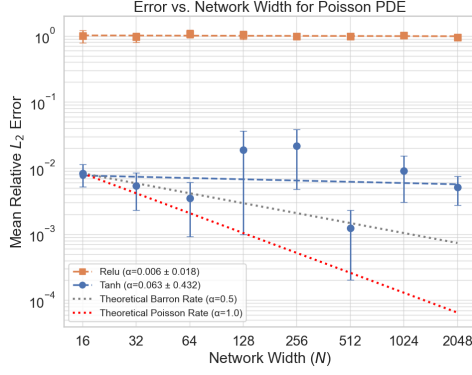
Figure 1: Error vs. Network Width (N) for the Poisson PDE. `Tanh` networks find low-error solutions but exhibit high variance and no clear scaling ($\alpha \approx 0.06 \pm 0.4$). `ReLU` networks fail to learn ($\alpha \approx 0.01 \pm 0.01$). The gray and red lines give the theoretical error decay rates of $\mathcal{O}(N^{-1/2})$ and $\mathcal{O}(N^{-1})$ respectively. It should be noted that the confidence intervals of the `tanh` error often intersect with theoretical estimates, but no consistent trend is observed.

4.2 Compounding Pathology: Non-Separable Scaling with Nonlinearity

Our central finding is that for nonlinear PDEs, the optimization failure is compounded by the problem's nonlinearity. This failure is not an artifact of averaging across different hardness levels. As shown in detail in Appendix B, fitting the scaling law error $\approx A \cdot N^{-\alpha}$ individually for each fixed κ consistently yields width exponents α that are near-zero or negative. This confirms the baseline pathology is pervasive across all conditions tested.

To quantify the average effect of both width and hardness, we fit the multivariate separable power law. The results are presented in Tab. 1. The width exponent α is consistently near-zero or negative, confirming that wider networks do not help and can even be harmful. The hardness exponent γ is generally positive, confirming that problems become harder as κ increases, with the notable exception of the Allen-Cahn equation, which we discuss later.

Fig. 2 visualizes how this pathology is exacerbated as nonlinearity increases, using the Sine-Gordon equation as a representative example. We observe that the width scaling exponent α becomes a more complex, non-monotonic function of the hardness κ (Fig. 2a), demonstrating that a single scaling exponent is insufficient. While the final error generally increases with hardness, the relationship between network width and error is unstable and often detrimental. Notice, furthermore, that while a change in N typically results in less than an order of magnitude difference in error, a change in hardness κ can alter the error by several orders of magnitude (Fig. 2b). This suggests that for these problems, nonlinearity can be a far more dominant factor than network width.

Nonlinear terms in the PDE create a fundamentally more complex and non-convex loss landscape. As the problem becomes more nonlinear, the optimization fails not because of a lack of network capacity but because standard gradient-based methods find it more difficult to locate a good solution. This

Table 1: Overall scaling exponents from the separable power-law regression error $\approx A \cdot N^{-\alpha} \cdot \kappa^{\gamma}$. Here, α represents the average width scaling, while γ quantifies the error's sensitivity to the hardness parameter κ. Statistical significance based on the p-value of the fitted coefficient is indicated by: $p < 0.01$ (**), $p < 0.001$ (***).

