

PREDICTING SPECTRAL TRAINING DYNAMICS IN BILINEAR PINNS WITH A WEIGHT-SPACE MODAL TANGENT KERNEL

Roy Vaid

University of California, Santa Barbara (UCSB), CA, USA
royvaid@ucsb.edu

ABSTRACT

PINNs can be trained to learn one physical frequency fast while another physical frequency is learned slowly. We derive the weight-space modal tangent kernel (WMTK) for bilinear Fourier PINNs. On resolved cosine grids where the linear operator has a cosine-diagonal form, we prove that the operator-composed WMTK is equivalent to the full Fourier-projected residual NTK including mode-mode couplings. The complete calculation of the kernel gives more accurate predictions of the rate of modal learning than the diagonal kernel in our experiments. For 20 different random initializations, the frozen kernel prediction yields a mean per-run trajectory R^2 of 0.9993 on reaction-diffusion and 0.9994 on near-resonant Helmholtz problems. Rescaling of the factor weights via gauge transformations leaves the initial function unchanged but changes its training trajectory. The frozen kernel approximation becomes less accurate at later times. Evaluating the WMTK kernel does not depend on the grid size for a fixed architecture and modal dimension.

Index Terms— Physics-informed neural networks, neural tangent kernel, spectral bias, bilinear networks, Fourier analysis.

1. INTRODUCTION

PINNs approximate the solution to a PDE via fitting the residuals and boundary conditions [1, 2]. Training may be slow or imbalanced between the different parts of the loss function [3–5]. This behavior is related to the neural tangent kernel (NTK) [6] through parameter sensitivities [7]. The spectral bias results [8–10] have inspired Fourier features [11] and frequency-aware PINNs [12–14]. The PDE operator might affect the difficulty of training due to its conditioning [15].

Recent work addresses the issue of PINN training through frequency [16, 17]. Zhao et al. [16] quantify the diagonal Fourier-probe tangent energies ($\|\nabla_{\theta} a_k\|_2^2$) and an operator-composed version. Can the weights also represent the complete physical-mode tangent matrix, including the couplings between the different frequencies? The full kernel can also be calculated numerically from spatial residual Jacobians, using general-purpose finite-width NTK methods [18]. This calculation depends on the sampled grid.

Pearce et al. [19] understand bilinear MLPs via their weights. There has been analysis of the inductive biases in polynomial and Hadamard-product networks [20–22]. Here, we consider a cosine-basis bilinear PINN with homogeneous Neumann conditions. We focus on linear PDE operators that are diagonal in this basis.

Contribution. We derive the modal kernel of the weights and prove that its operator-composed form equals the projected residual eNTK on midpoint cosine grids with $N > 2K$. We check whether the mode coupling helps in improving finite-horizon predictions when compared to the diagonal kernel. Gauge rescaling allows fixing the

initial represented function while changing the training dynamics. Independent spatial autodiff tests validate the derivation. We quantify both the predictive performance and the computational cost. The operator-composed WMTK is equivalent to the projected residual eNTK and calculates the same information without collocation.

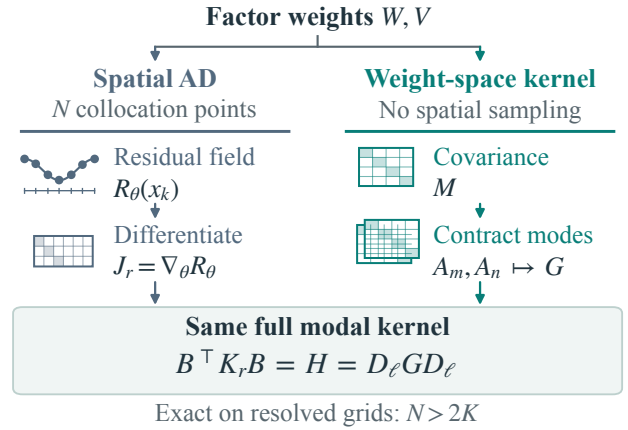


Fig. 1. Three routes to the modal kernel. Dense spatial AD forms $K_r = J_r J_r^\top$, while streamed AD projects Jacobian blocks without forming that matrix. WMTK uses factor weights and fixed mode tensors; at fixed modal dimension, its cost scales with parameter and modal dimensions rather than grid size N .

