EXP001: Finite-width MLPs converge to the NTK limit

Experiment Summary with Figures

Shreyas Kalvankar

November 2025

Goal and Context

The goal of these experiments is to study whether finite-width MLPs trained with gradient descent converge to the infinite-width NTK predictor and to identify finite-width or frequency-dependent effects.

The theoretical baseline (derived in the main manuscript) establishes (under constant kernel assumption):

$$\frac{df_t}{dt} = -K(f_t - y),$$

$$f_t = y + e^{-Kt}(f_0 - y).$$

1 Experiment 2.1: Reproduction of NTK kernel profile

Probe manifold: unit circle $x(\gamma) = [\cos\gamma, \sin\gamma]^\top$ with anchor $x_0 = (1, 0)$. Regression task for training: $f^*(x) = x_1 x_2$ on Gaussian inputs (the circle is only for probing). Model: ReLU MLP, depth $L = 4$, NTK parameterization. Widths $n \in \{100, 500, 2000\}$, training for 200 steps, $\eta = 1.0$, 10 seeds.

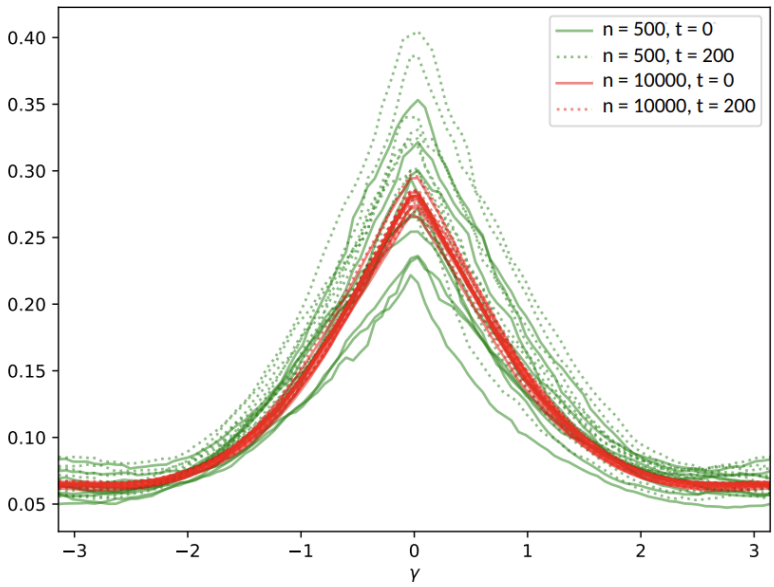

Figure 1: Convergence of the NTK to a fixed limit for two widths n and two times t.

Figure 1: Reference kernel profile from the NTK paper for comparison.

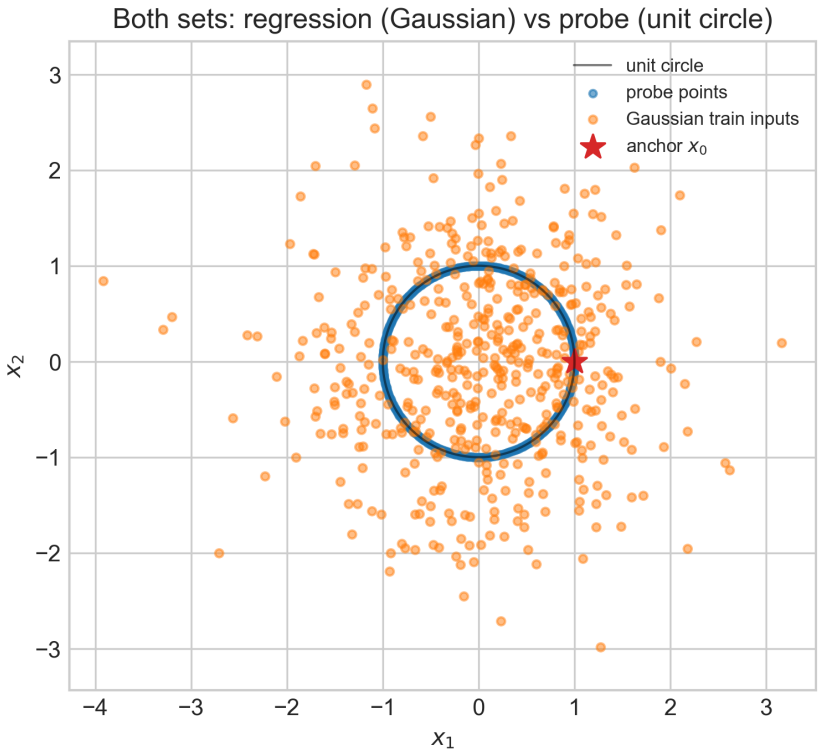

Figure 2: Training set (Gaussian inputs) vs probe manifold (unit circle).