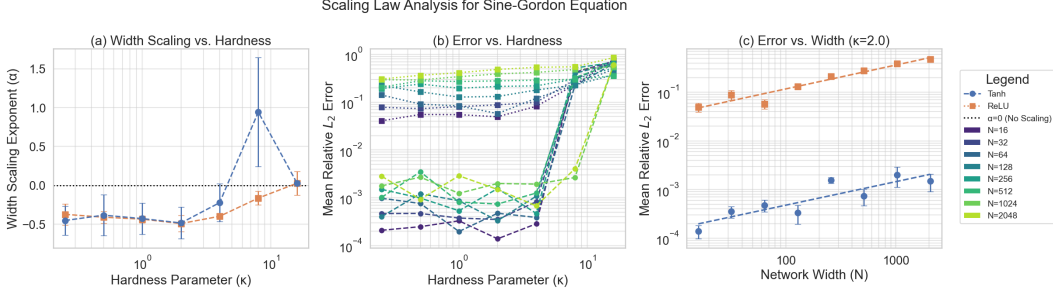| PDE	Activation	Width Exp. ($\alpha \pm 95\%$ CI)	Hardness Exp. ($\gamma \pm 95\%$ CI)	log(A)	Adj. R^2
KdV	ReLU	-0.05 ± 0.03**	0.17 ± 0.04***	-0.66	0.65
	Tanh	0.00 ± 0.01	0.18 ± 0.01***	-0.27	0.51
Sine-Gordon	ReLU	-0.32 ± 0.07***	0.28 ± 0.08***	-3.40	0.70
	Tanh	-0.14 ± 0.34	1.51 ± 0.39***	-7.30	0.52
Allen-Cahn	ReLU	-0.37 ± 0.08***	-0.44 ± 0.10***	-3.36	0.74
	Tanh	-0.02 ± 0.11	-0.03 ± 0.12	-6.78	-0.03

Figure 2: Scaling law analysis for the Sine-Gordon equation. **(a)** Width scaling exponent α vs. hardness κ. Often $\alpha < 0$, implying increasing network width also increases error. **(b)** Final error vs. hardness κ for different network widths N. The final error degrades significantly at a certain inflection point of hardness, indicating some kind of regime shift. **(c)** A representative example of error vs. width N for the median hardness value $\kappa = 2.0$ demonstrates that increased width fails to reduce error. Error bars denote standard error over 5 seeds.

aligns with recent theoretical work showing that standard intuitions from infinitely-wide networks (like the NTK) are misleading in these nonlinear regimes. [12]

This complex behavior, where the effect of width appears dependent on the level of nonlinearity, strongly suggests that the simple separable scaling law is insufficient. We test this hypothesis directly by fitting a more complex, non-separable model with an interaction term. The full analysis, presented in Appendix C, quantitatively confirms this suspicion and reveals a key mechanistic difference between the activation functions, providing strong evidence for our second hypothesis. The interaction term is statistically significant in every case for `ReLU` but in no case for `tanh`, suggesting an inheritance of κ-dependent stiffness in the former case.

We observe similar scaling pathologies for the KdV and Allen-Cahn equations albeit with their own subtleties (see Tab. 1). While the dispersive (KdV) and hyperbolic (Sine-Gordon) equations show the expected trend of error increasing with hardness ($\gamma > 0$), the reactive-parabolic Allen-Cahn equation is a notable exception. For `ReLU`, γ is significantly negative, suggesting a qualitatively different failure mechanism. For `tanh`, the model fits poorly (negative Adj. R^2), and it can be visually observed in Appendix A that the error is lower but skewed by the final $N = 1024$ point.

5 Conclusion and Future Work

In this work, we introduced a methodology for establishing empirical scaling laws for SLN-PINNs, and identified a dual optimization failure from non-convexity and spectral bias.

First, we demonstrated a width scaling pathology: standard, gradient-trained PINNs fail to achieve the positive scaling with width ($\alpha \approx 0.5$) conjectured by approximation theory. We find $\alpha \approx 0$ (wider does nothing) or $\alpha < 0$ (wider is worse), a clear sign that optimization, not approximation capacity, is the bottleneck. Second, we identified a compounding effect of nonlinearity, where increasing nonlinearity κ (and, for our set of PDEs, solution frequency) causes the scaling law itself to break down. We showed (Appendix C) that for `ReLU`, this manifests as a non-separable interaction, while for `tanh`, network width ceases to be a statistically significant factor at all. We further argue that this effect depends heavily on the specific PDE.