2. WEIGHT-SPACE MODAL DYNAMICS

2.1. Bilinear Fourier Architecture and DC Removal

Define $x \in [0, \pi]$, where the cosine basis is truncated at order K . Its vector representation is

$$\phi(x) = [1, \cos(x), \cos(2x), \dots, \cos(Kx)]^\top. \quad (1)$$

Every basis element and product of basis elements has a zero derivative at both ends, hence the network satisfies Neumann homogeneous boundary conditions. Cooley et al. [23] consider adaptive Fourier expansions in PINNs. Our choice of the fixed cosine basis allows us to analyze the factor weights exactly.

Consider the bilinear network $u_\theta(x)$, which has a width H_{dim} and factor matrices $W, V \in \mathbb{R}^{H_{\text{dim}} \times (K+1)}$:

$$u_\theta(x) = \frac{1}{\sqrt{H_{\text{dim}}}} \sum_{j=1}^{H_{\text{dim}}} (w_j^\top \phi(x)) (v_j^\top \phi(x)), \quad (2)$$

where $w_j, v_j \in \mathbb{R}^{K+1}$ are the transposes of the j th rows of W, V , correspondingly. As in Pearce et al. [19], the exact symmetric interaction matrix $Q_\theta \in \mathbb{R}^{(K+1) \times (K+1)}$ implements the same function:

$$Q_\theta = \frac{1}{2\sqrt{H_{\text{dim}}}} (W^\top V + V^\top W), \quad u_\theta(x) = \phi(x)^\top Q_\theta \phi(x). \quad (3)$$

Our target functions have zero mean value, hence we remove the generated DC term:

$$\bar{u}_\theta(x) = u_\theta(x) - c_0(\theta), \quad c_0(\theta) = \langle A_0, Q_\theta \rangle_F, \quad (4)$$

We introduce A_0 later. Both the target and the source lie in the same resolved modal space, which implies that the residual will be in $\text{span}\{\cos(mx)\}_{m=1}^{2K}$.

2.2. Physical Mode Extraction via Product Tensor Decomposition

Using the product-to-sum identity, we obtain

$$\cos(ix)\cos(jx) = \frac{1}{2} [\cos(|i-j|x) + \cos((i+j)x)], \quad (5)$$

Hence, products of the basis functions span physical frequencies $m \in \{0, 1, \dots, 2K\}$. For every mode m , we define a symmetric matrix $A_m \in \mathbb{R}^{(K+1) \times (K+1)}$ as follows:

$$(A_m)_{ij} = \frac{1}{2} [\mathbf{1}\{|i-j|=m\} + \mathbf{1}\{i+j=m\}], \quad (6)$$

Here, $\mathbf{1}\{\cdot\}$ is the indicator function. We calculate the Fourier coefficients of the mean-centered network from its factor weights. For $m \geq 1$,

$$c_m(\theta) = \frac{1}{\sqrt{H_{\text{dim}}}} \sum_{j=1}^{H_{\text{dim}}} w_j^\top A_m v_j = \langle A_m, Q_\theta \rangle_F. \quad (7)$$

In Equation (7), the brackets denote the Frobenius inner product, and we use no sampled points or numerical transforms. Spectral-informed networks learn in spectral coordinates [24]. We use the factor weights to examine the tangent dynamics of these modes. Under the grid, source, and operator conditions below, the spatial and modal losses coincide.

