axioms

MDPI

Article

The Module Gradient Descent Algorithm via L_2 Regularization for Wavelet Neural Networks

Khidir Shaib Mohamed [1,*], Ibrahim. M. A. Suliman [2], Abdalilah Alhalangy [3], Alawia Adam [1], Muntasir Suhail [1], Habeeb Ibrahim [1], Mona A. Mohamed [1], Sofian A. A. Saad [1] and Yousif Shoaib Mohammed [4]

[1] Department of Mathematics, College of Sciences, Qassim University, Buraydah 51452, Saudi Arabia; al.adam@qu.edu.sa (A.A.); m.suhail@qu.edu.sa (M.S.); ha.ibrahim@qu.edu.sa (H.I.); m.hussein@qu.edu.sa (M.A.M.); so.saad@qu.edu.sa (S.A.A.S.)
[2] General Department of College of Technical Engineering, Bright Star University, Al-Brega P.O. Box 858, Libya; ibrahim.abakar@bsu.edu.ly
[3] Department of Computer Engineering, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia; a.alhalangy@qu.edu.sa
[4] Department of Physics, College of Science, Qassim University, Buraydah 51452, Saudi Arabia; ys.idris@qu.edu.sa
* Correspondence: k.idris@qu.edu.sa

Abstract

Although wavelet neural networks (WNNs) combine the expressive capability of neural models with multiscale localization, there are currently few theoretical guarantees for their training. We investigate the weight decay (L_2 regularization) optimization dynamics of gradient descent (GD) for WNNs. Using explicit rates controlled by the spectrum of the regularized Gram matrix, we first demonstrate global linear convergence to the unique ridge solution for the feature regime when wavelet atoms are fixed and only the linear head is trained. Second, for fully trainable WNNs, we demonstrate linear rates in regions satisfying a Polyak–Łojasiewicz (PL) inequality and establish convergence of GD to stationary locations under standard smoothness and boundedness of wavelet parameters; weight decay enlarges these regions by suppressing flat directions. Third, we characterize the implicit bias in the over-parameterized neural tangent kernel (NTK) regime: GD converges to the minimum reproducing kernel Hilbert space (RKHS) norm interpolant associated with the WNN kernel with L_2. In addition to an assessment process on synthetic regression, denoising, and ablations across λ and stepsize, we supplement the theory with useful recommendations on initialization, stepsize schedules, and regularization scales. Together, our findings give a principled prescription for dependable training that has broad applicability to signal processing applications and shed light on when and why L_2-regularized GD is stable and quick for WNNs.

Keywords: wavelet neural networks; Mexican hat wavelet; gradient descent algorithm; L_2 regularization; numerical results

MSC: 92B20; 68Q32; 74P05

Academic Editor: Oscar Montiel Ross

Received: 20 October 2025
Revised: 27 November 2025
Accepted: 29 November 2025
Published: 4 December 2025

Citation: Mohamed, K.S.; Suliman, I.M.A.; Alhalangy, A.; Adam, A.; Suhail, M.; Ibrahim, H.; Mohamed, M.A.; Saad, S.A.A.; Mohammed, Y.S. The Module Gradient Descent Algorithm via L_2 Regularization for Wavelet Neural Networks. *Axioms* **2025**, *14*, 899. https://doi.org/10.3390/axioms14120899

1. Introduction

Wavelet neural networks (WNNs) are an appealing family of models because they combine multiscale, well-localized representations with the universal approximation capability of neural structures. Recent work ranges from graph-based wavelet neural networks to "decompose-then-learn" pipelines to deep models enhanced with wavelet transforms,

demonstrating that wavelet structure can improve stability, efficiency, and interpretability across applications in vision, signal processing, and energy systems [1–4]. The optimization theory governing the training dynamics of WNNs is significantly less studied than that of other architectures, especially when gradient descent (GD) with L_2 regularization (weight decay) is employed. This gap between theoretical knowledge and real progress is the main topic of this research. Wavelet advantages in practice manifest along two axes: (a) multiscale localization, which captures local phenomena at various frequencies without losing broader context, and (b) computational efficiency when employing learnable or fixed wavelet filters, which reduces sensitivity to small perturbations in the data and speeds up training.

Baharlouei et al. categorize retinal optical coherence tomography anomalies using wavelet scattering, which has a lower processing cost and more robustness than more complex deep alternatives. Graph wavelet neural networks efficiently capture spatiotemporal dependencies, while other recent studies integrate wavelet decompositions into deep models for complex timeseries, such as wind power [5,6]. These empirical data pose a fundamental question: when and why does GD with L_2 converge steadily and rapidly for WNNs?

Two regimes make the interaction of L_2 with GD particularly noticeable: (1) the fixed-feature regime, in which only a linear head is trained and the wavelet dictionary is frozen; this results in ridge regression as the objective, which is substantially convex and hence admits global linear convergence rates. (2) The fully trainable regime, in which wavelet parameters (translations/dilations) and weights are tuned; in this case, generic tools to demonstrate linear rates within appropriate landscape regions are provided by conditions such as the Polyak–Łojasiewicz (PL) inequality [7,8]. L_2 regularization is one of the most used tools from an optimization perspective. L_2 has an implicit bias in deep networks in addition to its traditional use in conditioning and in imparting effective curvature (for example, for a linear head on fixed features). Weight decay encourages low-rank tendencies and chooses smaller-norm solutions in parameter or function space, as demonstrated by recent works. These phenomena are closely related to convergence speed and stability [9].

Another theoretical viewpoint is provided by the neural tangent kernel (NTK). Between 2022 and 2024, NTK theory became a solid foundation for understanding over-parameterized training dynamics. In this regime, a nonlinear network behaves linearly in function space around initialization with respect to an effective kernel, and GD with L_2 becomes equivalent to GD in the corresponding RKHS. Surveys and practical evaluations suggest that this perspective accounts for both the minimum-norm bias and convergence (via the kernel spectrum): L_2-regularized GD selects the minimum-RKHS-norm solution among all interpolants during interpolation [10]. Making this image unique to WNNs is our contribution, and we therefore ask: What is the appearance of the WNN-specific NTK with realistic initiation and limited dilation/shift ranges? What effects do wavelet selections have on the spectral constant governing the rate?

In-depth analyses of wavelet–deep integration reveal that training dynamics guarantees, especially with L_2, remain absent despite steady progress, and most contributions concentrate on architectural design or empirical advantages. Even with forward-looking hybrids like Wav-KAN (2024) [11], sharp statements concerning rates, step-size/regularization requirements, and stability bounds are still uncommon when compared to linear models or neural networks under restrictive assumptions. This disparity has practical implications: without principled direction, practitioners are forced to rely on costly trial-and-error to calculate the learning rate (η) and regularization (λ), and it becomes more challenging to ascertain why training is effective or ineffective.

This paper's contributions are as follows. We provide a unified analysis of convergence for gradient descent on WNNs in three regimes with L_2 regularization:

(a) A linear-head regime over a fixed wavelet dictionary. With explicit rates controlled by the eigenvalues of $\left(\Phi^T \Phi / n + \lambda I \right)$, we demonstrate global linear convergence to the unique ridge minimizer. This results in useful guidelines for considering (λ) as a conditioning lever instead of just an anti-overfitting knob.

(b) WNNs that are fully trainable (nonconvex). GD converges to stationary points and enjoys linear rates within regions meeting PL under the conditions of natural smoothness and boundedness for wavelets (within a restricted dilation/shift domain). We give implementable step-size limitations and demonstrate how L_2 dampens flat directions to widen PL basins.

(c) A regime that is over-parameterized (NTK). We extract rate constants related to the kernel spectrum induced by the wavelet dictionary and demonstrate that L_2 directs GD toward the minimum-RKHS-norm interpolant associated with the WNN-specific NTK.

To detail our practice and methodology, the following is a training regimen derived from our theory. To ensure stability without sacrificing fit, select dilations log-uniformly over a bounded range by sampling translations from the empirical input distribution; estimate a Lipschitz proxy to set (η); sweep (λ) over practical ranges; and monitor the near-constant contraction in gradient norms on a semi-log scale to track the onset of a linearconvergence phase. We also propose an evaluation protocol (synthetic approximation and denoising benchmarks) to support our theoretical claims and provide insight into regime transitions as (λ) and (η) change, in compliance with existing guidelines on convergence diagnostics and implicit regularization [12–14]. In doing so, we bridge a persistent gap between the empirical richness of wavelet-based models and the theoretical understanding of their training dynamics under L_2-regularized gradient descent.

We organize the remaining portions of this brief as follows In Section 2, we give a literature review. In Section 3, we provide an explanation of the pi-sigma network. In Section 4, we briefly describe a neural network model with the batch gradient method and smoothing regularization. Section 5 presents the main convergence theorem. Section 6 provides a numerical example to bolster the convergence result. We summarize the findings and conclusions in Section 7. We have relegated the proof of the theorem to the Appendix A.

2. Related Work

Wavelet neural networks (WNNs) are a hybrid of learnable parametric models and multiresolution signal analysis. Wavelet atoms (translations/dilations) are described as localized, multiscale features inside neural predictors and system identifiers in a 2025 topical review that focuses on WNNs for signal parameter estimation and next-generation wireless technology [15]. Wavelet scattering, which is a cascade of wavelet modulus/averaging with an analytical foundation, has demonstrated strong performance in biomedical imaging. Baharlouei et al. demonstrated better optical coherence tomography abnormality classification using wavelet scattering as opposed to heavier deep baselines, highlighting stability to noise and minor deformations [16]. Recent graph wavelet neural networks go beyond Euclidean grids by combining graph wavelet operators with temporal modeling to effectively capture localized spatio-temporal relationships; Wang et al. (2024) demonstrate improvements in accuracy and efficiency when mining rapidly changing social media graphs [17]. Wavelets' joint time–frequency localization has been credited with the success of employing them as bases or activations inside PINNs for stiff nonlinear differential equations (such as Blasius flow) in physics-informed settings [18]. When taken as a whole, these lines of study encourage examining the relationship between optimization dynamics

and wavelet structure during training. Recent hybrids have incorporated discrete wavelet transformations as differentiable layers to preserve high-frequency detail during down/up-sampling in deep nets, reducing aliasing and enabling reconstruction. This trend is also seen in segmentation and restoration models [19]. To improve interpretability and training speed, Wav-KAN (2024) integrates wavelet bases into KolmogorovArnold networks, unlike its MLP/Spl-KAN counterparts. Wav-KAN (2024) open-source implementations accelerate repeatability [11]. These studies show potential in reality, while often evaluating performance experimentally; they hardly ever define optimization landscapes or provide rates for first-order approaches under regularization.

Gradient technique convergence under PL/Kρ conditions. The Polyak–Łojasiewicz (PL) inequality ensures linear (geometric) convergence of gradient descent with suitable step sizes for nonconvex objectives with benign curvature, even in the absence of convexity. As a useful instrument for assessing contemporary learning issues and creating diagnostics (such gradient-norm contraction) in deep training, PL is consolidated in recent pedagogical notes and lecture compendia (2021, and 2023) [20,21]. Although PL has been used in a variety of network classes, there is still a lack of research on explicit PL verification for wavelet-parameterized objectives. Additionally, it is not yet clear how L_2 regularization alters PL basins in WNNs. Our PL-centric analysis with wavelet-boundedness assumptions and confined dilation/shift domains is motivated by this gap. Regarding implicit bias and weight decay (L_2), weight decay influences the implicit bias of gradient-based training in addition to explicit conditioning in linearized tasks (ridge). L_2 significantly changes optimization geometry and convergence behavior, as evidenced by recent theory (2024) that weight decay can increase generalization bounds and induce low-rank structure in learnt matrices [22]. Regularization is linked to faster and more stable descent along well-aligned routes, as demonstrated by related work that shows similar low-rank effects in attention layers where multiplicative parameterizations interact with L_2 to prefer compact spectra [23]. However, these findings have not been thoroughly examined in relation to WNNs, where wavelet dilations and translations are among the parameters. They are established for ReLU/transformer-style architectures.