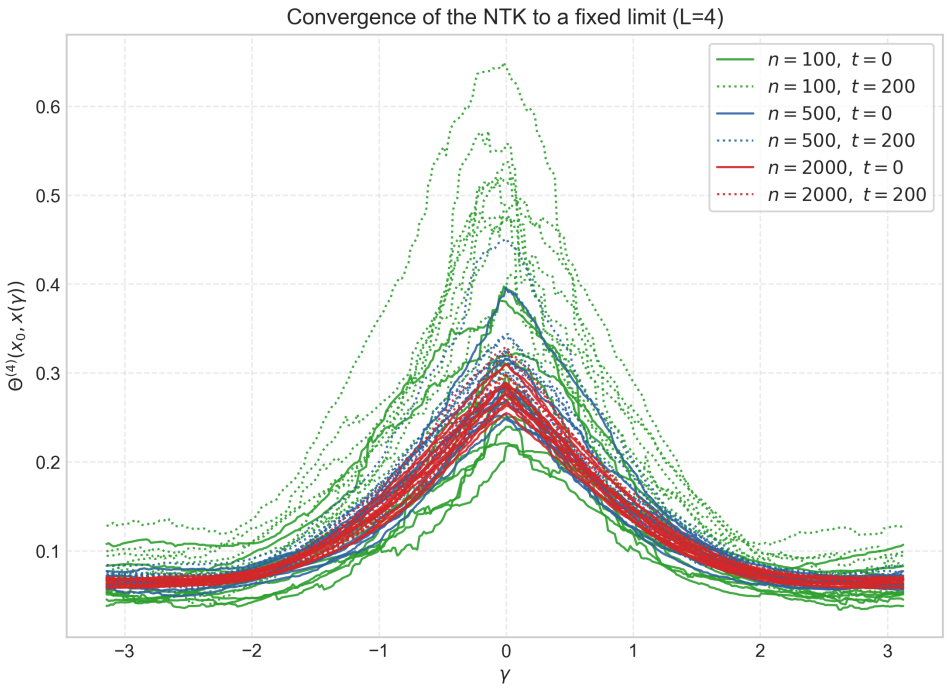

Figure 3: Empirical NTK profile $\gamma \mapsto \Theta^{(4)}_{\theta_t}(x_0, x(\gamma))$ at $t = 0$ (solid) and $t = 200$ (dotted). Variance decreases as width increases.

Observation. Variance across seeds decreases with width. The peak near $\gamma = 0$ corresponds to self-similarity ($x_0 \cdot x(\gamma) \approx 1$). As width grows, the kernel concentrates toward a deterministic limit.

2 Experiment 2.2: Function-space convergence on S^1 (simple target)

Inputs: $x(\gamma) = (\cos\gamma, \sin\gamma)$, target $f^*(x) = x_1 x_2 = \frac{1}{2}\sin(2\gamma)$ (low-frequency). Depth $L = 1$, widths $n \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$, full-batch GD with $\eta = 10^{-2}$, trained for 30,000 steps. Analytic NTK predictor built with the same activation and initialization.

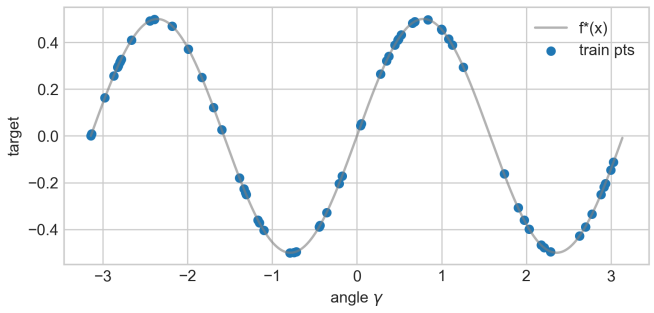

Figure 4: Simple regression target on the circle: $\frac{1}{2}\sin(2\gamma)$.