2.3. Weight-Space Modal Tangent Kernel (WMTK)

Let $\theta = \text{vec}([W, V]) \in \mathbb{R}^P$, where $P = 2H_{\text{dim}}(K+1)$. The symmetry of A_m gives the parameter gradients for $m \geq 1$:

$$\nabla_{w_j} c_m = \frac{1}{\sqrt{H_{\text{dim}}}} A_m v_j, \quad \nabla_{v_j} c_m = \frac{1}{\sqrt{H_{\text{dim}}}} A_m w_j. \quad (8)$$

Each entry of the weight-space modal tangent kernel $G \in \mathbb{R}^{2K \times 2K}$ is an inner product of mode gradients. For $m, n \in \{1, \dots, 2K\}$, we set $G_{mn} = \langle \nabla_\theta c_m, \nabla_\theta c_n \rangle$:

$$G_{mn} = \frac{1}{H_{\text{dim}}} \sum_{j=1}^{H_{\text{dim}}} [(A_m v_j)^\top (A_n v_j) + (A_m w_j)^\top (A_n w_j)]. \quad (9)$$

Let the linear operator $\mathcal{L}$ satisfy $\mathcal{L} \cos(mx) = \ell_m \cos(mx)$. We choose $u^*(x) = \sum_{m=1}^{2K} c_m^* \cos(mx)$ and set $f = \mathcal{L} u^*$. The residual in mode m is $r_m = \ell_m (c_m - c_m^*)$. With $D_\ell = \text{diag}(\ell_1, \dots, \ell_{2K})$, the residual-space kernel $H \in \mathbb{R}^{2K \times 2K}$ is

$$H = D_\ell G D_\ell, \quad H_{mm} = \ell_m^2 G_{mm}. \quad (10)$$

Set $M = (V^\top V + W^\top W)/H_{\text{dim}}$. Then $G_{mn} = \text{tr}(A_m A_n M)$. Dense contraction takes $\mathcal{O}(H_{\text{dim}} K^2 + K^4)$ flops. It uses $\mathcal{O}(H_{\text{dim}} K + K^3)$ storage. Neither cost depends on the grid size N .

2.4. Collocation Points and Discrete GD Dynamics

We use $N > 2K$ midpoint cosine (DCT-II) collocation points:

$$x_k = \frac{(k - \frac{1}{2})\pi}{N}, \quad k = 1, \dots, N. \quad (11)$$

Set $\Phi \in \mathbb{R}^{N \times 2K}$ with $\Phi_{km} = \cos(mx_k)$, and let $B = (2/N)\Phi$. Orthogonality on this grid gives $\Phi^\top \Phi = (N/2)I$ and $B^\top \Phi = I$ for modes $1, \dots, 2K$. The mean-squared spatial residual loss then equals the modal loss:

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{k=1}^N |\mathcal{L} \bar{u}_\theta(x_k) - f(x_k)|^2 = \frac{1}{2} \|r_\theta\|_2^2. \quad (12)$$

The spatial residual has Jacobian $J_r = \nabla_\theta R_\theta = \Phi J_{\text{modal}} \in \mathbb{R}^{N \times P}$, where $J_{\text{modal}} = D_\ell \nabla_\theta c$. Its empirical residual NTK is $K_r = J_r J_r^\top \in \mathbb{R}^{N \times N}$.

Proposition 1. *For the centered bilinear model, assume a cosine-diagonal linear operator, a source in the resolved modal space, and $N > 2K$. Then the Fourier-projected residual eNTK equals the operator-composed WMTK H for any factor weights:*

$$B^\top K_r B = (B^\top \Phi) J_{\text{modal}} J_{\text{modal}}^\top (\Phi^\top B) = H. \quad (13)$$

The last equality follows from $B^\top \Phi = I$. At $N = 2K$, the highest cosine mode vanishes at every midpoint, so we require $N > 2K$. For gradient descent (GD) with step size η , we linearize around θ_0 [6, 25]. The predicted modal trajectory is

$$\hat{r}_t = (I - \eta H_0)^t r_0, \quad H_0 = H(\theta_0). \quad (14)$$

The frozen trajectory approximates the finite-width run. For the full update, set $g_j^W = \nabla_{w_j} \mathcal{J}$ and $g_j^V = \nabla_{v_j} \mathcal{J}$. Bilinearity gives the exact one-step residual update:

$$r_{t+1} = r_t - \eta H_t r_t + \eta^2 q_t, \quad (q_t)_m = \frac{\ell_m}{\sqrt{H_{\text{dim}}}} \sum_j (g_j^W)^\top A_m g_j^V. \quad (15)$$

Changes in H_t and the quadratic term q_t limit predictions at later steps. In the positive eigenspace of H_0 , the frozen update is stable when $0 < \eta \lambda_{\max}(H_0) < 2$.