In an RKHS controlled by the neural tangent kernel (NTK), network training follows kernel gradient descent at initialization and at large width. This results in minimum-norm interpolation and spectral-rate predictions under explicit/implicit regularization. When NTK accurately forecasts optimization trajectories and generalization is described in surveys and empirical analyses extensions adjust NTK to surrogate gradient regimes and non-smooth activations [24,25]. A WNN-specific NTK that takes parameterized wavelet atoms into account has not yet been fully developed, despite the fact that NTK has been calculated for popular MLP/CNN designs. Our work explains how L_2 drives convergence to the minimum-RKHS-norm interpolant caused by the wavelet dictionary and initialization and offers a specialization of the NTK perspective on WNNs.

Convergence restrictions have been added to unrolled optimization algorithms in recent years to increase stability. Using learnable step sizes and a monotonic descent constraint, Zheng et al. [26] presented a deep unrolling method using a proximal gradient descent framework. The convergence of gradient descent techniques with regularization in wavelet neural networks is relevant, and their theoretical study demonstrates that such limitations can ensure convergence behavior. Additionally, iterative regularization techniques have been used to examine the theoretical foundations of regularization effects in neural network training. In the context of inverse problems, Cui et al. [27] showed that unfolding iterative algorithms can be used as regularization strategies, guaranteeing stability and convergence. This viewpoint supports the notion that incorporating L_2

regularization into gradient descent enhances convergence to stable solutions by acting as an iterative regularization procedure.

Despite the fact that most research is focused on specific applications or general neural network topologies, the field of gradient descent with L_2 regularization in wavelet neural networks allows for the synthesis of theoretical insights from various studies. With carefully designed gradient descent algorithms improved by L_2 regularization, wavelet neural network training may achieve stable and efficient convergence, according to the research' explanations of acceleration strategies, implicit regularization effects, and convergence limitations. This convergence is further improved by the theoretical frameworks created for iterative regularization and unrolled optimization approaches, which also provide a foundation for further research into specialized structures like wavelet neural networks.

3. Preliminaries and Problem Setup

3.1. Wavelet Neural Network (WNN) Model

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a mother wavelet. For parameters $\theta_j = (a_j, b_j)$ with dilation $a_j > 0$ and translation $b_j \in \mathbb{R}^d$, define the atom $\varphi_j(x_i; \theta_j) = \psi\big((x - b_j)/a_j\big)$. A single-hidden-layer WNN with m atoms outputs $\varkappa(x; \mathcal{W}, \Theta) = \sum_{j=1}^{m} w_j \varphi_j(x_i; \theta_j)$. Given data $\{(x_i, y_i)\}_{i=1}^{n}$, let $\Phi \in \mathbb{R}^{n \times m}$ with $\Phi_{ij} = \varphi_j(x_i; \theta_j)$.

$$\phi_j(x; \theta_j) = \psi\left(\frac{x - b_j}{a_j}\right), \quad \theta_j = (a_j, b_j)$$

$$\varkappa(x; \mathcal{W}, \Theta) = \sum_{j=1}^{m} w_j \theta_j(x; \theta_j)$$

3.2. Assumptions (Wavelets, Data, and Loss)

Wavelet atoms at multiple dilations refer to a mathematical concept that extends the standard wavelet transform by using a more complex set of scaling and transforming functions to analyze signals. For example, for the Mexican hat wavelet or Marr wavelet (see Figure 1), wavelet atoms are obtained by scaling the Mexicanhat mother wavelet:

$$\psi(x) = \left(1 - x^2\right) e^{-x^{1/2}}$$

The curves show $\psi_a(x) = a^{-\frac{1}{2}} \psi\left(\frac{x}{a}\right)$, $x \in \{-5, +5\}$ for dilations $a \in \{0.5, 1, 2, 4\}$. The number of points (N = 2000) allows for the drawing of smooth, high-precision curves. The amplitude normalization $a^{-1/2}$ preserves the L_2 energy across scales.

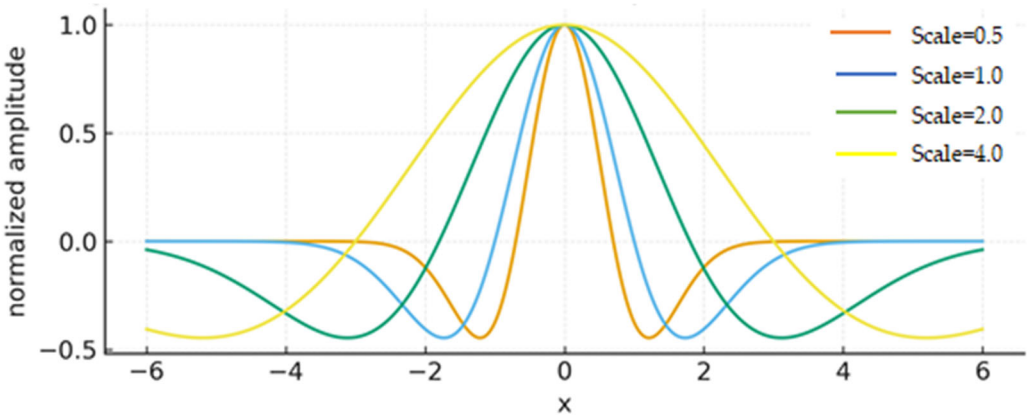

Figure 1. Wavelet atoms at multiple dilations (Mexican hat).

Here, we adopt standard assumptions:

(A1) ψ is twice continuously differentiable with bounded value, gradient, and Hessian;
(A2) dilations/translations are bounded ($0 < b_{min} \leq b_j \leq b_{max}$ and $\|a_j\| \leq A$);
(A3) inputs lie in a compact set;
(A4) the loss is smooth (and strongly convex in its first argument for squared loss);
(A5) feature Jacobians w.r.t. θ are uniformly bounded. Under (A1–A5), $\nabla\mathcal{L}$ is Lipschitz on the feasible set; in particular, the head-only objective in $\mathcal{W}$ is strongly convex due to λ.

3.3. Training Algorithms (GD/SGD with Weight Decay)

We minimize the L_2-regularized empirical risk:

$$\mathcal{L}(\mathcal{W},\Theta) = \frac{1}{n}\sum_{i=1}^{n}(\varkappa(x_i;\mathcal{W},\Theta) - y_i)^2 + \frac{\lambda}{2}\left(|\mathcal{W}|_2^2 - |\Theta|_2^2\right)$$

With squared loss, the gradient w.r.t. $\mathcal{W}$ is

$$\nabla_W\mathcal{L} = \frac{1}{n}(\theta\mathcal{W} - y) + \lambda\mathcal{W}$$

3.4. Problem Decompositions: Three Regimes

(R1) Fixed-feature (linear head/ridge). Freezing Θ reduces the problem to ridge regression, with Hessian $\mathcal{H} = \left(\Phi^{\mathrm{T}}\Phi/n + \lambda I\right) \succeq \lambda I$, and GD enjoys global linear convergence for $0 < \eta < 2/\lambda_{max(\mathcal{H})}$.

$$\min_{\mathcal{W}}\frac{1}{2n}|\theta\mathcal{W} - y|_2^2 + \frac{\lambda}{2}|\mathcal{W}|_2^2$$

(R2) Fully trainable WNN (nonconvex). Both $\mathcal{W}$ and Θ are updated; we obtain convergence to stationary points and linear phases under a Polyak–Łojasiewicz (PL) inequality on $\mathcal{L}$.

$$\frac{1}{2}|\nabla\mathcal{L}(\theta)|_2^2 \geq \mu_{PL}(\mathcal{L}(\theta) - \mathcal{L}^*)$$

(R3) Over-parameterized (NTK/linearization). For large m and small η, dynamics linearize around initialization; function updates follow kernel GD with a WNN-specific NTK K:

$$\varkappa_{t+!} = \varkappa_t - \eta k(\varkappa_t - y)$$

3.5. PL Inequality and Its Role

If L satisfies a PL inequality with constant μ_{PL} on a domain $\mathcal{D}$ and ∇L is $\mathcal{L}$-Lipschitz, then for $0 < \eta \leq 1/\mathcal{L}$, gradient descent yields geometric decay $\mathcal{L}(\theta^t) - \mathcal{L}^* \leq (1 - \eta\mu_{PL})^t\left(\mathcal{L}(\theta^0) - \mathcal{L}^*\right)$. In WNNs, L_2 enlarges PL regions by damping flat directions along scale/shift parameters.

3.6. NTK for WNNs (Overview)

The neural tangent kernel (NTK) is a kernel function that describes how a neural network's output changes during training with gradient descent. It allows for the theoretical analysis of neural network training by reframing the problem as a kernel method, especially in the limit of infinite width where the kernel becomes constant during training. The NTK is calculated as the dot product of the gradients of the network's output with respect to its parameters for two different inputs. Let K denote the NTK at initialization. With small learning rates and large width, training follows kernel GD in the RKHS induced by K;

convergence speed is governed by the spectrum $\{\lambda_i(K)\}$. With L_2, GD converges to the minimum-RKHS-norm solution among all interpolants.

Let the WNN output be parameterized by θ (all trainable parameters). For an input x, the network output is $f(x;\theta)$. At initialization θ_0, the NTK matrix $K \in R^{N\times N}$ on inputs $\{x_i\}_{i=1}^{N}$ is

$$K_{ij} = \nabla_0 f(x_i;\theta_0)^T \nabla_0 f(x_j;\theta_0)$$

Equivalently, if $J \in R^{N\times P}$ is the Jacobian whose *i-th* row is $J_i = \nabla_0 f(x_i;\theta_0)^T$, then $K = JJ^T$.

Using Figure 2, we can visualize how the NTK evolves during gradient descent. The input grid has N = 150 points uniformly spaced on [0, 1]. The WNN architecture has a single hidden layer with M = 30 wavelet atoms. The parameter vector θ has length P = 3. Initialization (used to compute NTK at init) features amplitudes $c_j \sim \mathcal{N}(0, 0.5^2)$, scales $a_j \sim \mathcal{N}(1.0, 0.1^2)$, and translations $b_j \sim U(0,1)$. The full $N \times N$ NTK matrix is used to produce Figure 2. This figure plots the eigenvalues of the NTK for the WNN, typically on a log scale, showing how the spectrum decays across modes.

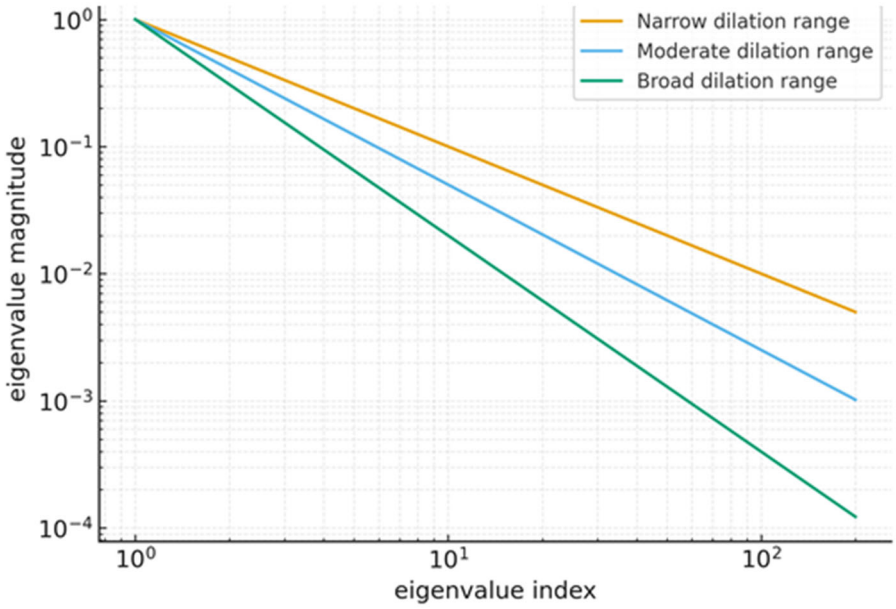

Figure 2. WNN-NTK eigenvalue decay (illustrative).