This study provides a quantitative benchmark on the practical limits of standard PINNs. It suggests that a "brute-force" approach of simply using wider shallow networks is an inefficient strategy. We wish to emphasize that the main purpose of this work is a call-to-action for such scaling studies in other settings. This work uses Adam, SLN-PINNs, and a small set of PDEs with one spatial dimension to make the benefits of this approach clear. Future work should aim to benchmark and identify architectures (e.g., multi-layer, Fourier features, attention-based) and optimizers (e.g., adaptive weighting, second-order methods) that can (1) close the gap between theoretical and empirical scaling of width and (2) demonstrate robustness to problem hardness and spectral bias. Overall, while it is reasonable to assume that nonlinearity causes difficulty for PINNs, it should be further investigated whether different types of nonlinear equations have different scaling laws.

5

References

[1] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[3] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 1989.

[4] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 1989.

[5] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3), 1993.

[6] Aditi S Krishnapriyan, Amir Gholami, Dillon McQueen, Agueda Lapedriza, Oxana Tort-Cases, Julius Buhler, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Sifan Wang, Xinling Yu, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 44(1), 2022.

[8] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

[9] Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. Finite basis physics-informed neural networks (fbpinns): a scalable domain decomposition approach for solving differential equations. *Advances in Computational Mathematics*, 49(4), 2023.

[10] Yang You, Igor Gitman, and Boris Ginsburg. The effect of network width on the performance of large-batch training. *arXiv preprint arXiv:1812.04252*, 2018.

[11] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[12] Andrea Bonfanti, Giacomo Bruno, and Carlo Cipriani. The challenges of the nonlinear regime for physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[13] Bryan Reyes, Amanda A Howard, Paris Perdikaris, and S Karni. Learning physics-based models of fluid dynamics from sparse data. In *AAAI Conference on Artificial Intelligence*, volume 35, 2021.

[14] Amir Arzani, Jian-Xun Wang, and Roshan M D'Souza. Uncovering near-wall blood flow from sparse data with physics-informed neural networks. *Physics of Fluids*, 33(7), 2021.

[15] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 2017.

[16] Ronald A DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30, 2021.

A Allen-Cahn and KdV Scaling Analysis

In this section, we present the full scaling law analysis for the Allen-Cahn and KdV equations.

Fig. 3 shows the scaling analysis for the Allen-Cahn equation. For both `tanh` and `ReLU` activations, the width scaling exponent α is consistently negative or near-zero, indicating that increasing network width provides no benefit and often harms performance. This pathology is especially severe for `ReLU` networks, where the scaling exponent becomes progressively more negative as the problem hardness κ increases (Fig. 3a). Furthermore, there is a performance gap between the activation functions; `tanh` networks achieve a low error that remains remarkably stable across all hardness values, whereas `ReLU` networks perform orders of magnitude worse with a width scaling exponent that becomes more negative as hardness increases.
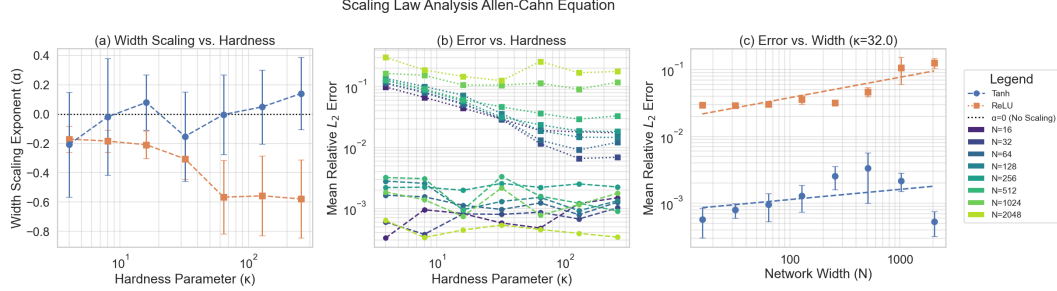

Figure 3: Scaling law analysis for the Allen-Cahn equation. **(a)** Width scaling exponent α vs. hardness $\kappa = 1/D$. For both activations, α is consistently negative, indicating wider networks perform worse. **(b)** Final error vs. hardness κ. `Tanh` networks (dashed lines) achieve low error that is remarkably stable against increasing hardness, while `ReLU` networks (dotted lines) perform orders of magnitude worse. **(c)** A representative example of error vs. width N for the median hardness value $\kappa = 32.0$, clearly showing the negative scaling trend ($\alpha < 0$). Error bars denote standard error over 5 seeds.