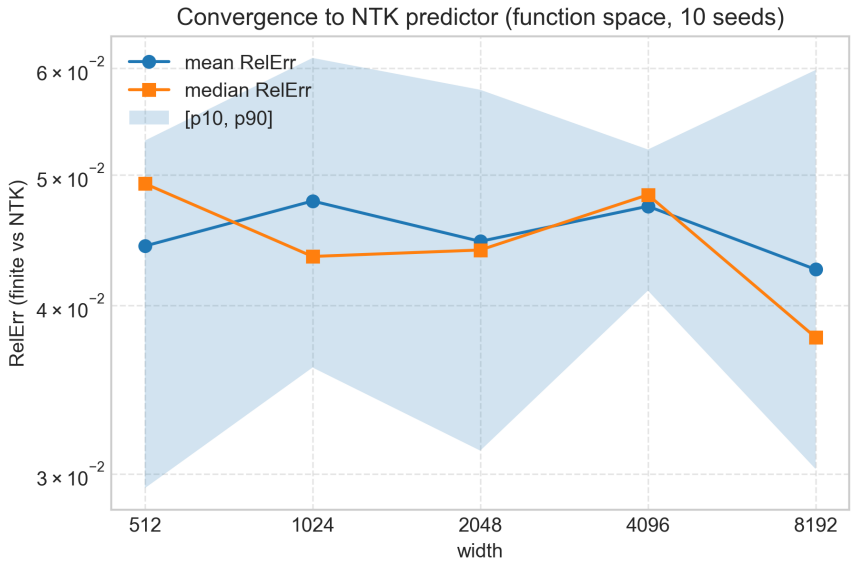

Figure 5: Relative error (finite vs NTK predictor) for the simple task.

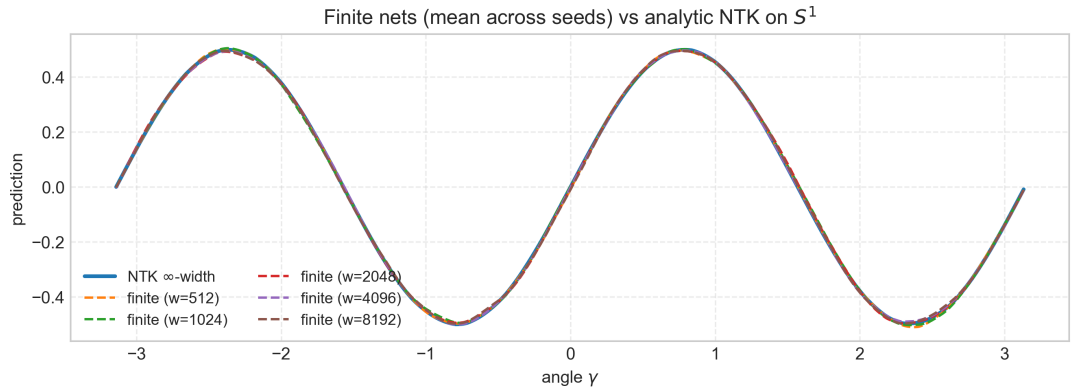

Figure 6: Overlays of finite-width predictions vs analytic NTK. Excellent match across all widths.

Result. Finite-width networks match the analytic NTK predictor almost exactly; RelErr is flat. This task lies within a single low-frequency eigenmode that even small networks approximate well.

3 Experiment 2.3: Function-space convergence on a Fourier mixture

Target:

$$y(\gamma) = \sum_{k\in\{2,4,7,11,16,23,32\}} a_k \sin(k\gamma + \phi_k),$$

trained for 30,000 steps with $\eta = 0.01$.

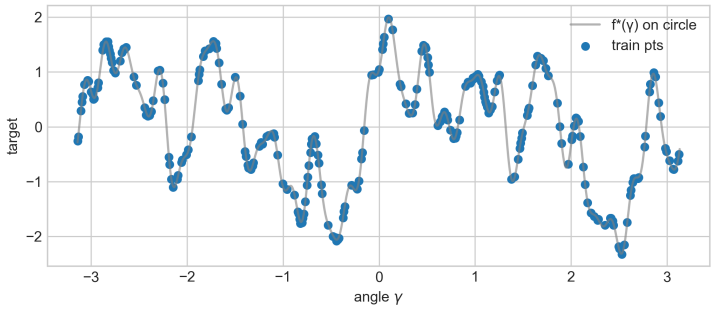

Figure 7: Complex Fourier mixture target with seven harmonics.