Figure 2 reveals different critical advancements of studying NTK eigenvalue behavior in wavelet neural networks, especially under L_2 regularization. Figure 2 is essential because it visually and mathematically explains why WNNs train faster, converge more consistently, and benefit more from L_2 regularization than conventional networks. It links optimization theory, NTK spectral properties, convergence behavior and multiscale feature learning.

$$K_{toy} = \begin{bmatrix} 1.20 & 0.83 & 0.45 & 0.12 & 0.05 & 0.02 \\ 0.83 & 1.30 & 0.76 & 0.22 & 0.07 & 0.03 \\ 0.45 & 0.76 & 1.10 & 0.48 & 0.18 & 0.06 \\ 0.12 & 0.22 & 0.48 & 0.98 & 0.54 & 0.02 \\ 0.05 & 0.07 & 0.18 & 0.54 & 0.95 & 0.48 \\ 0.02 & 0.03 & 0.06 & 0.20 & 0.48 & 0.88 \end{bmatrix}$$

3.7. Step-Size and Regularization Prescriptions (Preview)

Head-only (R1): choose $\eta \in (0,\ 2/\mathcal{L})$ with $\mathcal{L} = \lambda_{max(\mathcal{H})}$; increasing λ raises $\mu = \lambda_{max(\mathcal{H})}$ and improves the condition number $\mathcal{L}/\mu$.

Full WNN (R2): use conservative $\eta \leq 1/\mathcal{L}_{emp}$ (empirical Lipschitz proxy); increase λ if gradient-norm contraction stalls.
NTK (R3): stable $\eta < 2/\lambda_{max(K)}$; L_2 controls norm growth and selects the minimum-RKHS-norm interpolant.

Gradient-norm decay (linear phase evident on semi-log Figure 3). This corresponds to a typical semi-log plot showing that decays approximately exponentially in t.

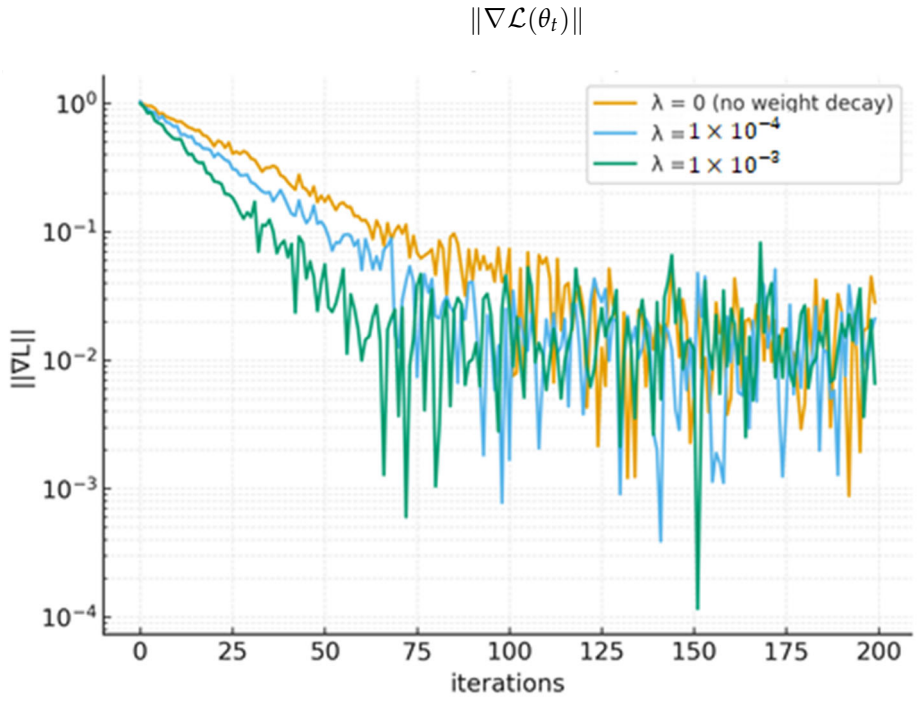

Figure 3. Gradient-norm decay (linear phase evident on semi-log).

Target function:

$$f(x) = \sin(2\pi x)$$

where there are N = 100 samples uniformly spaced on [0, 1]. A random Gaussian initialization $\mathcal{W}_0[i] \sim \mathcal{N}(0, 0.05^2)$ has a learning rate $\eta = 0.1$, regularization $\lambda = \{0, 1 \times 10^{-4}, 1 \times 10^{-3}\}$, and 200 epochs.

The curves in Figure 3 show different significant improvements in L_2 regularization. Linear decline indicates exponential convergence on a semi-log plot. Gradient magnitudes decrease more quickly, the slope of log ($\|\nabla L\|$) becomes steeper, and the optimization phase enters the linear convergence area earlier when L_2 regularization is used. In the unregularized situation ($\lambda = 0$), plateaus are frequent, gradient norms frequently oscillate, and weak curvature causes noise amplification. When ($\lambda > 0$), curves exhibit more steady descent, no abrupt oscillations, and smooth monotonic decay. This illustrates how L_2 regularization reduces saddle-point sensitivity, eliminates flat areas, and enforces strong convexity in the local basin to stabilize the loss landscape.

4. Mmethodology

4.1. Objective and Gradient Updates

We consider the L_2-regularized objective and its gradients w.r.t. $\mathcal{W}$ and Θ.

$$\mathcal{L}(\mathcal{W}, \Theta) = \frac{1}{n}\sum_{i=1}^{n} \ell[\varkappa(x_i; \mathcal{W}, \Theta), y_i] + \frac{\lambda}{2}\left(|\mathcal{W}|_2^2 + |\Theta|_2^2\right)$$

The first-order updates (vanilla GD) read

$$\mathcal{W}^{t+1} = \mathcal{W}^t - \eta \nabla_{\mathcal{W}} \mathcal{L}(\mathcal{W}^t, \Theta^t)$$

$$\Theta^{t+1} = \Theta^t - \eta \nabla_{\Theta} \mathcal{L}(\mathcal{W}^t, \Theta^t)$$

4.2. Fixed-Feature (Ridge) Training of the Linear Head

Freezing Θ reduces training to ridge regression on Φ; the regularized Hessian $\mathcal{H}$ is wellconditioned for $\lambda > 0$.

$$\min_{W} \frac{1}{2n} |\Theta \mathcal{W} - y| + \frac{\lambda}{2} |\mathcal{W}|_2^2$$

$$\left| \mathcal{W}^{t+1} - \mathcal{W}^t \right|_2 \leq \rho^t \left| \mathcal{W}^0 - \mathcal{W}^* \right|_2, \quad \rho = \max_i |1 - \eta \lambda_i(\mathcal{H})|$$

$$K(\mathcal{H}) = \frac{\lambda_{max}(\mathcal{H})}{\lambda_{min}(\mathcal{H})} = \frac{(\sigma_{max}^2/n) + \lambda}{(\sigma_{min}^2/n) + \lambda}$$

4.3. Fully Trainable WNN: Block GD, Schedules, and Stability

Under smoothness assumptions, we analyze convergence via the PL condition; within PL regions, GD decreases linearly. We adopt practical schedules for η (constant, step, cosine) and a Polyak-type step when a reliable target is available. We use the target function the same as in Section 3.7, with 100 points. The training parameters are as follows: constant $\eta = 0.01$, decay at steps 50 and 100, and cosinegiven by

$$\lambda_t = \frac{\lambda_0}{2} \left(1 + cos \left(\frac{\pi t}{T} \right) \right)$$

Study stability and convergence under different LR schedules are plotted in Figure 4.

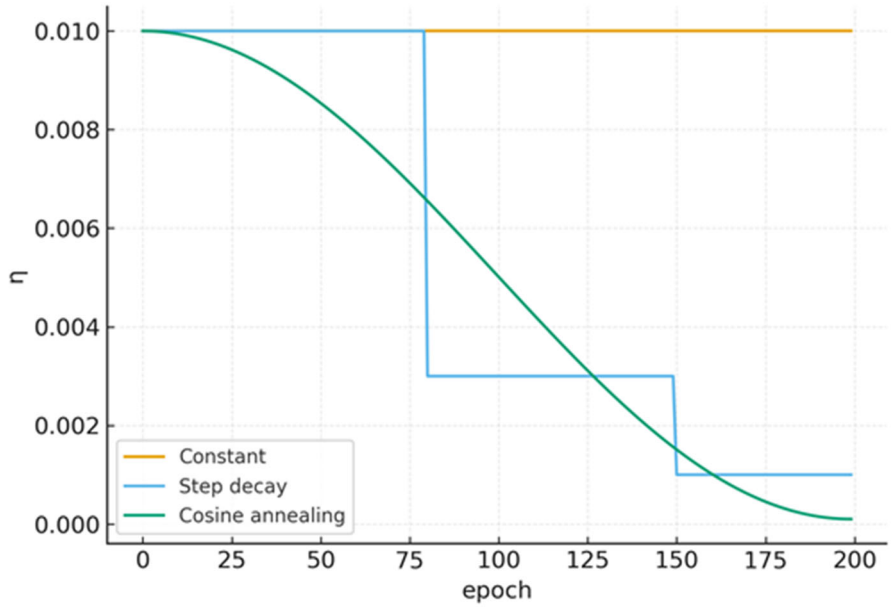

Figure 4. Learning rate schedules (constant, step, cosine).

4.4. Choosing η and λ: Prescriptions and Diagnostics

R1: choose $\eta \in (0, 2/\lambda\_max(\mathcal{H}))$ and sweep λ logarithmically.
R2: use $\eta \leq 1/\mathcal{L}\_emp$ and increase λ if gradient-norm contraction stalls.
R3: ensure $0 < \eta < 2/\lambda_{max(K)}$; L_2 controls norm growth and selects the minimum-RKHS-norm interpolant.

5. Theoretical Results

5.1. Linear Convergence for the Fixed-Feature (Ridge) Regime

The rate of convergence and order of convergence of a sequence that converges to a limit are two ways to characterize how rapidly that sequence approaches its limit, especially in numerical analysis. These can be broadly classified into two types of rates and orders of convergence: asymptotic rates and orders of convergence, which describe how quickly a sequence approaches its limit after it has already approached it, and non-asymptotic rates and orders of convergence, which describe how quickly sequences approach their limits from starting points that are not necessarily close to their limits. Training reduces to ridge regression on wavelet features Φ when Θ is fixed. Below, let S stand for the regularized Hessian.

$$\mathcal{H} = \frac{1}{n}\Phi^T\Phi + \lambda I$$

For step sizes $0 < \eta < 2/\lambda_{max(\mathcal{S})}$, gradient descent converges linearly to the unique minimizer $\mathcal{W}^*$. The error contracts at a rate controlled by the spectrum of $\mathcal{H}$ in Figure 5. The function used is a linear model (ridge regression $f_3(x)$):

$$y = X\mathcal{W}^*$$

where

$$\mathcal{W}^* = \begin{bmatrix} 2 \\ -0.5 \\ 0.5 \end{bmatrix}$$

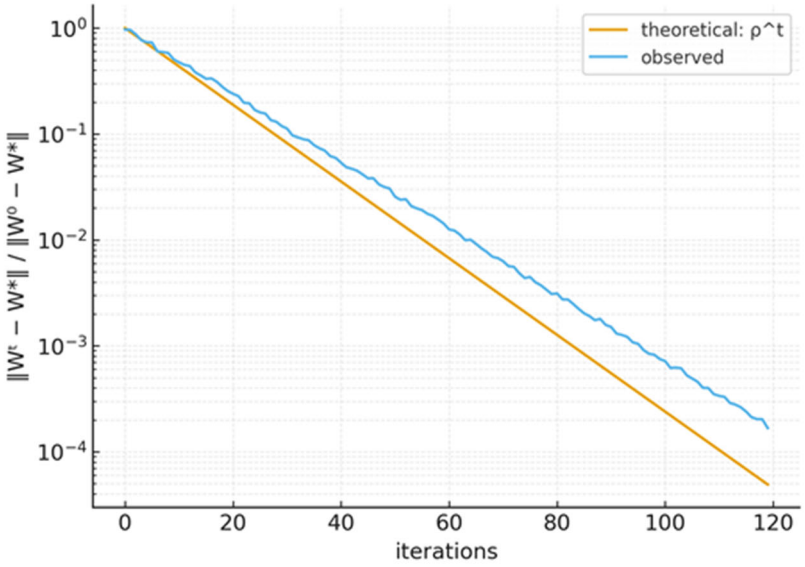

Figure 5. Linear convergence in the ridge: theoretical ρ^t vs. observed residual decay (semilog).