Fig. 4 shows the scaling analysis for the KdV equation. A sharp degradation in solution quality is observed as the hardness parameter κ increases (Fig. 4b). The analysis of the width scaling exponent α again reveals a failure to benefit from larger networks. For `tanh`, α remains consistently close to zero across all hardness values, indicating no performance gain from increasing width. For `ReLU`, α begins as negative and trends toward zero as κ increases (Fig. 4a). While `tanh` generally achieves a lower error than `ReLU`, their overall performance is more comparable than in the Allen-Cahn case, suggesting some interaction between activation function and the type of PDE nonlinearity.
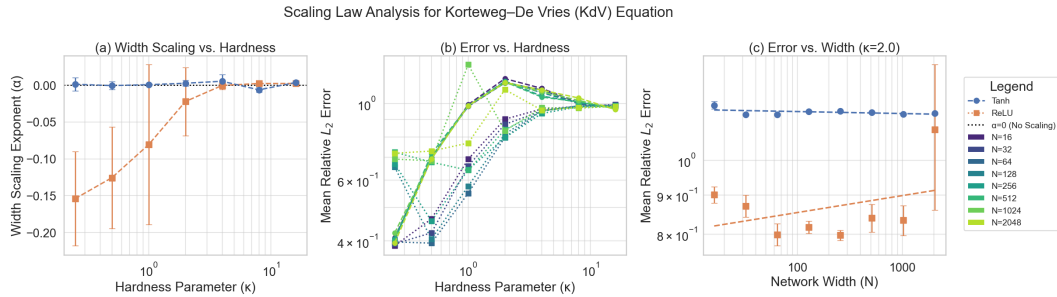

Figure 4: Scaling law analysis for the Korteweg-de Vries (KdV) equation. **(a)** Width scaling exponent α vs. hardness $\kappa = A$. For both `ReLU` and `tanh`, α is consistently near or below zero, indicating no performance gain from increasing network width. **(b)** Final error vs. hardness κ. Error increases with hardness for both activations, though `tanh` generally performs better. **(c)** A representative example of error vs. width N for $\kappa = 2.0$, visualizing the flat ($\alpha \approx 0$) or negative scaling. Error bars denote standard error over 5 seeds.

B Table of Univariate Regressions on Width

Table 2: Fitted width scaling exponents α for each hardness parameter κ and activation function. The exponent is derived from fitting the mean error over 5 seeds to the power law error $\approx A \cdot N^{-\alpha}$.

PDE	Hardness (κ)	ReLU ($\alpha \pm 95\%$ CI)	Tanh ($\alpha \pm 95\%$ CI)
Poisson	N/A	0.01 ± 0.02	0.06 ± 0.43
KdV	0.25	-0.15 ± 0.06	0.00 ± 0.01
	0.50	-0.13 ± 0.07	0.00 ± 0.01
	1.00	-0.08 ± 0.11	0.00 ± 0.01
	2.00	-0.02 ± 0.05	0.00 ± 0.01
	4.00	0.00 ± 0.01	0.01 ± 0.01
	8.00	0.00 ± 0.01	-0.01 ± 0.01
	16.00	0.00 ± 0.01	0.00 ± 0.01
Sine-Gordon	0.25	-0.37 ± 0.14	-0.45 ± 0.19
	0.50	-0.41 ± 0.08	-0.38 ± 0.26
	1.00	-0.43 ± 0.04	-0.43 ± 0.20
	2.00	-0.49 ± 0.10	-0.48 ± 0.20
	4.00	-0.40 ± 0.03	-0.22 ± 0.24
	8.00	-0.16 ± 0.09	0.94 ± 0.70
	16.00	0.03 ± 0.16	0.03 ± 0.01
Allen-Cahn	4.00	-0.17 ± 0.09	-0.21 ± 0.36
	8.00	-0.18 ± 0.08	-0.02 ± 0.40
	16.00	-0.21 ± 0.09	0.08 ± 0.19
	32.00	-0.31 ± 0.14	-0.15 ± 0.30
	64.00	-0.57 ± 0.25	0.00 ± 0.27
	128.00	-0.56 ± 0.27	0.05 ± 0.25
	256.00	-0.58 ± 0.27	0.14 ± 0.25