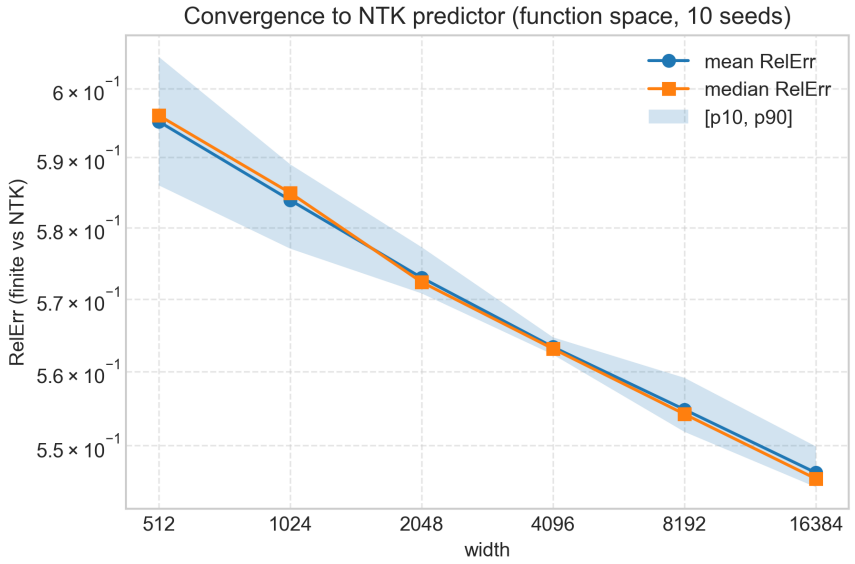

Figure 8: RelErr vs width for the Fourier mixture task.

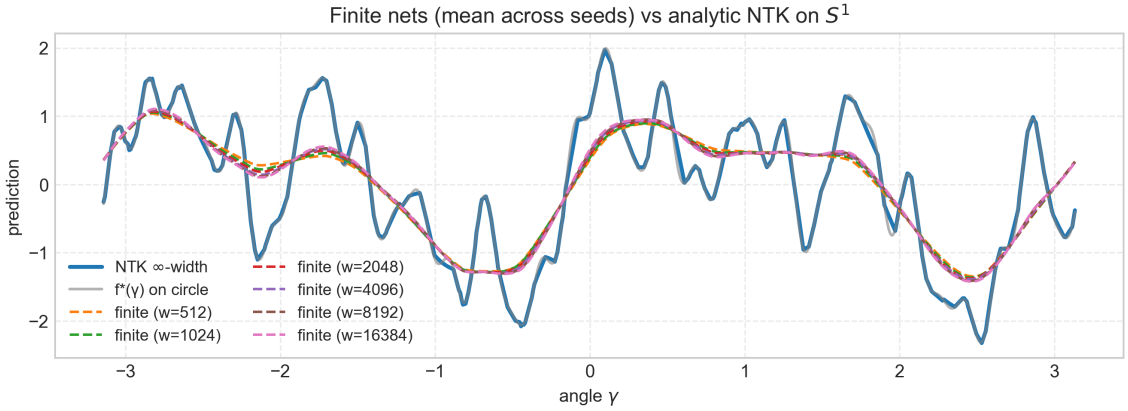

Figure 9: Finite nets vs analytic NTK on the Fourier mixture. All widths collapse to the second harmonic.

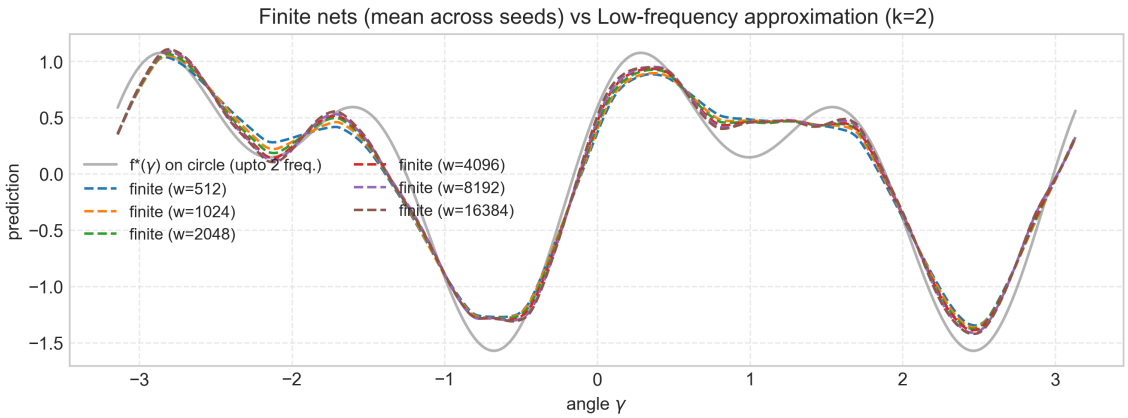

Figure 10: Partial overlay highlighting near-identical outputs across widths.