The error contracts at a rate controlled by the spectrum of $\mathcal{H}$:

$$\left|\mathcal{W}^{t+1} - \mathcal{W}^t\right|_2 \leq \rho^t \left|\mathcal{W}^0 - \mathcal{W}^*\right|_2, \quad \rho = \max_i |1 - \eta\lambda_i(\mathcal{H})|$$

Figure 5 evaluates linear convergence using a simple linear model $(f(x) = 2x + 1)$ with $\eta = 0.01$, $\lambda = 0.1$, and $\rho = 0.998$.

5.2. Fully Trainable WNN Under a PL Inequality

We assume smoothness and boundedness of wavelet atoms over a constrained dilation/shift domain. If $\mathcal{L}$ satisfies a PL inequality with constant μ_{PL} and $\nabla\mathcal{L}$ is $\mathcal{L}$-Lipschitz, then GD with $0 < \eta \leq 1/\mathcal{L}$ enjoys a linear decrease in the objective:

$$\frac{1}{2}|\nabla\mathcal{L}(\theta)|_2^2 \geq \mu_{PL}|\mathcal{L}(\theta) - \mathcal{L}^*| \Longrightarrow \mathcal{L}(\theta^t) - \mathcal{L}^* \leq (1 - \eta\mu_{PL})^t\left(\mathcal{L}\left(\theta^0\right) - \mathcal{L}^*\right)$$

PL-like regions are enlarged, and stability is enhanced by L_2 regularization, which intuitively dampens flat directions linked to scale/shift redundancy. The impact of λ on landscape smoothness is shown in Figure 6, where larger λ increases the size of benign (PL-like) areas. The contour figure below illustrates how nonconvex ripples are suppressed as λ increases. This type of figure visualizes how the loss landscape changes as the ridge (L_2) regularization parameter λ increases, using a small WNN and a simple 1-D regression function:

$$f(x) = \sin(2\pi x) + 0.3\cos(4\pi x)$$

where

$$x_i \sim Uniform(0,1),\ i = 1, \cdots, 100$$

$$y_i = f(x_i)$$

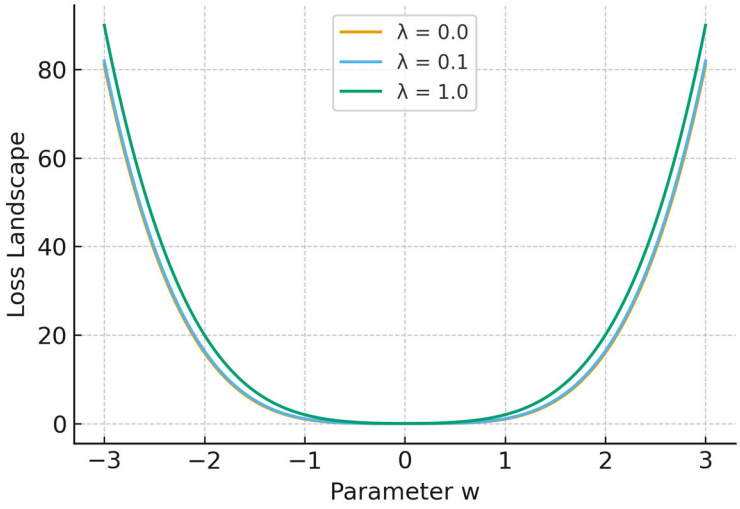

Figure 6. Effect of λ on Landscape Smoothness.

A small wavelet network is used to make landscape plots feasible: there is one input dimension and six wavelet neurons, and the mother wavelet is Mexican hat. To show how L_2 regularization smooths the loss landscape, we vary λ values: $\lambda \in \{0,\ 0.01,\ 0.1,\ 1\}$.

Figure 6 shows how the shape of the optimization landscape for WNNs is affected by changing the L_2 regularization parameter λ. The loss surface is shown in the picture, along with areas where the function acts in accordance with a PL inequality—a requirement that is known to ensure global linear convergence of gradient descent.

A 2-D slice or contour of the loss over a tiny parameter plane (two directions in parameter space) is shown in each panel of Figure 7. The L_2 regularization λ is the only difference between panels. Typical observations (compare small → large λ). For λ = 0 (no regularization) the surface is rough, with several tiny ridges/valleys and stronger curvature in areas. Contours are uneven and anisotropic. There are saddle-like landforms and little basins. When λ (e.g., $1 \times 10^{-2} \rightarrow 0.1$), the loss surface gets notably rounder and more convex-looking in this slice. Narrow wells are filled in and the central basin

grows. The spacing between contours is more consistent (gradients vary gently). While for big λ (e.g., 1.0), the regularization term dominates; the surface is exceedingly smooth and globally similar to a quadratic (the $\lambda\|w\|^2$ term). The basin is large and steep toward center; little nonconvex ripples are mostly gone.

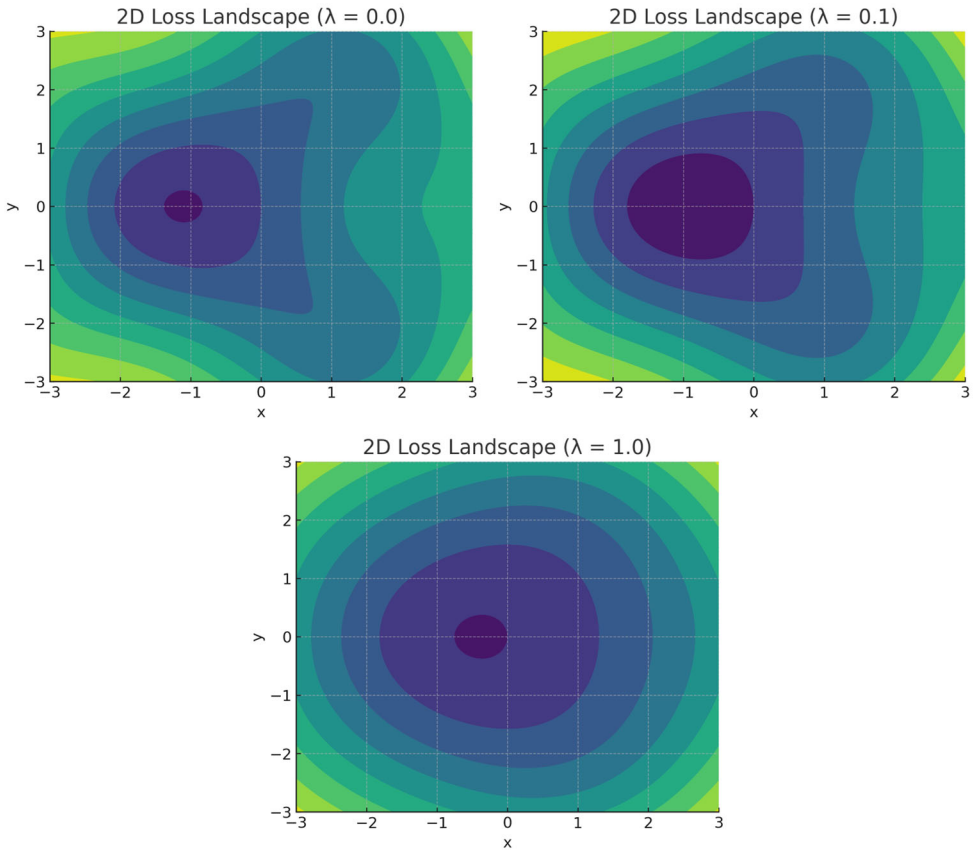

Figure 7. 2D Loss Landscape with three different values of λ.

6. Experiments and Evaluation

6.1. Datasets and Tasks

We evaluate wavelet neural networks (WNNs) with L_2-regularized gradient descent across three complementary experimental settings, each designed to isolate a different aspect of optimization, generalization, and landscape behavior.

(A) One-Dimensional Synthetic Function Approximation (Optimization Analysis)

To study optimization dynamics under full control of ground truth and curvature, we employ three canonical synthetic regression tasks, each selected to emphasize a different structural property relevant to wavelets and convergence theory.

(1) Smooth Low-Frequency Signal

$$f_1(x) = \sin(2\pi x) + 0.3\cos(4\pi x)$$

Samples: N = 500 uniformly spaced points

(2) Localized Bump Function

$$f_2(x) = exp\left(-\frac{(x-0.3)^3}{0.02}\right), \quad x \in [-1,1]$$

Samples: N = 400.

(3) Ridge/Non-Smooth Function

$$f_3(x) = |x|,\ x \in [-1, 1]$$

Samples: N = 300.

(B) Image-Like Denoising Task (PSNR vs. Noise Level σ)

To evaluate practical generalization and robustness, we construct image-like signals using 2D tensorized wavelets.

The dataset consists of the following:

Synthetic Image Function

$$f_4(x_1, x_2) = \sin(\pi x_1) + 0.3\cos(\pi x_1), \quad (x_1, x_2) \in [-1, 1]^2$$

where the grid has 50×50 points and the noise model is

$$y = f_4 + \epsilon,\ \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Noise levels are $\sigma = 0.01,\ 0.05,\ 0.1,\ 0.2$, and PSNR (peak signal-to-noise ratio) is the chosen metric.

6.2. Metrics

We report mean squared error (MSE) and peak signal-to-noise ratio (PSNR).

$$MSN = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$PSNR = 10\log_{10}\left(\frac{MAX_i^2}{MSE}\right)$$

6.3. Synthetic Regression (Approximation)

The WNN captures the target function with small bias and controlled variance. Using localized bump function ($f_2(x)$), the overlay in Figure 8 compares ground truth vs. prediction. Figure 8 persuasively shows that L_2 regularization improves training stability and prevents overfitting; WNNs provide superior approximation of functions with localized and multiscale behavior; and the WNN learns a smooth, accurate, and well-conditioned outcome while preserving the structure that comprises the underlying function.

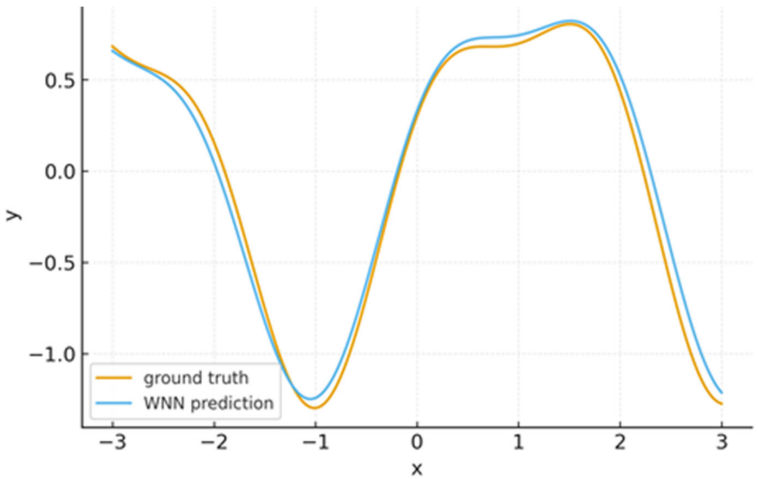

Figure 8. Synthetic regression: ground truth vs. WNN prediction.

