Global Convergence and Better Spectral Bias in Low-Rank Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks often struggle to learn highly oscillatory functions at finite training time: low-frequency components are fitted first, while high-frequency modes can remain poorly recovered even when the model is expressive enough to represent them. Low-rank networks are usually introduced as compressed alternatives to dense models, but this view overlooks a more useful possibility: rank can act as a structural control on what the network learns first. In this paper, we show that this is the case. We first prove that low-rank random-feature networks in the mean-field limit converge to a global minimizer of the population risk whenever their limiting dynamics converge. We then show that the rank is not merely a compression parameter: choosing it correctly can reduce the number of trainable degrees of freedom while also improving the fit of highly oscillatory targets. The key practical message is that the best rank is typically intermediate. If the rank is too small, the model lacks expressivity; if it is too large, it recovers the finite-time bias of the dense model. Controlled geometric and Fourier diagnostics, together with high-frequency regression experiments, show that an appropriate low rank can lower test loss, improve high-frequency recovery, and that the optimal rank shifts with the target spectrum and training objective.

1 Introduction

Low-rank neural networks replace dense matrices by factored or bottlenecked operators. The standard motivation is computational: fewer trainable parameters and cheaper memory traffic. This paper argues that the more important question is mathematical. Can rank be used as a principled control on training dynamics rather than only as a compression parameter?

This question is important for scientific machine learning. Many tasks in AI for science require learning functions with oscillations, sharp transitions, multiscale structure, or high-frequency Fourier content: wave fields, PDE solution operators, molecular potentials, turbulent signals, and inverse problems all contain information at small spatial or temporal scales. Neural networks can represent such functions, but finite-time training often learns smooth, low-frequency components first. This frequency bias can be modified by the data distribution or by replacing the loss with a Sobolev norm [23]; our question is whether the architecture itself, through rank, can provide another control parameter. This makes the problem difficult: expressivity alone is not enough, and a model that is globally trainable in principle can still fail to recover the high-frequency part under a realistic training budget.

We answer the first question positively for a low-rank random-feature architecture. The model freezes a rich random-feature map and trains low-rank channel weights through mean-field gradient flow. Since the frozen features keep a dense span throughout training, the usual obstruction in deep mean-field convergence proofs is removed: at a limit point, zero gradient implies the conditional

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

first-order condition for the loss, and the loss-identifiability assumption forces global optimality. Low rank enters through the mixing matrix and channel index, changing constants but not the logic of the proof.

The second question is more subtle. Global convergence is an asymptotic statement: it says what happens if the population dynamics settle. Spectral bias is a finite-time phenomenon: it asks which parts of the target are reached first under a limited optimization budget. These two views can disagree on oscillatory targets, because a model may be globally trainable and still spend most of training on the low-frequency part of the signal. We show that rank changes this finite-budget behavior by changing the geometry of ReLU networks, hence which Fourier modes are recovered early. The practical outcome is a rank-selection rule: choose the smallest rank that avoids the approximation bottleneck while improving high-frequency recovery.

Contributions. The paper makes three contributions:

- We give a globally convergent anchor model for low-rank networks. Freezing a rich random-feature map keeps the feature span from collapsing, while the trainable low-rank channels still allow a mean-field feature-learning interpretation.
- We explain why rank can change high-frequency learning. In ReLU networks, lowering rank reduces the number of independent path constraints, which changes the piecewise-affine geometry that controls Fourier energy.
- We provide controlled diagnostics for choosing rank. The experiments show that the useful rank is usually intermediate and moves with the target spectrum or training objective, so rank should be selected from spectral recovery rather than parameter count alone.

2 Related Work

This section positions the paper relative to mean-field convergence, low-rank adaptation, bottleneck-rank theory, and spectral bias.

Mean-field convergence. Mean-field theory analyzes neural network training through the deterministic evolution of parameter distributions in the large-width limit. Foundational work established global convergence mechanisms for two-layer models and Wasserstein gradient flows under suitable convexity, homogeneity, or support assumptions [17, 5]. Later multilayer mean-field analyses clarified how deep feature maps evolve and when stationary points can be related to global optima [20, 19]. Our result follows this line but focuses on low-rank random-feature networks: the existence and uniqueness estimates use the same mean-field template, while the convergence proof must be adapted to channel-wise low-rank couplings.