C   Scaling Laws for Interaction Between Nonlinearity and Width

To test our hypothesis that the simple, separable scaling law (Tab. 1) is insufficient, we fit a more complex, non-separable interaction model. Previously the simple model was $\log(\text{error}) = \beta_0 + \beta_N \log(N) + \beta_\kappa \log(\kappa)$. Tab. 3 gives the results for fitting a model with some interaction term, which is now defined as:

$$\log(\text{error}) = \beta_0 + \beta_N \log(N) + \beta_\kappa \log(\kappa) + \nu(\log(N) \cdot \log(\kappa)). \tag{2}$$

A statistically significant interaction coefficient ν (i.e., $p < 0.05$) would suggest that the scaling with width N is dependent on the nonlinearity κ.

Table 3: Scaling exponents for the interaction model defined in (2). Statistical significance based on the p-value of the fitted coefficient is indicated by: $p < 0.01$ (**), $p < 0.001$ (***).

PDE	Activation	β_0 (Intercept)	β_N (Width)	β_κ (Hardness)	ν (Interaction)	Adj. R^2
KdV	ReLU	-0.809***	0.083***	0.386***	**-0.042 *****	**0.741**
	Tanh	-0.271**	-0.001	0.178**	0.000	0.498
Sine-Gordon	ReLU	-3.716***	0.382***	0.740***	**-0.089 *****	**0.758**
	Tanh	-8.094***	0.296	2.666***	-0.222	0.542
Allen-Cahn	ReLU	-1.202**	-0.048	-1.061***	**0.120 *****	**0.811**
	Tanh	-7.800***	0.212	0.260	-0.057	-0.008

The results in Tab. 3 provide strong evidence for a non-separable scaling relationship, particularly for `ReLU` networks. For every nonlinear PDE, the interaction term ν for `ReLU` is statistically significant, and the adjusted R^2 of the interaction model is notably higher than that of the simple separable model in Table 1. This is the statistical signature of a brittle failure mode: the scaling of `ReLU` with network width is actively exacerbated by problem hardness. This is consistent with the spectral bias theory, as the non-smooth `ReLU` activation creates a particularly difficult optimization landscape when forced to represent the high-frequency solutions associated with higher κ [8].

In contrast, both the width term β_N and the interaction term ν for `tanh` are not statistically significant for any of the PDEs. For `tanh` networks, network width appears to be an irrelevant variable for predicting the final error. The failure mode is not one of complex interaction, but rather a consistent failure to benefit from increased capacity. This points to a more "graceful" failure mode. The inherent smoothness of the `tanh` function, while still subject to spectral bias, leads to a more stable, albeit suboptimal, optimization dynamic.

The Allen-Cahn equation serves as a particularly interesting case study that highlights the limits of this power-law model. For the `tanh` network, the extremely poor fit of the model (negative adjusted R^2) is itself a result; it confirms the visual evidence from Fig. 3, where the error is low and stable across all N and κ. For the `ReLU` network on Allen-Cahn, we note the unusual negative, statistically-significant hardness coefficient β_κ, a behavior not seen in the other PDEs that suggests its failure mechanism may be qualitatively different in this reactive-parabolic setting.