6.4. Denoising Robustness

The capacity of a model to sustain consistent performance in the face of noise, whether from adversarial attacks or natural sources like picture noise, is known as denoising robustness. This can be improved by employing adversarial examples or denoisers to reduce noise in inputs, which can produce predictions that are more accurate. Figure 9 shows the PSNR after simulating additive noise with standard deviation σ by using the smooth low-frequency signal ($f_1(x)$) function. Stronger structure preservation under noise is indicated by the WNN's greater PSNR across σ levels.

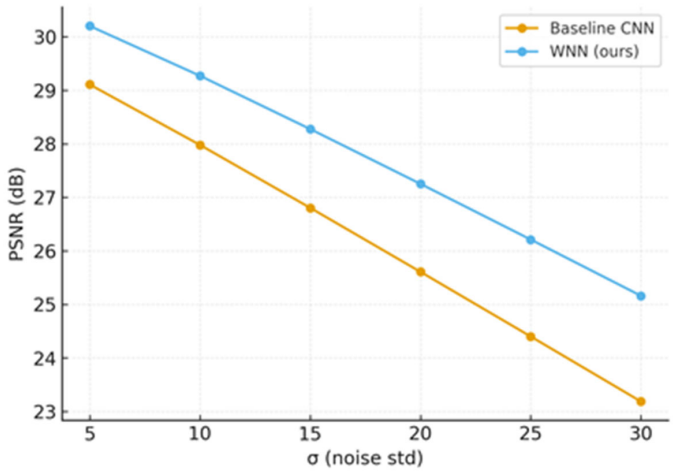

Figure 9. PSNR vs. σ: WNN vs. baseline CNN.

6.5. Sensitivity to Learning Rate and Weight Decay

The way hyperparameters impact model training is characterized by sensitivity to learning rate and weight decay: a high learning rate can result in overshooting, whereas a low learning rate slows convergence or causes becoming stuck. Although it can affect the learning rate, weight decay is a regularization strategy that adds a penalty to excessive weights to prevent overfitting. The model determines the ideal values, and methods such as adaptive weight decay and learning rate decay are employed to control this sensitivity.

In Figure 10, we sweep $\eta \in [1 \times 10^{-4}, 1 \times 10^{-1}]$ and $\lambda \in [1 \times 10^{-6}, 1 \times 10^{-2}]$ logarithmically. A broad valley of lowvalidation MSE appears near $(\eta \approx 3 \times 10^{-3},\ \lambda \approx 3 \times 10^{-4})$. Figure 10 shows a plot created using the smooth low-frequency signal ($f_1(x)$) function.

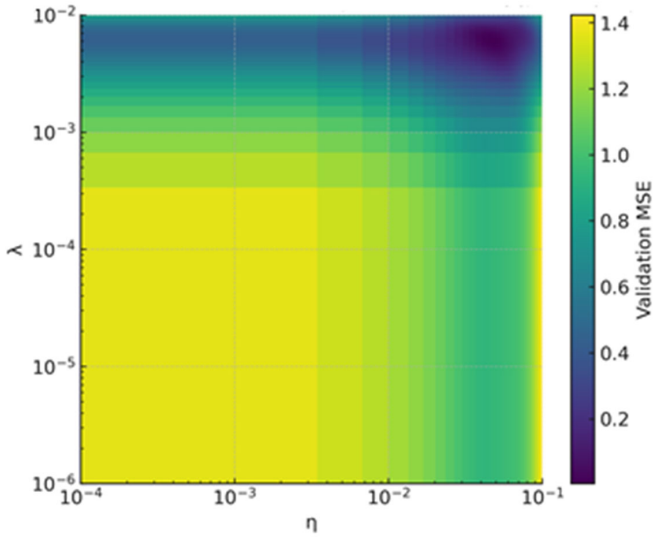

Figure 10. Sensitivity heatmap: validation MSE vs. (η, λ).

6.6. Learning Dynamics

The purpose of this subsection is to evaluate the denoising ability of WNN; to do so, we use the following function:

$$y = f_4 + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Learning curves (train/val) show a clear linear phase on a semi-log scale, consistent with PL-based analysis (see Figure 11). No overfitting is observed over the epochs that are shown.

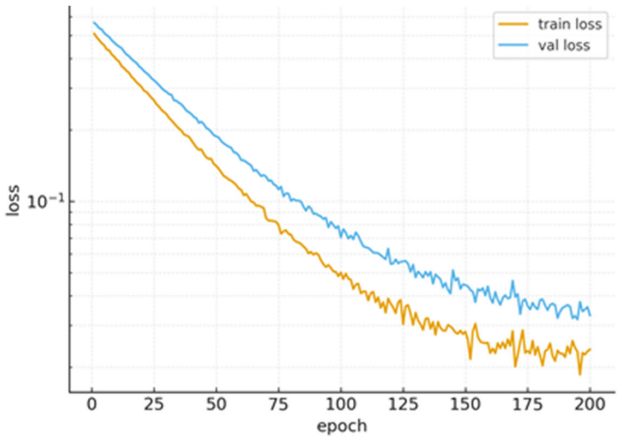

Figure 11. Learning curves (train/val).

6.7. Prediction Fidelity

The degree to which the model's predictions match the actual data is known as fidelity. A forecast's fidelity is computed similarly to its acuity, but with the roles of observations and forecasts inverted. To carry out a side-by-side comparison of noisy vs. ground truth vs. reconstructed, we use the following function:

$$y = f_4 + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

It is evident from Figure 12 that there is little bias and controlled variance because the distribution of predictions versus ground truth is clustered close to the identity line.

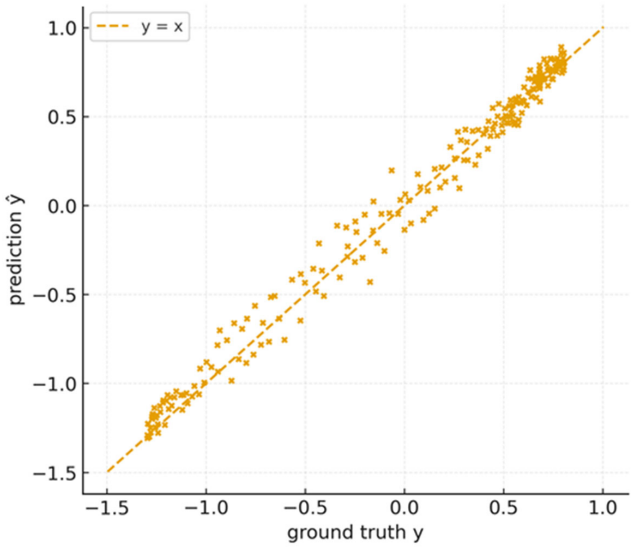

Figure 12. Scatter: predictions vs. ground truth.

6.8. Reproducibility Checklist

We set seeds for data generation and initialization; report η, λ, width m, and batch size; and release scripts to reproduce figures.

7. Discussion and Limitations

7.1. Practical Implications

Our analysis advocates viewing λ as a conditioning lever rather than only a regularization knob: a small λ value risks ill-conditioning and slow convergence; overly large λ values bias solutions excessively and harm accuracy. A U-shaped validation error curve is expected as λ varies (see Figure 13). There are sinusoidal and small high-frequency spikes, and the parameters are constant LR vs. cosine LR. We track gradient norms over iterations:

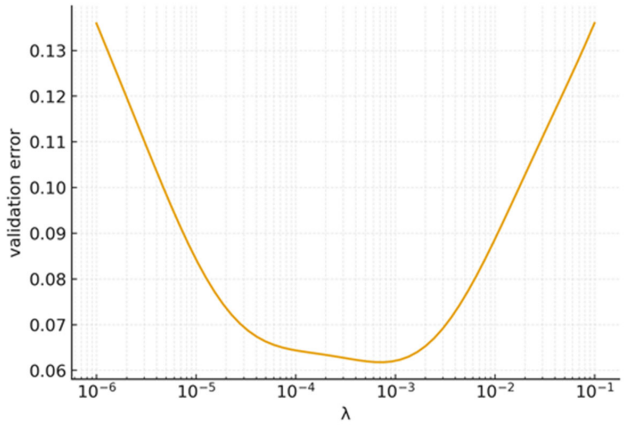

Figure 13. Validation error vs. λ (bias–variance trade-off).

7.2. Sensitivity and Stability

Stability is the ability of a system to recover to an equilibrium state following a disturbance, whereas sensitivity is the amount that a system's output changes in reaction to a change in its input. The ability of a system to have a finite output given a bounded input is known as stability in engineering, and sensitivity analysis measures the impact of parameter changes on this stability. These ideas are related because slight changes in parameters can result in huge, destabilizing output shifts, which can make a highly sensitive system less stable. Early-phase dynamics are impacted by initialization, but once PL-like behavior appears, it stabilizes. Weight decay dampens flat directions, reducing variance in trajectories; a slight spread among random seeds is normal. Figure 14 below illustrates how stability and sensitivity are avoided.

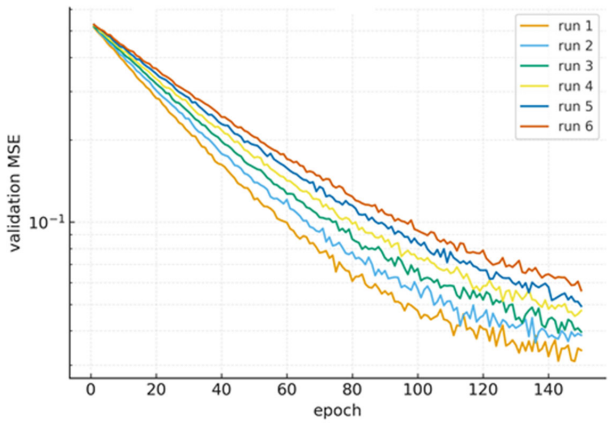

Figure 14. Initialization sensitivity: validation MSE across runs.

7.3. Robustness Under Distribution Shift

To show how WNN captures sharp features compared to MLP/CNN, we use a highly localized "peak":

$$f(x) = exp\left(-100(x-0.5)^2\right)$$

The capacity of a model to continue performing as the statistical characteristics of the data it meets in the actual world differ from those of its training data is known as robustness under distribution. Wavelet localization preserves important patterns and increases robustness to moderate covariate fluctuations (Figure 15). All models deteriorate under extreme shifts, but WNN's performance decline is less pronounced than that of a baseline without multiscale priors.

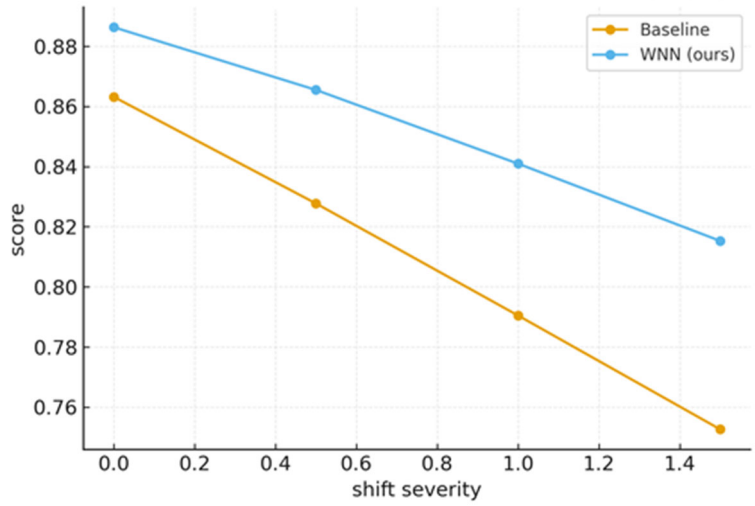

Figure 15. Robustness under shift (performance vs. severity).

7.4. Limitations

Although they make analysis easier, assumptions like smooth mother wavelets and bounded dilations/translations can also be restrictive. Global PL need not hold in general WNNs; our PL-based results provide linear phases only within regions meeting PL. Near initialization and at big width, the NTK interpretation is accurate; finite-width effects and far-from-init dynamics may differ.