Observation. All widths converge to similar low-frequency functions dominated by $k = 2$. RelErr decreases only slightly (for example, $0.60 \to 0.55$). Finite-width networks appear trapped in the low-frequency mode.

4 Experiment 2.4: Same task at low widths

Up to this point, all widths seemed to produce almost identical output functions, none of which matched the analytic NTK prediction. This contradicted the initial hypothesis that increasing width would make network predictions converge toward the NTK predictor.

However, since the input features are $(\sin\gamma, \cos\gamma)$, it is not difficult for a network with hundreds or thousands of neurons (e.g., 512, 1024, 2048) to approximate the lower harmonics of a Fourier mixture with only seven components. Consequently, the prediction functions for all large widths looked nearly identical.

To visualize the convergence more clearly, the idea was to start from very low widths. If we use only two neurons, approximating seven harmonics is extremely difficult; as width increases, the network's representational capacity should gradually improve, and we should observe visible differences in the predicted functions. This is indeed what we found.

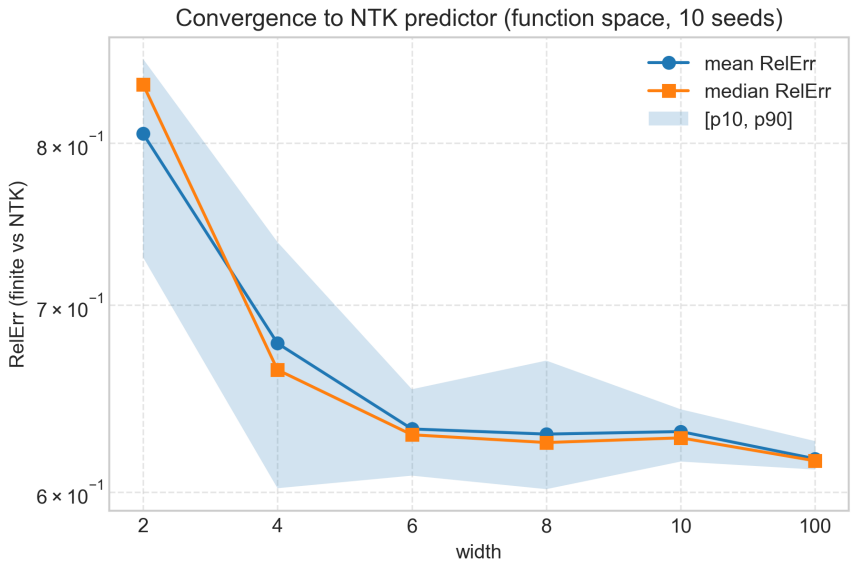

Figure 11: RelErr vs width for very small widths on the Fourier mixture.

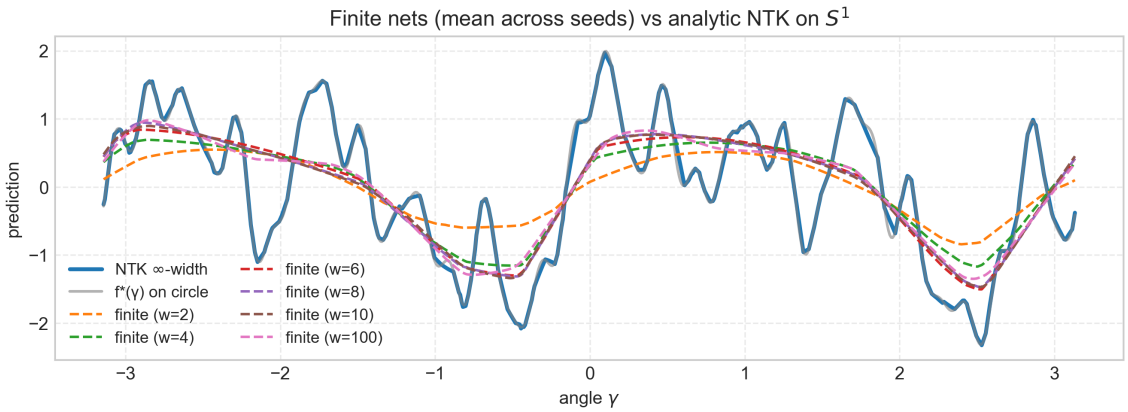

Figure 12: Visible progression of function-space convergence with width.