2.5. Representation vs. Trainability: Gauge Invariance

We rescale the factors as $W \rightarrow cW$ and $V \rightarrow V/c$ for $c > 0$. The matrix Q_θ and the mean-centered network $\bar{u}_\theta$ stay the same. Equation (8) shows that the W -gradient scales by c^{-1} and the V -gradient by c :

$$G(c) = c^{-2} G_V + c^2 G_W, \quad (16)$$

Here, $(G_V)_{mn} = H_{\text{dim}}^{-1} \sum_j (A_m v_j)^\top (A_n v_j)$, and G_W has the same form with w_j . Since W and V have the same initialization distribution, $\mathbb{E}[G_V] = \mathbb{E}[G_W] = \frac{1}{2} \mathbb{E}[G(1)]$. Thus,

$$\mathbb{E}[G(c)] = \frac{1}{2} [c^2 + c^{-2}] \mathbb{E}[G(1)] \succeq \mathbb{E}[G(1)]. \quad (17)$$

The re-scaling of the gauge leaves Q and the function itself unchanged. However, GD acts on W and V separately. Equation (17) shows the difference in the expectation of the tangent kernel. The residual and the step size affect the training trajectory.

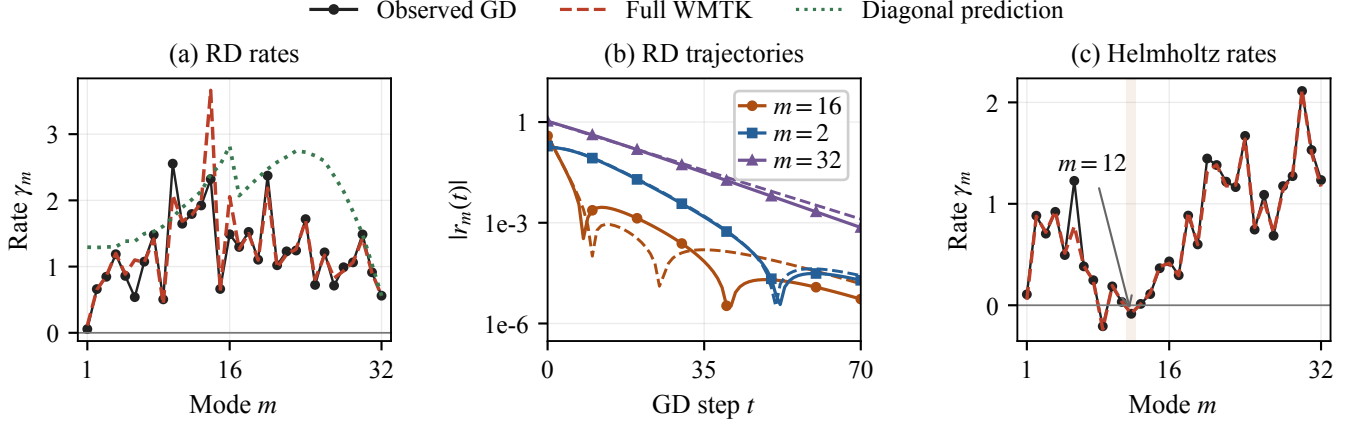


Fig. 2. Results for one initialization; Table 1 shows the averages over 20 initializations. (a) Signed RD rates from GD, full WMTK, and the diagonal kernel. (b) Magnitudes of the RD residuals from GD (solid) and the frozen kernel (dashed). (c) Signed Helmholtz rates, where negative values around the operator notch at $m = 12$ indicate transient growth.

2.6. Spectral Bias at Network Initialization

For the Gaussian initialization with variance σ^2 of W and V , $\mathbb{E}[G_{mn}] = 2\sigma^2 \text{tr}(A_m A_n)$. For $K = 16$, $\|A_m\|_F^2 = 9.50$ for $m = 1$, 5.75 for $m = 16$, and 0.25 for $m = 32$. Fewer pairs of basis functions give rise to higher modes. Nevertheless, the expected kernel correlates different modes: for $K \geq 2$, $\text{tr}(A_1 A_3) = 1/2$. After averaging over random initialization, such correlations persist. They do not arise due to finite-width noise.