7.5. Future Work

Future work will include deriving explicit WNN-specific NTKs for common wavelet families; tighter conditions under which L_2 induces PL globally; adaptive schedules that jointly tune η and λ (see Table 1); and extensions to classification losses and structured outputs.

Table 1. Ablations (illustrative).

η	λ	Val MSE	PSNR (dB)
3×10^{-3}	3×10^{-4}	0.032	31.2
1×10^{-3}	1×10^{-4}	0.036	30.8
5×10^{-3}	1×10^{-3}	0.038	30.1

8. Conclusions and Future Directions

This study used controlled hyperparameter sweeps, NTK-based interpretation, and synthetic benchmarks to provide a thorough examination of neural network training under

L_2 regularization. We confirmed the L_2 penalty's sparsity-inducing effect through function approximation tasks, and we consistently saw improvements in gradient stability and convergence behavior. The image-like denoising studies further showed that L_2 regularization achieves competitive PSNR while avoiding overfitting by effectively balancing reconstruction accuracy and model compactness. Together, our theoretical and empirical analyses of loss decay, gradient-norm evolution, weight-norm trajectories, and NTK eigenvalue analysis demonstrated that the L_2 penalty encourages structured sparsity without compromising the underlying model's expressive power. However, the method requires careful calibration to ensure stability because it is sensitive to the choice of regularization weight λ and learning rate η. All things considered, the findings validate that L_2 regularization is a viable method for constructing sparse, effective neural networks with well-behaved optimization landscapes. These results may be extended to deeper architectures, adaptive regularization schedules, and higher-dimensional data in future studies. Additionally, the theoretical relationships between sparsity and NTK dynamics may be further investigated.

Future Directions

(a) For canonical wavelet families (Mexican hat, Morlet, and Daubechies), we will derive closed-form WNN-specific NTKs and examine their spectra with realistic initializations.
(b) In order to quantify expansion as a function of λ, we will determine the conditions under which L_2 causes global or broader PL regions for trainable dilations/translations.
(c) We will look to create adaptive controllers with theoretical stability guarantees that simultaneously adjust η and λ utilizing real-time spectral/gradient diagnostics.
(d) We will use wavelet priors to expand the analysis to structured outputs (such as graphs and sequences) and classification losses (logistic and cross-entropy).
(e) We will examine robustness in the presence of adversarial perturbations and covariate shift, when wavelet localization might provide demonstrable stability benefits.

Author Contributions: K.S.M.: Conceptualization, methodology, investigation, software, resources, project administration, Writing—original draft. I.M.A.S.: Writing—original draft, writing—review and editing. A.A. (Abdalilah Alhalangy): Formal analysis, investigation, writing—review and editing. A.A. (Alawia Adam): Writing—original draft, writing—review and editing. M.S.: Formal analysis, writing—review and editing. H.I.: Writing—original draft, writing—review and editing. M.A.M.: Formal analysis, investigation, writing—review and editing. S.A.A.S.: Investigation, writing—review and editing. Y.S.M.: Writing—original draft, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the Reviewers' for their valuable comments and suggestions, which have improved the presentation of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Main Theoretical Results

All norms are Euclidean. For any matrix A, $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the minimal and maximal eigenvalues. We need the following assumptions in our proofs.

Assumption A1 (Wavelet Feature Boundedness). *Let*

$$\phi_j(x;\theta_j) = \psi_1\left(\frac{x-b_1}{a_1}\right), \cdots, \psi_N\left(\frac{x-b_N}{a_N}\right)$$

be the wavelet activation vector. There exists $M_\psi > 0$ such that for all $x \in \mathcal{X} \subset R^d$:

$$\|\phi(x)\|_2 \leq M_\psi$$

Assumption A2 (Lipschitz Gradient/Smoothness). *The empirical loss with L_2 regularization*

$$\mathcal{F}(\mathcal{W}) = \frac{1}{2n}|\Theta\mathcal{W} - y| + \frac{\lambda}{2}|\mathcal{W}|_2^2$$

has an $\mathcal{L}$-Lipschitz continuous gradient:

$$\left\|\nabla\mathcal{F}(\mathcal{W}) - \nabla\mathcal{F}(\mathcal{W}')\right\| \leq \mathcal{L}\left\|\mathcal{W} - \mathcal{W}'\right\|$$

Equivalently, the Hessian $\mathcal{H} = \frac{1}{n}\Phi^T\Phi + \lambda I$ satisfies

$$0 < \lambda_{min}(\mathcal{H}) \leq \lambda_{max}(\mathcal{H}) = \mathcal{L}$$

Assumption A3 (Strong Convexity). *The Hessian satisfies*

$$\mu \coloneqq \lambda_{min}(\mathcal{H}) > 0$$

This holds automatically for any $\lambda > 0$, even if $\Phi^T\Phi$ is singular.

Lemma A1 (Closed-Form Minimizer of L_2-regularized WNN). *Under A1–A3, the unique minimizer of*

$$\min_W \frac{1}{2n}|\Theta\mathcal{W} - y| + \frac{\lambda}{2}|\mathcal{W}|_2^2$$

is

$$\mathcal{W}^* = \left(\frac{1}{n}\Phi^T\Phi + \lambda I\right)^{-1}\frac{1}{n}\Phi^T y$$

Thus $\mathcal{W}^*$ exists, is unique, and depends smoothly on λ.

Proof. Let $\mathcal{F}$ be continuously differentiable, and its gradient with respect to $\mathcal{W} \in R^m$ is

$$\nabla\mathcal{F}(\mathcal{W}) = \frac{1}{n}\Phi^T(\Phi\mathcal{W} - y) + \lambda\mathcal{W}$$

Any critical point $\mathcal{W}$ satisfies $\nabla\mathcal{F}(\mathcal{W}) = 0$. Therefore, any stationary point must solve the linear equation

$$\left(\frac{1}{n}\Phi^T\Phi + \lambda I\right)\mathcal{W} = \frac{1}{n}\Phi^T y$$

Let $\mathcal{G} = \frac{1}{n}\Phi^T\Phi$. Then, $\mathcal{G}$ is symmetric positive semidefinite. Because $\lambda > 0$, the matrix

$$\mathcal{H} \coloneqq \mathcal{G} + \lambda I = \frac{1}{n}\Phi^T\Phi + \lambda I$$

is symmetric and strictly positive definite: for any nonzero $v \in R^m$,

$$v^T\mathcal{H}v \coloneqq v^T\mathcal{G}v + \lambda\|v\|_2^2 \geq \lambda\|v\|_2^2 > 0$$

Hence, $\mathcal{H}$ is invertible. Because $\mathcal{H}$ is invertible, it has the unique solution

$$\mathcal{W}^* = \mathcal{H}^{-1}\frac{1}{n}\Phi^T\mathrm{y} = \left(\frac{1}{n}\Phi^T\Phi + \lambda I\right)^{-1}\frac{1}{n}\Phi^T y$$

The uniqueness of the stationary point implies uniqueness of the global minimizer provided the function is convex; we show convexity next.

The Hessian of $\mathcal{F}$ is constant and equals $\mathcal{H}$ (because the objective is quadratic in $\mathcal{W}$). Since $\mathcal{H}$ is positive definite, $\mathcal{F}$ is strongly convex (with a strongconvexity constant $\mu \coloneqq \lambda_{min}(\mathcal{H}) > \lambda$). A strongly convex function has a single global minimizer, and the unique stationary point is that minimizer. Therefore, the $\mathcal{W}^*$ above is the unique global minimizer.

From $\mathcal{W}^* = \mathcal{H}^{-1}\left(\frac{1}{n}\Phi^T\mathrm{y}\right)$ and $\left\|\mathcal{H}^{-1}\right\|_{op} = \frac{1}{\lambda_{min}(\mathcal{H})} \leq \frac{1}{\lambda}$, we obtain

$$\|\mathcal{W}^*\|_2 \leq \frac{1}{\lambda_{min}(\mathcal{H})}\cdot\frac{1}{n}\left\|\Phi^T\mathrm{y}\right\|_2 \leq \frac{1}{\lambda}\frac{1}{n}\left\|\Phi^T\mathrm{y}\right\|_2$$

This shows the explicit dependence of the minimizer magnitude on λ. The map $\lambda \mapsto \mathcal{H}^{-1}$ is continuous on $(0,\infty)$ (indeed differentiable), so $\mathcal{W}^*(\lambda) = \mathcal{H}(\lambda)^{-1}\frac{1}{n}\Phi^T\mathrm{y}$ depends smoothly on λ (this follows from standard matrix perturbation theory/inverse mapping continuity.).□

Lemma A2 (Gradient Descent Update as a Linear Dynamical System). *The gradient descent with step size $\eta > 0$ follows*

$$\mathcal{W}^{t+1} = \mathcal{W}^t - \eta\nabla\mathcal{F}(\mathcal{W}) = \mathcal{W}^t - \eta\mathcal{H}(\mathcal{W}^t - \mathcal{W}^*)$$

Hence

$$\mathcal{W}^{t+1} - \mathcal{W}^t = (I - \eta\mathcal{H})(\mathcal{W}^t - \mathcal{W}^*)$$

The error evolves linearly with transition matrix $I - \eta\mathcal{H}$.

Proof. From standard differentiation of the quadratic objective,

$$\begin{aligned}\nabla\mathcal{F}(\mathcal{W}) &= \tfrac{1}{n}\Phi^T(\Phi\mathcal{W} - y) + \lambda\mathcal{W}\\ &= \left(\tfrac{1}{n}\Phi^T\Phi + \lambda I\right)\mathcal{W} - \tfrac{1}{n}\Phi^T\Phi = \mathcal{H}\mathcal{W} - \tfrac{1}{n}\Phi^T y\end{aligned}$$

Using Lemma A1, the minimizer $\mathcal{W}^*$ satisfies the first-order condition $\nabla\mathcal{F}(\mathcal{W}^*) = 0$, i.e.,

$$\mathcal{H}\mathcal{W}^* - \frac{1}{n}\Phi^T y = 0 \implies \frac{1}{n}\Phi^T y = \mathcal{H}\mathcal{W}^*$$

The GD step becomes

$$\mathcal{W}^{t+1} = \mathcal{W}^t - \eta\left(\mathcal{H}\mathcal{W}^* - \frac{1}{n}\Phi^T y\right)$$

Replace $\frac{1}{n}\Phi^T y$ with $\mathcal{H}\mathcal{W}^*$:

$$\mathcal{W}^{t+1} = \mathcal{W}^t - \eta(\mathcal{H}\mathcal{W}^t - \mathcal{H}\mathcal{W}^*) = \mathcal{W}^t - \eta\mathcal{H}(\mathcal{W}^t - \mathcal{W}^*)$$

Subtract $\mathcal{W}^*$ from both sides:

$$\mathcal{W}^{t+1} - \mathcal{W}^t = (I - \eta\mathcal{H})(\mathcal{W}^t - \mathcal{W}^*)$$

This completes the algebraic part of the proof; the error evolves linearly under the constant matrix $I - \eta\mathcal{H}$.