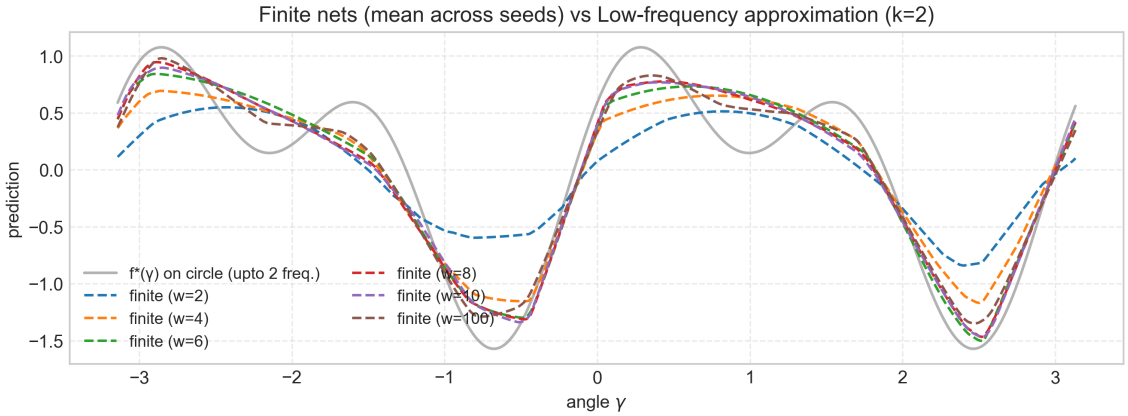

Figure 13: Partial overlay: higher harmonics gradually restored as width increases.

Observation. Narrow networks capture only coarse, low-order harmonics. Increasing width improves representation, approaching the NTK predictor.

5 Experiment 2.5: Higher learning rate and more steps

The previous results suggested that all widths were effectively learning the same dominant low-frequency mode, consistent with the *frequency principle*: gradient descent first fits low-frequency components, while higher-frequency ones converge much more slowly. This raised the question of whether wider networks might be capable of learning the higher-frequency harmonics if trained for longer or with a faster effective rate.

To test this, training was extended to 100,000 steps with a learning rate of $\eta = 1.0$ for widths $w \in \{2, 4, 6, 8, 10, 100, 1000, 10000\}$. The expectation was that larger ηt would allow higher-frequency modes (associated with smaller NTK eigenvalues) to decay and that the networks would begin to approach the NTK limit more closely.

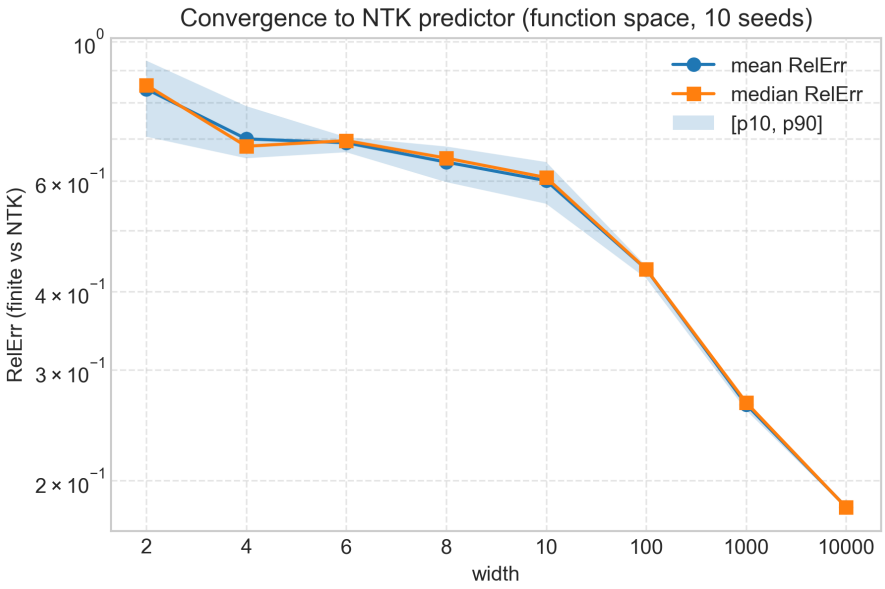

Figure 14: RelErr vs width with higher LR and longer training.