3. EXPERIMENTAL RESULTS

3.1. Experimental Design and Independent Verification

At $K = 16$, the network has 32 physical modes; we set $H_{\text{dim}} = 128$ and $N = 512$ midpoint collocation points. The factors are initialized as $W, V \sim \mathcal{N}(0, 1/K)$, and the target coefficient for each mode is $c_m^* = s_m / \sqrt{2K}$, where $s_m \in \{-1, +1\}$ and is drawn randomly. For seeds 0–19, we perform 70 full-batch GD steps with $\eta = 0.50 / \lambda_{\max}(H_0)$. We test two operators. Reaction-diffusion employs $\mathcal{L} = I - v \partial_{xx}$, $v = 0.003$, and $\ell_m = 1 + v m^2$; Helmholtz employs $\mathcal{L} = I + \kappa^{-2} \partial_{xx}$, $\kappa = 12.3$, and $\ell_m = 1 - (m/\kappa)^2$. The cosine basis observes homogeneous Neumann boundary conditions.

We perform training with the modal loss and verify its gradients using spatial-residual GD with automatic differentiation (AD). The second implementation evaluates the network in space, computes two spatial derivatives and then differentiates the residual with respect to the factor weights. This implementation does not involve A_m and analytic modal gradients. For 20 seeds of each operator at $N = 512$, the difference between the projected kernel and WMTK does not exceed 10^{-13} relative to the former. In an additional initialization, the 70-step trajectory matches the modal solver within 2×10^{-14} relative error. Such checks support the correctness of the derivation and implementation.

3.2. Rate Prediction and Cross-Mode Coupling

For each mode, we fit $\log(|r_m(t)| + 10^{-12})$ against $t\eta$ over $t = 0, \dots, 19$ and calculate the negative slope as γ_m . For the frozen prediction, we use the same fit to compute $\hat{\gamma}_m$. If the fitted rate is negative, it means that the fitted log-residual magnitude has an upward

Table 1. Benchmark results over 20 seeds. Results are reported as mean $\pm$ population standard deviation. Predictors are based on data from initialization. Trajectories consist of 71 states and 32 modes.

Predictor	RD $\rho \uparrow$	Helmholtz $\rho \uparrow$
Frequency $-m$	-0.22 ± 0.10	-0.60 ± 0.05
Operator ℓ_m^2	0.22 ± 0.10	0.86 ± 0.05
Capacity $(G_0)_{mm}$	-0.22 ± 0.10	-0.60 ± 0.05
Diagonal $(H_0)_{mm}$	0.40 ± 0.14	0.86 ± 0.06
Expected full kernel	0.76 ± 0.11	0.95 ± 0.04
Full WMTK	0.95 ± 0.03	0.99 ± 0.02
Trajectory error $\downarrow$	RD (%)	Helmholtz (%)
Diagonal trajectory	16.6 ± 3.8	16.5 ± 4.3
Expected full kernel	8.08 ± 1.92	8.14 ± 1.97
Full WMTK	2.49 ± 0.70	2.32 ± 0.69

trend during this period of time. We evaluate rates' agreement with Spearman correlation $\rho(\gamma, \hat{\gamma})$ and calculate trajectories with $\epsilon_{\text{traj}} = \|r - \hat{r}\|_F / \|r\|_F$. For each run, $R^2 = 1 - \|r - \hat{r}\|_F^2 / \|r - \text{mean}(r)\|_F^2$, using all states and modes.

RD has two forces affecting its modal rates: as m grows, ℓ_m^2 increases, but the network capacity decreases. This results in the non-monotonic rate profile dependent on the initial residual and cross-mode coupling in H_0 . Over 20 seeds, the full kernel gives $\rho = 0.95$, in comparison to 0.40 for the diagonal predictor. The mean trajectory error of the full kernel over 71 states and 32 modes is 2.49%, while the mean trajectory error of the diagonal trajectory is 16.6%. Proposition 1 guarantees the equality between the projected residual eNTK and WMTK for this frozen prediction.