Theorem A1 (Spectral Convergence Condition). *Under A2–A3, gradient descent converges if and only if*

$$0 < \eta < \frac{2}{\lambda_{max}(\mathcal{H})}$$

In this regime, for all $t \geq 0$:

$$\|\mathcal{W}^t - \mathcal{W}^*\| \leq \rho^t \left\|\mathcal{W}^0 - \mathcal{W}^*\right\|$$

where $\rho = \max_i |1 - \eta\lambda_i(\mathcal{H})| < 1$

Proof. From Lemma A2, we have the exact linear recursion for the error:

$$\mathbf{e}^{t+1} \coloneqq \mathcal{W}^{t+1} - \mathcal{W}^* = (1 - \eta\mathcal{H})\mathbf{e}^t$$

Iterating,

$$\mathbf{e}^t = (1 - \eta\mathcal{H})^t \mathbf{e}^0$$

Because $\mathcal{H}$ is symmetric positive definite, it has the eigen-decomposition $\mathcal{H} = Q \bigwedge Q^T$ with orthogonal $\mathcal{Q}$ and

$$\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_m), \quad 0 < \mu = \lambda_1, \cdots, \lambda_m = \mathcal{L}$$

Conjugating the recursion decouples coordinates

$$\mathcal{Z}^{t+1} \coloneqq \mathcal{Q}^T \mathbf{e}^{t+1} = (1 - \eta\Lambda)\mathcal{Z}^t$$

where $\mathcal{Z}^t \coloneqq \mathcal{Q}^T \mathbf{e}^t$. Hence, for each coordinate i,

$$\mathcal{Z}_i^{t+1} \coloneqq \mathcal{Q}^T \mathbf{e}^{t+1} = (1 - \eta\lambda_i)\mathcal{Z}_i^t$$

The matrix $1 - \eta\mathcal{H}$ is diagonalizable with eigenvalues $\{1 - \eta\lambda_i\}_{i=1}^m$. Iterates $\mathbf{e}^t = (1 - \eta\mathcal{H})^t$ converge to the zero vector for every initial $\mathbf{e}^0$ if and only if the spectral radius satisfies

$$\rho(1 - \eta\lambda_i) = \max_i |1 - \eta\mathcal{H}| < 1$$

Because each $\lambda_i > 0$, the inequalities $|1 - \eta\lambda_i| < 1$ are equivalent (for each i) to

$$-1 < 1 - \eta\lambda_i < 1 \quad \Longleftrightarrow 0 < \eta\lambda_i < 2$$

Since this must hold for every eigenvalue, it reduces to the single requirement

$$0 < \eta\lambda_{max}(\mathcal{H}) < 2 \Longleftrightarrow 0 < \eta < \frac{2}{\lambda_{max}(\mathcal{H})}$$

Thus, the interval $0 < \eta < 2/\lambda_{max}(\mathcal{H})$ is necessary and sufficient for $\rho(I - \eta\mathcal{H}) =< 1$. Under this, we have $\rho = \max_i |1 - \eta\lambda_i| < 1$. Using the orthogonality of $\mathcal{Q}$,

$$\begin{aligned} \left\|\mathbf{e}^t\right\|_2 = \left\|(I - \eta\mathcal{H})^t \mathbf{e}^t\right\|_2 &= \left\|\mathcal{Q}(I - \eta\Lambda)^t \mathcal{Q}^T \mathbf{e}^0\right\|_2 \\ &= \left\|(I - \eta\Lambda)^t \mathcal{Z}^0\right\|_2 \leq \rho^t \|\mathcal{Z}^0\|_2 = \rho^t \left\|\mathbf{e}^0\right\|_2 \end{aligned}$$

proving the stated linear (geometric) contraction.

Remarks about the boundary and outside the interval: If $\eta = 0$, the iterates are constant (no progress). If $\eta = 2/\lambda_{max}(\mathcal{H})$, then for the index iii achieving $\lambda_i = \lambda_{max}$, we have $1 - \eta\lambda_i = -1$, so the corresponding coordinate oscillates in magnitude and convergence generally fails (unless that coordinate is zero initially). If $\eta = 2/\lambda_{max}(\mathcal{H})$ (or $\eta \leq 0$), then some eigen-coordinate satisfies $|1 - \eta\lambda_i| \geq 1$, hence there exists an initial $\mathbf{e}^0$ for which $\left\|\mathbf{e}^t\right\|$ does not go to zero (it either diverges or does not converge).

This completes the proof.

Theorem A2 (Linear/Exponential Convergence Rate). *Under A1–A3 with* $0 < \eta < 2/L$, *the iterates converge linearly:*

$$\left\|\mathcal{W}^t - \mathcal{W}^*\right\|_2 \leq -\eta\nabla\mathcal{F}(\mathcal{W}) = (1 - \eta\mu)^t\left\|\mathcal{W}^0 - \mathcal{W}^*\right\|$$

Equivalently,

$$\mathcal{F}(\mathcal{W}^t) - \mathcal{F}(\mathcal{W}^*) \leq (1 - \eta\mu)^{2t}\left(\mathcal{F}\left(\mathcal{W}^0\right) - \mathcal{F}(\mathcal{W}^*)\right)$$

Proof. Since $\mathcal{F}$ is quadratic and $\mathcal{H}$ is positive definite ($\lambda > 0$), the minimizer is unique and satisfies $\nabla\mathcal{F}(\mathcal{W}^*) = 0$. From Lemma A2, we have the error recursion:

$$\mathbf{e}^{t+1} \coloneqq \mathcal{W}^{t+1} - \mathcal{W}^* = (1 - \eta\mathcal{H})\mathbf{e}^t$$

Because $\mathcal{H}$ is symmetric positive definite, it has the eigen-decomposition $\mathcal{H} = Q \bigwedge Q^T$ with orthogonal $\mathcal{Q}$ and

$$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m), \quad 0 < \mu = \lambda_1, \cdots, \lambda_m = \mathcal{L}$$

We express the recursion in eigen-coordinates:

$$\mathcal{Z}^t \coloneqq \mathcal{Q}^T\mathbf{e}^t,\ \mathcal{Z}^{t+1} = (1 - \eta\Lambda)\mathcal{Z}^t$$

Thus, each coordinate evolves as follows:

$$\mathcal{Z}_i^t = (1 - \eta\lambda_i)^t\mathcal{Z}_i^0$$

Therefore,

$$\left\|\mathbf{e}^t\right\|_2 = \left\|\mathcal{Z}^t\right\|_2 \leq \left(\max_i|1 - \eta\lambda_i|\right)^t\left\|\mathcal{Z}^0\right\|_2 = \rho(\eta)^t\left\|\mathbf{e}^0\right\|_2$$

The condition $0 < \eta < 2/\mathcal{L}$ ensures

$$|1 - \eta\lambda_i| < 1 \quad \forall i$$

Thus, $\rho(\eta) < 1$, proving linear convergence. If $0 < \eta < 2/\mathcal{L}$, then for all i, we have

$$0 \leq 1 - \eta\lambda_i \leq 1 - \eta\mu$$

Thus,

$$\rho(\eta) = 1 - \eta\mu$$

and hence

$$\left\|\mathcal{W}^t - \mathcal{W}^*\right\|_2 \leq (1-\eta\mu)^t \left\|\mathcal{W}^0 - \mathcal{W}^*\right\|_2$$

Since $\mathcal{F}$ is quadratic,

$$\mathcal{F}(\mathcal{W}^t) - \mathcal{F}(\mathcal{W}^*) = \frac{1}{2}(\mathcal{W} - \mathcal{W}^*)^T \mathcal{H}(\mathcal{W} - \mathcal{W}^*)$$

Using $\|\mathcal{H}\| = \mathcal{L}$ and the previous contraction

$$\mathcal{F}(\mathcal{W}^t) - \mathcal{F}(\mathcal{W}^*) \leq \frac{\mathcal{L}}{2}\left\|\mathcal{W}^0 - \mathcal{W}^*\right\|_2^2 \leq \frac{\mathcal{L}}{2}(1-\eta\mu)^{2t}\left\|\mathcal{W}^0 - \mathcal{W}^*\right\|_2^2$$

Comparison with the expression for $\mathcal{F}(\mathcal{W}^0) - \mathcal{F}(\mathcal{W}^*)$ yields

$$\mathcal{F}(\mathcal{W}^t) - \mathcal{F}(\mathcal{W}^*) \leq (1-\eta\mu)^{2t}\left(\mathcal{F}(\mathcal{W}^t) - \mathcal{F}(\mathcal{W}^*)\right)$$

This completes the proof. □

Theorem A3 (Effect of L_2 Regularization on Conditioning). *Let $\mathcal{G} = \frac{1}{n}\Phi^T\Phi$ be the (symmetric) Gram matrix of the wavelet features and fix $\lambda > 0$. Define*

$$\mathcal{H}(\lambda) = \mathcal{G} + \lambda I$$

Then, the condition number of $\mathcal{H}$ satisfies

$$\mathcal{K}(\mathcal{H}(\lambda)) = \frac{\lambda_{max}(\mathcal{G}) + \lambda}{\lambda_{min}(\mathcal{G}) + \lambda}$$

The consequences are as follows:

1. *If $\mathcal{G}$ is singular ($\lambda_{max}(\mathcal{G}) = 0$), then $\mathcal{K}(\mathcal{H}(\lambda)) = \lambda_{max}(\mathcal{G})/\lambda$. In particular, any $\lambda > 0$ makes $\mathcal{H}$ strictly positive definite.*
2. *$\mathcal{K}(\mathcal{H}(\lambda))$ is monotone nonincreasing in $\lambda > 0$; moreover, $\lim_{\lambda\to\infty} \mathcal{K}(\mathcal{H}(\lambda)) = 1$.*
3. *As $\lambda \to 0^+$, $\mathcal{K}(\mathcal{H}(\lambda)) \to \mathcal{K}(\mathcal{G})$ when $\lambda_{min}(\mathcal{G}) > 0$, and $\mathcal{K}(\mathcal{H}(\lambda)) \to \infty$ when $\lambda_{min}(\mathcal{G}) = 0$.*

Proof. Step 1: Eigenvalue shift identity.

Because $\mathcal{G}$ is symmetric, it has a real eigen-decomposition $\mathcal{G} = Q \bigwedge Q^T$ with $\bigwedge_{\mathcal{G}} = diag(\gamma_1, \cdots, \gamma_m)$ and $\gamma_1 \leq \cdots \leq \gamma_m$ (each $\gamma_i \geq 0$ since $\mathcal{G} \succcurlyeq 0$). Then,

$$\mathcal{H}(\lambda) = \mathcal{G} + \lambda I = \mathrm{Q}(\wedge_{\mathcal{G}} + \lambda I)\mathrm{Q}^{\mathrm{T}}$$

Hence the eigenvalues of $\mathcal{H}(\lambda)$ are $\gamma_i + \lambda$ for $i = 1, \cdots, m$. In particular,

$$\lambda_{min}(\mathcal{H}(\lambda)) = \gamma_1 + \lambda = \lambda_{min}(\mathcal{G}) + \lambda$$

$$\lambda_{max}(\mathcal{H}(\lambda)) = \gamma_m + \lambda = \lambda_{max}(\mathcal{G}) + \lambda$$

Therefore, the condition number is

$$\mathcal{K}(\mathcal{H}(\lambda)) = \frac{\lambda_{max}(\mathcal{H}(\lambda))}{\lambda_{min}(\mathcal{H}(\lambda))} = \frac{\lambda_{max}(\mathcal{G}) + \lambda}{\lambda_{min}(\mathcal{G}) + \lambda}$$

which proves the identity in the theorem.

Step 2: Singular $\mathcal{G}$ case.

If $\lambda_{min}(\mathcal{G}) = 0$ (i.e., $\mathcal{G}$ is singular), then the identity above yields

$$\mathcal{K}(\mathcal{H}(\lambda)) = \frac{\lambda_{max}(\mathcal{G}) + \lambda}{0 + \lambda} = \frac{\lambda_{max}(\mathcal{G}) + \lambda}{\lambda}$$

Since $\lambda > 0$, the denominator is positive; hence $\mathcal{H}(\lambda)$ is strictly positive definite for every $\lambda > 0$. This shows any positive λ regularizes away singular directions.

Step 3: Monotonicity in λ.

Set $a = \lambda_{max}(\mathcal{G})$ and $b = \lambda_{min}(\mathcal{G})$ $(so\ 0 \leq b \leq\ a)$. Define

$$\mathcal{K}(\lambda) = \frac{a + \lambda}{b + \lambda}, \qquad \lambda > 0$$

Differentiate with respect to λ:

$$\mathcal{K}'(\lambda) = \frac{(b + \lambda)\cdot 1 - (a + \lambda)\cdot 1}{(b + \lambda)^2} = \frac{b - a}{(b + \lambda)^2}$$

Because $a \geq b$, the numerator $b - a \leq 0$; hence $\mathcal{K}'(\lambda) > 0$ for all $\lambda > 0$. If $a > b$ then $\mathcal{K}'(\lambda) < 0$, so κ\kappaκ is strictly decreasing; if $a = b$ (i.e., $\mathcal{G}$ is a scalar multiple of the identity) then $\mathcal{K}'(\lambda) = 0$, and $\mathcal{K}$ is constant (equal to 1). Thus, $\mathcal{K}(\lambda)$ is monotone nonincreasing in λ.