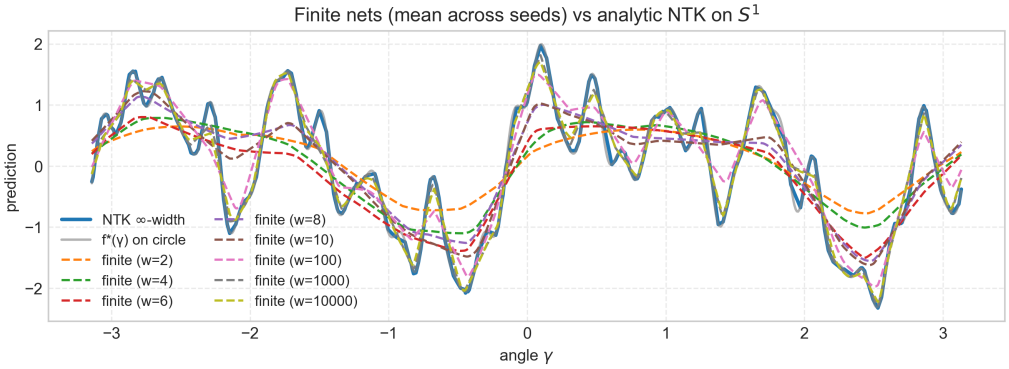

Figure 15: Convergence across widths improves substantially with larger ηt.

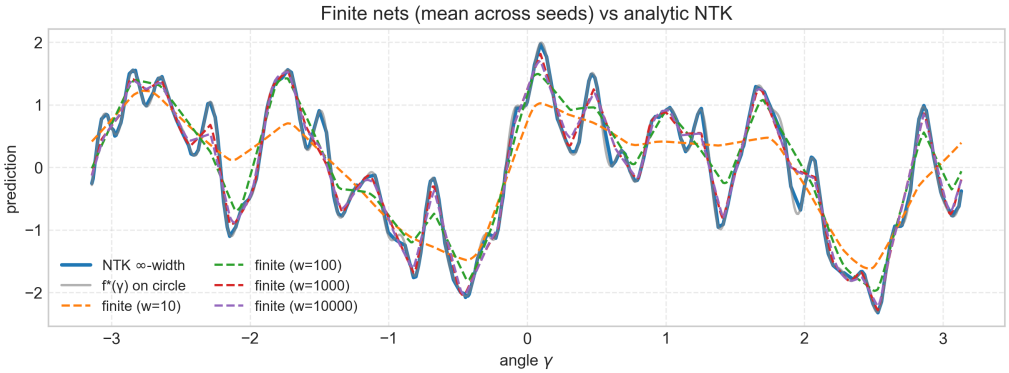

Figure 16: High-width overlays $(10, 100, 1000, 10000)$: clear approach toward the NTK predictor.

Results. With longer training and a higher learning rate, convergence became visibly stronger. High-frequency harmonics began to appear, particularly at larger widths, and the relative error decreased more clearly, aligning with NTK theory.

Interpretation. The lower widths appear limited by spectral bias—they cannot move beyond the dominant low-frequency mode—while the wider networks, given sufficient training time and step size, begin to approximate higher frequencies. This indicates that the networks are capable of approaching the NTK predictor, but doing so requires longer effective training (larger ηt), reflecting the discrete-time nature of gradient descent dynamics.

6 From Continuous Gradient Flow to Discrete Gradient Descent

In the Fourier–mixture experiments, simply increasing width did not move the finite networks toward the NTK predictor, whereas increasing the learning rate and the number of steps did. I try to derive the discrete GD recurrence in function space and relate its mode factors $(1-\eta\lambda_j)^t$ to the NTK spectrum and to the trends seen in Experiments 2.4 and 2.5.