Then we substitute $\mathbb{E}[H_0]$ for H_0 for each run, keeping its r_0 and η . The expected kernel preserves cross-mode coupling but loses the difference in factor weights. The mean trajectory error grows to 8.08% for RD and 8.14% for Helmholtz. The particular factor weight of a run affects the prediction.

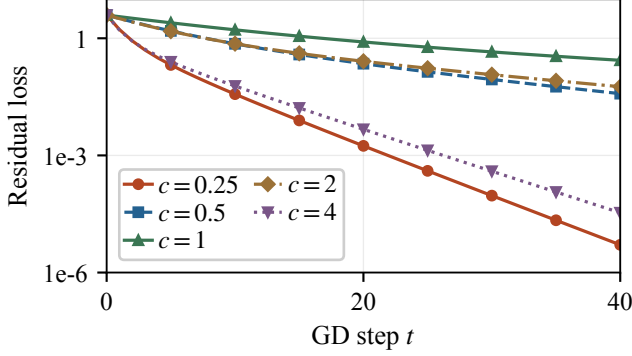


Fig. 3. Gauge rescaling using $\eta = 0.03$. All five runs begin with the same function and Q . After 40 steps, their losses span 4.73 orders of magnitude.

For Helmholtz, $\ell_{12} \simeq 0.048$ produces the notch in the operator spectrum at mode 12. In Fig. 2(c), the fitted rates at modes 9 and 12 are negative. The positivity-semidefiniteness of the kernel does not guarantee the decrease of the magnitude of each residual component at each step; the coupling can lead to the increase of one of them while training the full system.

3.3. Robustness and Prediction Limits

We test widths 32, 64, 128, and 256 with $\eta\lambda_{\max}(H_0) = 0.1, 0.25$, or 0.5, providing 480 runs over 20 seeds and two operators. The targets stay fixed across widths and step sizes. The minimum configuration mean R^2 is 0.9917. The minimum R^2 of a single run is 0.9771.

The other 360 runs vary K in 8, 16, and 32 and the target spectrum. The targets have flat, $1/m$, or high-band-only ($m > K$) envelopes, with random signs and unit coefficient norm. We fix the width to 128 and $\eta\lambda_{\max}(H_0)$ to 0.5. See Table 2 for the minimum configuration means. In all 18 configurations, the mean full-kernel rate correlation is higher than the diagonal one.

Scores over whole trajectories hide the late-time errors. In 70 steps, the per-run median modal error is 1.4% for RD and 1.7% for Helmholtz. In the last 20 states, RD gives the mean per-run relative error of 96.1% although its mean RMS discrepancy is 0.0756% of the initial residual norm. The tail error of Helmholtz is 2.7%. In 500 steps, the mean whole-trajectory errors are 2.49% and 2.32%, while the tail errors are 102.1% and 3.6% for RD and Helmholtz, respectively. The frozen kernel has less accuracy for late residuals.

3.4. Computational Complexity and Gauge Rescaling

We measure the time of WMTK against two spatial-AD routes: dense AD calculates $B^\top J_r J_r^\top B$, while streamed AD accumulates $B^\top J_r$ in blocks before calculating the modal Gram matrix. All three

Table 2. Bandwidth and target sweep. For each envelope, the table shows the minimum configuration mean R^2 over both operators and three bandwidths. It also shows the minimum mean full-kernel rate correlation for each operator. All 18 configurations have 20 seeds.