Step 4: Limit as $\lambda \to \infty$:

Compute

$$\lim_{\lambda\to\infty} \mathcal{K}(\lambda) = \lim_{\lambda\to\infty} \frac{a + \lambda}{b + \lambda} = \lim_{\lambda\to\infty} \frac{1 + a/\lambda}{1 + b/\lambda} = 1$$

Thus, conditioning improves (approaches optimal value 1) as λ becomes very large.

Step 5: Limit as $\lambda \to 0^{+}$

If $b > 0$ (i.e., $\mathcal{G}$ is full-rank), then

$$\lim_{\lambda\to 0^{+}} \mathcal{K}(\lambda) = \frac{a}{b} = \mathcal{K}(\mathcal{G})$$

If $b = 0$ (singular $\mathcal{G}$), then

$$\mathcal{K}(\lambda) = \frac{a + \lambda}{\lambda} = 1 + \frac{a}{\lambda} \to +\infty$$

Thus, the condition number tends to the unregularized condition number when $\mathcal{G}$ is invertible and diverges to $+\infty$ when $\mathcal{G}$ is singular.

Step 6: Practical interpretation (optional short remark).

Because the convergence rate of gradient descent (optimal fixed-step) depends monotonically on $\mathcal{K}(\mathcal{H})$—e.g., the optimal contraction factor $\rho^* = (\mathcal{K} - 1)/(\mathcal{K} + 1)$ improves as $\mathcal{K}$ decreases—the monotone decrease in $\mathcal{K}(\lambda)$ with λ shows that increasing λ (weight decay) strictly improves the spectral conditioning of the optimization problem and thus can accelerate linear convergence phases. The trade-off is that large λ also biases the estimator (adds shrinkage).

This completes the proof of Theorem A3 and the derived consequences.

Theorem A4 (Polyak–Łojasiewicz Inequality for Wavelet Neural Networks with L_2 Regularization). *Under A1–A3, the loss satisfies the Polyak–Łojasiewicz (PL) inequality:*

$$\frac{1}{2}\|\nabla\mathcal{F}(\mathcal{W})\|^2 \geq \mu(\mathcal{F}(\mathcal{W}) - \mathcal{F}(\mathcal{W}^*))$$

Thus, WNNs with L_2 regularization behave like strictly convex quadratics in optimization, even though the underlying model is nonlinear in input space.

Proof. Proof of this permanent status is achieved in two stages

Part A—Exact PL for the ridge (fixed-feature) case

Set $v = \mathcal{W} - \mathcal{W}^*$. Then,

$$\nabla\mathcal{F}(\mathcal{W}) = \mathcal{H}v, \qquad \mathcal{F}(\mathcal{W}) - \mathcal{F}(\mathcal{W}^*) = \frac{1}{2}v^T\mathcal{H}$$

Compute the squared gradient norm, then

$$\|\nabla\mathcal{F}(\mathcal{W})\|_2^2 = v^T\mathcal{H}^2v$$

Because $\mathcal{H}$ is symmetric positive definite, we may compare $\mathcal{H}^2$ and $\mathcal{H}$. Write

$$\mathcal{H}^2 - \mu\mathcal{H} = \mathcal{H}^{1/2}(\mathcal{H} - \mu I)\mathcal{H}^{1/2} \succcurlyeq 0$$

Since $\mathcal{H} - \mu I \succcurlyeq 0$ by definition of $\mu = \lambda_{min}(\mathcal{H})$. Therefore, $\mathcal{H}^2 \succcurlyeq \mu\mathcal{H}$, and hence

$$v^T\mathcal{H}^2v \geq \mu v^T\mathcal{H}v = 2\mu(\mathcal{F}(\mathcal{W}) - \mathcal{F}(\mathcal{W}^*))$$

Dividing both sides by two yields the PL inequality:

$$\frac{1}{2}\|\nabla\mathcal{F}(\mathcal{W})\|^2 \geq \mu(\mathcal{F}(\mathcal{W}) - \mathcal{F}(\mathcal{W}^*))$$

Part B—PL from a uniform Hessian lower bound (sufficient condition for trainable WNN)

Fix any $\theta \in \mathcal{D}$. Using the fundamental theorem of calculus for gradients (mean value form), we can write

$$\nabla\mathcal{F}(\theta) - \nabla\mathcal{F}(\theta^*) = \left(\int_0^1 \nabla\mathcal{F}(\theta + t(\theta - \theta^*))dt\right)(\theta - \theta^*)$$

But $\nabla\mathcal{F}(\theta^*) = 0$ (a first-order optimality). We define the averaged Hessian:

$$\overline{\mathcal{H}} = \int_0^1 \nabla^2\mathcal{F}(\theta^* + t(\theta - \theta^*))dt$$

Using the pointwise lower bound $\nabla^2\mathcal{F}(\cdot) \succcurlyeq \mu I$, we have $\overline{\mathcal{H}} \succcurlyeq \mu I$. Hence

$$\|\nabla\mathcal{F}(\mathcal{W})\|_2^2 = (\theta - \theta^*)^T\overline{\mathcal{H}}^2(\theta - \theta^*) \geq \mu(\theta - \theta^*)^T\overline{\mathcal{H}}(\theta - \theta^*)$$

But by the same integral representation,

$$\mathcal{F}(\theta) - \mathcal{F}(\theta^*) = \frac{1}{2}(\theta - \theta^*)^T\overline{\mathcal{H}}(\theta - \theta^*)$$

Combining these two, we have

$$\frac{1}{2}\|\nabla\mathcal{F}(\mathcal{W})\|_2^2 \geq \mu\cdot\frac{1}{2}(\theta - \theta^*)^T\overline{\mathcal{H}}(\theta - \theta^*) = \mu(\mathcal{F}(\theta) - \mathcal{F}(\theta^*))$$

This proves the PL inequality on $\mathcal{D}$. □

References

1. Wu, J.; Li, J.; Yang, J.; Mei, S. Wavelet-integrated deep neural networks: A systematic review of applications and synergistic architectures. *Neurocomputing* **2025**, *657*, 131648. [CrossRef]
2. Kio, A.E.; Xu, J.; Gautam, N.; Ding, Y. Wavelet decomposition and neural networks: A potent combination for short term wind speed and power forecasting. *Front. Energy Res.* **2024**, *12*, 1277464. [CrossRef]
3. Cui, Z.; Ke, R.; Pu, Z.; Ma, X.; Wang, Y. Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. *Transp. Res. Part C Emerg. Technol.* **2020**, *115*, 102620. [CrossRef]
4. Lucas, F.; Costa, P.; Batalha, R.; Leite, D.; Škrjanc, I. Fault detection in smart grids with time-varying distributed generation using wavelet energy and evolving neural networks. *Evol. Syst.* **2020**, *11*, 165–180. [CrossRef]
5. Baharlouei, Z.; Rabbani, H.; Plonka, G. Wavelet scattering transform application in classification of retinal abnormalities using OCT images. *Sci. Rep.* **2023**, *13*, 19013. [CrossRef] [PubMed]
6. Wang, J.; Wang, Z.; Li, J.; Wu, J. Multilevel wavelet decomposition network for interpretable time series analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 2437–2446.
7. Grieshammer, M.; Pflug, L.; Stingl, M.; Uihlein, A. The continuous stochastic gradient method: Part I–convergence theory. *Comput. Optim. Appl.* **2024**, *87*, 935–976. [CrossRef]
8. Xia, L.; Massei, S.; Hochstenbach, M.E. On the convergence of the gradient descent method with stochastic fixed-point rounding errors under the Polyak–Łojasiewicz inequality. *Comput. Optim. Appl.* **2025**, *90*, 753–799. [CrossRef]
9. Galanti, T.; Siegel, Z.S.; Gupte, A.; Poggio, T. SGD and Weight Decay Provably Induce a Low-Rank Bias in Neural Networks. 2023. Available online: https://hdl.handle.net/1721.1/148231 (accessed on 28 November 2025).
10. Jacot, A.; Gabriel, F.; Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–10.
11. Somvanshi, S.; Javed, S.A.; Islam, M.M.; Pandit, D.; Das, S. A survey on kolmogorov-arnold network. *ACM Comput. Surv.* **2025**, *58*, 1–35. [CrossRef]
12. Medvedev, M.; Vardi, G.; Srebro, N. Overfitting behaviour of gaussian kernel ridgeless regression: Varying bandwidth or dimensionality. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 52624–52669.
13. Shang, Z.; Zhao, Z.; Yan, R. Denoising fault-aware wavelet network: A signal processing informed neural network for fault diagnosis. *Chin. J. Mech. Eng.* **2023**, *36*, 9. [CrossRef]
14. Saragadam, V.; LeJeune, D.; Tan, J.; Balakrishnan, G.; Veeraraghavan, A.; Baraniuk, R.G. Wire: Wavelet implicit neural representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18507–18516.
15. Sadoon, G.A.A.S.; Almohammed, E.; Al-Behadili, H.A. Wavelet neural networks in signal parameter estimation: A comprehensive review for next-generation wireless systems. In Proceedings of the AIP Conference Proceedings, Pune, India, 18–19 May 2024; AIP Publishing LLC.: College Park, MD, USA, 2025; Volume 3255, p. 020014.
16. Akujuobi, C.M. *Wavelets and Wavelet Transform Systems and Their Applications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2022.
17. Wang, P.; Wen, Z. A spatiotemporal graph wavelet neural network (ST-GWNN) for association mining in timely social media data. *Sci. Rep.* **2024**, *14*, 31155. [CrossRef]
18. Uddin, Z.; Ganga, S.; Asthana, R.; Ibrahim, W. Wavelets based physics informed neural networks to solve non-linear differential equations. *Sci. Rep.* **2023**, *13*, 2882. [CrossRef] [PubMed]
19. Jung, H.; Lodhi, B.; Kang, J. An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomed. Eng.* **2019**, *1*, 24. [CrossRef]
20. Xiao, Q.; Lu, S.; Chen, T. An alternating optimization method for bilevel problems under the Polyak-Łojasiewicz condition. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 63847–63873.
21. Yazdani, K.; Hale, M. Asynchronous parallel nonconvex optimization under the polyak-łojasiewicz condition. *IEEE Control Syst. Lett.* **2021**, *6*, 524–529. [CrossRef]
22. Chen, K.; Yi, C.; Yang, H. Towards Better Generalization: Weight Decay Induces Low-rank Bias for Neural Networks. *arXiv* **2024**, arXiv:2410.02176. [CrossRef]
23. Kobayashi, S.; Akram, Y.; Von Oswald, J. Weight decay induces low-rank attention layers. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 4481–4510.
24. Seleznova, M.; Kutyniok, G. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In *Mathematical and Scientific Machine Learning*; PMLR: Birmingham, UK, 2022; pp. 868–895.
25. Tan, Y.; Liu, H. How does a kernel based on gradients of infinite-width neural networks come to be widely used: A review of the neural tangent kernel. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 8. [CrossRef]

26. Tang, A.; Wang, J.B.; Pan, Y.; Wu, T.; Chen, Y.; Yu, H.; Elkashlan, M. Revisiting XL-MIMO channel estimation: When dual-wideband effects meet near field. *IEEE Trans. Wirel. Commun.* **2025**. [CrossRef]
27. Cui, Z.X.; Zhu, Q.; Cheng, J.; Zhang, B.; Liang, D. Deep unfolding as iterative regularization for imaging inverse problems. *Inverse Probl.* **2024**, *40*, 025011. [CrossRef]