Low-rank and random features. Low-rank neural networks are often motivated by parameter efficiency, with LoRA-style decompositions now standard in large-scale adaptation [15]. Recent theory has begun to explain why LoRA can be trainable: Jang, Lee, and Ryu [16] show that, in the NTK fine-tuning regime, sufficiently large LoRA rank removes spurious local minima and yields low-rank solutions with good generalization. Most theory for low-rank networks focuses on expressivity, approximation, kernel behavior, or LoRA fine-tuning rather than global mean-field training dynamics. Random features provide a tractable bridge between finite networks and kernel methods: fixing the feature map turns the representation into a stable basis, while training the channel weights retains a mean-field optimization problem. In our RF-LR setting, frozen random features prevent support collapse and preserve the dense-span property needed for the stationarity-to-optimality argument; the low-rank factors enter through bounded channel mixing rather than through a full dense contraction chain.

Bottleneck rank and implicit low-rank bias. Recent work gives a functional notion of optimal rank that is closer to our setting than the algebraic rank of a single matrix. Jacot [12] studies the implicit bias of large-depth homogeneous networks and shows that their representation cost converges toward a nonlinear rank notion, sandwiched between Jacobian rank and bottleneck rank. The bottleneck rank is the smallest inner dimension k for which a function can be factored as $f = h \circ g$ through $\mathbb{R}^k$. Follow-up work [13] proves that learned deep representations exhibit a bottleneck structure

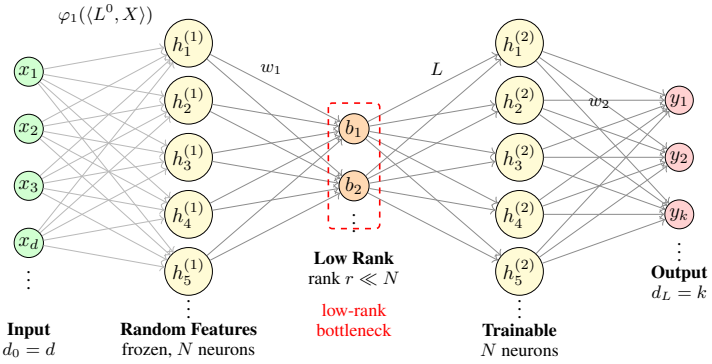

Figure 1: Architecture of a three-layer low-rank random-feature network with frozen random features, a rank-r bottleneck, and trainable output weights.

and introduces finite-depth corrections that balance low inner dimension against regularity. This suggests a theoretical target rank: the smallest rank that captures the intrinsic bottleneck dimension of the target without forcing an irregular or rank-underestimating interpolant. Complementarily, Bantzis, Simon, and Jacot [3] show that the first saddle escape of deep ReLU networks has a low-rank bias in deeper layers and propose saddle-to-saddle dynamics with increasing bottleneck rank. Our rank-selection principle is consistent with this picture: training should start from low effective rank, but useful learning requires increasing rank until approximation and spectral recovery saturate.

Spectral bias. The spectral-bias literature identifies why neural networks often fit low-frequency components before high-frequency components. Rahaman et al. [21] connect this learning order to Fourier properties of piecewise linear networks, while Yu, Yang, and Townsend [23] show that the bias can be tuned through nonuniform data and Sobolev-type losses. This paper keeps that Fourier viewpoint but changes the intervention: instead of reweighting data or losses, we study rank as an architectural parameter that changes the geometry seen by finite-time training.

3 Freezing Half the Weights Allows Global Convergence

This section defines the RF-LR mean-field model and states the conditional global convergence theorem.

Let $X \in \mathbb{R}^d$ and Y be drawn from a population distribution. The finite-width low-rank random-feature network used throughout the paper is the one from the original experiments:

$$h^{(0)}(x) = x, \qquad h^{(\ell)}(x) = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} w_j^{(\ell)} \varphi_\ell\Big(\langle L_j^{(\ell)}, h^{(\ell-1)}(x)\rangle + b_j^{(\ell)}\Big), \tag{1}$$

where the feature directions $L_j^{(\ell)}$ and biases $b_j^{(\ell)}$ are frozen, while the channel weights $w_j^{(\ell)}$ are trained. This is a low-rank factorization of a dense layer in which the left factor is fixed: if $W_\ell = U_\ell V_\ell^\top$ with $U_\ell \in \mathbb{R}^{n_\ell \times r_\ell}$ bounded and full column rank, then U_ℓ plays the role of a frozen random-feature or mixing basis and the trainable factor V_ℓ is absorbed into the channel weights. Thus one may avoid training U_ℓ entirely; drawing it as a bounded full-rank tall-skinny matrix is enough for the theory, while choosing $U_\ell \in \mathrm{St}(n_\ell, r_\ell)$ is an optional numerical normalization.