Target envelope	R^2	RD ρ	Helm. ρ
Flat	0.9978	0.961	0.974
$1/m$	0.9999	0.992	0.996
High-band-only	0.9923	0.928	0.951

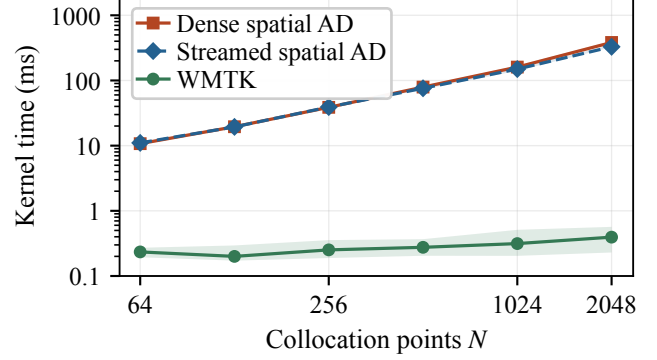


Fig. 4. Kernel time in float64. The points show the median and the interquartile range of 50 measurements for each method and grid size N . Both AD methods use spatial and parameter differentiation. Streamed AD avoids the dense $N \times N$ Gram matrix.

methods use PyTorch CPU float64 with one thread. Timing excludes fixed input preparation. See Fig. 4 for the results dependent on the grid size.

At $N = 2048$, $K = 16$, $H_{\text{dim}} = 128$, WMTK requires 0.39 ms, in comparison to 328.6 ms for streamed AD and 383.0 ms for dense AD. The speedup is $833\times$ compared to the streamed AD implementation tested here. The asymptotic cost of WMTK kernel calculation is independent of N for fixed K and width.

For the gauge experiment, we use $c \in \{0.25, 0.5, 1, 2, 4\}$ and $\eta = 0.03$; all five runs have the same function and interaction matrix Q . Their $\text{Tr}(G)$ values vary from 19.3 to 158.5, and after 40 updates the losses span 4.73 orders of magnitude (Fig. 3). But only the factorization has been changed. Hence, Q cannot predict the GD trajectory alone.

4. CONCLUSION

The WMTK serves as a model to study the spectral behavior of bilinear cosine PINNs before running gradient descent. Given the starting residual and weights, the WMTK can predict the decay/growth behavior of modes of interest, incorporating the PDE operator and the coupling effect in weight space. We show that for the operator-composed WMTK, this is equivalent to the full Fourier-projected residual eNTK in terms of modal predictions for linear operators diagonal in the cosine basis. For our reaction diffusion and near resonant Helmholtz equations, we find that the full kernel predicts modal behavior better than the diagonal score and finite horizon dynamics better than the diagonal kernel. This is accomplished without performing a kernel calculation that depends on the size of the collocation grid for a fixed architecture and mode number. Gauge rescaling demonstrates the role of weights, since one and the same function may exhibit a very different training trajectory. More generally, the above findings indicate that imbalances of PINN spectral behavior could be understood in terms of PDE operator sensitivity and mode coupling due to the network parameters.

ACKNOWLEDGMENT

OpenAI Codex (Astra) was used to develop and debug the experiment and verification code for Sections 2–3 and to prepare the plotting code for Figs. 1–4.

This work has not received any funding. There are no relevant financial or non-financial conflicts of interest to disclose.

COMPLIANCE WITH ETHICAL STANDARDS

Ethical Approval: This is a numerical simulation study for which no ethical approval was required.