Discrete GD in function space Let $f(\theta) \in \mathbb{R}^n$ be the stacked predictions on the training inputs and let $r := f(\theta) - y$. Consider

$$L(\theta) = \tfrac{1}{2}\|r\|_2^2.$$

Write the Jacobian explicitly as the parameter gradient of f,

$$\nabla_\theta f(\theta) \in \mathbb{R}^{n\times p},$$

so that

$$\nabla_\theta L(\theta) \;=\; \big(\nabla_\theta f(\theta)\big)^\top r.$$

A full–batch gradient descent step with step size η is

$$\theta_{t+1} = \theta_t - \eta\,\nabla_\theta L(\theta_t) \;=\; \theta_t - \eta\,\big(\nabla_\theta f(\theta_t)\big)^\top r_t, \qquad r_t := f(\theta_t) - y.$$

Linearize f at θ_t:

$$f(\theta_{t+1}) \;\approx\; f(\theta_t) - \eta\,\nabla_\theta f(\theta_t)\big(\nabla_\theta f(\theta_t)\big)^\top r_t \;=\; f_t - \eta\,K_t\,r_t, \qquad K_t := \nabla_\theta f(\theta_t)\big(\nabla_\theta f(\theta_t)\big)^\top.$$

Convert to residuals by subtracting y on both sides:

$$r_{t+1} \;=\; f(\theta_{t+1}) - y \;\approx\; (f_t - \eta K_t r_t) - y \;=\; (f_t - y) - \eta K_t r_t \;=\; r_t - \eta K_t r_t \;=\; (I - \eta K_t)\,r_t.$$

Hence the residual update is

$$\boxed{r_{t+1} \;\approx\; (I - \eta\,K_t)\,r_t}$$

When the kernel is effectively constant during training (lazy regime or features frozen at $t = 0$), $K_t \equiv K$ and the recurrence becomes a linear time–invariant system with exact closed form

$$\boxed{r_t = (I - \eta K)^t\,r_0, \qquad f_t = y + (I - \eta K)^t(f_0 - y).}$$

7 Open Questions and Next Steps

1. Continuous vs. Discrete Dynamics. Analyses based purely on gradient flow implicitly assume infinitesimal step size, thereby ignoring the learning rate as a hyperparameter. However, experiment 2.5 shows that η itself plays a crucial role in how higher–frequency components are learned. The discrete recurrence

$$r_t = (I - \eta K)^t\,r_0$$

captures this dependence explicitly and provides a more realistic description when training with finite η.

2. Fourier and Spectral Analysis. In the Fourier–mixture experiments, the networks consistently learned the low–frequency components of the target much faster than the high–frequency components. This is consistent with the *spectral bias* (or frequency principle) reported by Rahaman et al. (2019), where smoother, low–frequency features dominate early learning while higher frequencies require more training steps or a larger effective rate (ηt).

I think we can further investigate this by performing a Fourier analysis of the analytic NTK kernel on the circle to study its eigenvalue spectrum. In the infinite–width limit, this is expected to explain why different frequencies converge at different rates; Rahaman et al. (2019) does that I believe. However, I am not yet sure how to do this in the finite case.

3. Finite–Width Kernel Drift. Hanin and Nica (2019) showed that when depth and width co–scale ($\beta = \frac{depth}{width} \approx 1$), the NTK at initialization does not concentrate and it can evolve non–trivially during training. Building on this, Seleznova and Kutyniok (2022) also formalize and extend these effects, where they also show that the NTK need not remain deterministic nor constant as the depth–to–width ratio grows. In the current experiment setup, it might be interesting to check how different depth–to–width ratios affect convergence. Concretely, we can sweep a few ratios, repeat the kernel–profile plot from Experiment 2.1 at initialization and after training (with seed bands), and, in parallel, look at the function–space overlays. This would let us see whether changing the ratio has a noticeable impact on convergence behaviour.

Summary

- **Kernel profile replicated (Exp. 2.1):** The empirical NTK on the circle matches the paper's qualitative behaviour; variance across seeds shrinks with width and profiles are nearly time–invariant when wide.
- **Simple target matches NTK (Exp. 2.2):** For $f^*(x) = \frac{1}{2}\sin(2\gamma)$, finite-width MLPs essentially coincide with the analytic NTK predictor; RelErr is flat across widths.
- **Fourier mixture exposes frequency effects (Exp. 2.3–2.4):** With mixed frequencies, all widths initially settle on low harmonics; RelErr improvements with width are small unless capacity is very limited. Low-width sweeps reveal visible progression toward the NTK predictor as width increases.
- **Role of learning rate and training horizon (Exp. 2.5):** Increasing η and total steps (ηt) makes higher–frequency components emerge and reduces RelErr more clearly, indicating that discrete GD dynamics, not just width, govern the observed convergence.

References

Hanin, B. and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. (2019). On the spectral bias of neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Seleznova, M. and Kutyniok, G. (2022). Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.