5. REFERENCES

- [1] M. Raissi, P. Perdikaris, and G. 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*, vol. 378, pp. 686–707, 2019.
- [2] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, “Scientific machine learning through physics-informed neural networks: Where we are and what’s next,” *Journal of Scientific Computing*, vol. 92, no. 3, article 88, 2022.
- [3] S. Wang, Y. Teng, and P. Perdikaris, “Understanding and mitigating gradient flow pathologies in physics-informed neural networks,” *SIAM Journal on Scientific Computing*, vol. 43, no. 5, pp. A3055–A3081, 2021.
- [4] A. Krishnapriyan, A. Gholami, S. Zhe, R. Kirby, and M. W. Mahoney, “Characterizing possible failure modes in physics-informed neural networks,” in *Proc. NeurIPS*, 2021, vol. 34, pp. 26548–26560.
- [5] S. Wang, S. Sankaran, H. Wang, and P. Perdikaris, “An expert’s guide to training physics-informed neural networks,” *arXiv preprint arXiv:2308.08468*, 2023.
- [6] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Proc. NeurIPS*, 2018, vol. 31, pp. 8571–8580.
- [7] S. Wang, X. Yu, and P. Perdikaris, “When and why PINNs fail to train: A neural tangent kernel perspective,” *Journal of Computational Physics*, vol. 449, article 110768, 2022.
- [8] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, “On the spectral bias of neural networks,” in *Proc. ICML*, 2019, pp. 5301–5310.
- [9] R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman, “Frequency bias in neural networks for input of non-uniform density,” in *Proc. ICML*, 2020, pp. 685–694.
- [10] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, “Frequency principle: Fourier analysis sheds light on deep neural networks,” *Communications in Computational Physics*, vol. 28, no. 5, pp. 1746–1767, 2020.
- [11] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, “Fourier features let networks learn high frequency functions in low dimensional domains,” in *Proc. NeurIPS*, 2020, vol. 33, pp. 7537–7547.
- [12] S. Wang, H. Wang, and P. Perdikaris, “On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks,” *Computer Methods in Applied Mechanics and Engineering*, vol. 384, article 113938, 2021.
- [13] G. Jin, J. C. Wong, A. Gupta, S. Li, and Y.-S. Ong, “Fourier warm start for physics-informed neural networks,” *Engineering Applications of Artificial Intelligence*, vol. 132, article 107887, 2024.
- [14] Y. Liu, H. Gu, X. Yu, and P. Qin, “Diminishing spectral bias in physics-informed neural networks using spatially-adaptive Fourier feature encoding,” *Neural Networks*, vol. 182, article 106886, 2025.
- [15] T. De Ryck, F. Bonnet, S. Mishra, and E. de Bézenac, “An operator preconditioning perspective on training in physics-informed machine learning,” in *Proc. ICLR*, 2024.
- [16] B. Zhao, Y. Wang, J. T. Barron, K. L. Bouman, D. Verbin, and P. P. Srinivasan, “Fourier feature pyramids for physics-informed neural networks,” *arXiv preprint arXiv:2605.24278*, 2026.
- [17] S. Khodakarami, V. Oommen, N. A. Daryakenari, M. Beekenkamp, and G. E. Karniadakis, “Spectral bias in physics-informed and operator learning: Analysis and mitigation guidelines,” *Computer Methods in Applied Mechanics and Engineering*, vol. 461, article 119156, 2026.
- [18] R. Novak, J. Sohl-Dickstein, and S. S. Schoenholz, “Fast finite width neural tangent kernel,” in *Proc. ICML*, 2022, pp. 17018–17044.
- [19] M. T. Pearce, T. Dooms, A. Rigg, J. M. Oramas, and L. Sharkey, “Bilinear MLPs enable weight-based mechanistic interpretability,” in *Proc. ICLR*, 2025.
- [20] G. G. Chrysos, S. Moschoglou, G. Bouritsas, Y. Panagakis, J. Deng, and S. Zafeiriou, “PI-nets: Deep polynomial neural networks,” in *Proc. CVPR*, 2020, pp. 7325–7335.
- [21] Y. Wu, Z. Zhu, F. Liu, G. G. Chrysos, and V. Cevher, “Extrapolation and spectral bias of neural nets with Hadamard product: A polynomial net study,” in *Proc. NeurIPS*, 2022, vol. 35, pp. 26980–26993.
- [22] M. Choraria, L. T. Dadi, G. Chrysos, J. Mairal, and V. Cevher, “The spectral bias of polynomial neural networks,” in *Proc. ICLR*, 2022.
- [23] M. Cooley, V. Shankar, R. M. Kirby, and S. Zhe, “Fourier PINNs: From strong boundary conditions to adaptive Fourier bases,” *Transactions on Machine Learning Research*, 2025.
- [24] T. Yu, Y. Qi, I. V. Oseledets, and S. Chen, “Spectral informed neural networks,” *Journal of Computational and Applied Mathematics*, vol. 477, article 117178, 2026.
- [25] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, “Wide neural networks of any depth evolve as linear models under gradient descent,” in *Proc. NeurIPS*, 2019, vol. 32, pp. 8572–8583.