3.1 Mean-Field Formulation

The mean-field viewpoint is the relevant one for global convergence. Pure kernel or neural-tangent analyses give convexity by freezing the representation, but they largely remove feature learning. Finite-width nonconvex analyses allow feature learning but rarely provide global convergence guarantees for deep low-rank networks. Mean-field theory is currently the most developed framework that can keep a feature-learning interpretation while still proving that limiting gradient-flow trajectories reach global minimizers under identifiability assumptions. We use a frozen-feature low-rank specialization

because it is the cleanest setting in which the dense-span argument is preserved; it should be viewed as the globally convergent anchor case for more general low-rank feature-learning models.

In the mean-field limit, frozen feature maps are written $L^0(C_\ell)$ and trainable channel weights are written $w_\ell(t, C_\ell)$. The channel mixer $L \in \mathbb{R}^{N_2 \times r}$ is a bounded full-column-rank tall-skinny factor, not necessarily an orthogonal matrix. We write the second-layer operator as $W_2 = LA_2^\top$, with L fixed and A_2 trainable, then absorb the small right factor $A_2^\top$ into the rank-channel functions below. One may additionally choose $L \in \mathrm{St}(N_2, r)$, or periodically re-orthogonalize it, as a numerical design choice; this is natural for matrix-structured optimizers such as Muon, which orthogonalize matrix updates by Newton–Schulz or polar iterations [14, 4]. In the two-hidden-layer case, the first partial functions are

$$f_k(t, x) = \mathbb{E}_{C_1}\left[w_1(t, C_1, k)\, \varphi_1\left(\langle L^0(C_1), x\rangle\right)\right], \qquad k = 1, \dots, r, \tag{2}$$

and the second-layer preactivation is

$$H_2(t, c_2; x) = \sum_{k=1}^{r} L_{c_2,k} f_k(t, x). \tag{3}$$

Thus H_2 is the reconstruction of the second-layer signal from a bounded rank-r channel basis. The output is obtained by averaging the trainable top weights against $\varphi_2(H_2)$. For deeper networks, the same alternating pattern is used: trainable channel weights are averaged against frozen random features and mixed through bounded low-rank channel matrices.

The population objective is

$$\mathcal{L}(W) = \mathbb{E}_{(X,Y)}\left[\mathcal{L}\left(Y, \hat{y}(X; W)\right)\right]. \tag{4}$$

The mean-field gradient flow evolves the trainable weights by

$$\partial_t w_\ell(t, \cdot) = -\xi_\ell(t)\, \nabla_{w_\ell} \mathcal{L}(W(t)), \tag{5}$$

where ξ_ℓ are bounded learning-rate schedules. The exact formulas are standard backpropagation in the mean-field variables; the only low-rank modification is the channel mixing through $L_{c_\ell,k}$.

Under the technical regularity, diversity, loss-identifiability, and convergence assumptions stated in Appendix A, freezing the feature directions gives a well-defined mean-field system and allows the usual stationarity argument to conclude global optimality.

Theorem 3.1 (Well-posedness and conditional global convergence of RF-LR training)**.** *Under Assumptions A.1–A.6, the low-rank mean-field gradient-flow system has a unique solution on every finite interval $[0, T]$ and therefore on $[0, \infty)$. Moreover, if the resulting low-rank mean-field dynamics converges to $W^\star$ in the modified $\mathcal{W}_4$ coupling topology defined in Appendix A through Assumption A.6, then $W^\star$ is a global minimizer of the population loss:*

$$\mathcal{L}(W^\star) = \inf_W \mathcal{L}(W). \tag{6}$$

For every depth $L \geq 2$, this holds with standard independent initialization of the trainable weights.

The convergence statement is conditional: as in the corresponding multilayer mean-field results, we do not prove that every trajectory converges. The theorem says that convergence cannot occur to a bad stationary point under the stated dense-span, non-degeneracy, and loss-identifiability assumptions. The well-posedness part only justifies that the limiting dynamics are genuine ODEs rather than formal flows. This is the natural Picard argument of Nguyen and Pham [19]: nothing structural changes in the proof, except that dense-layer operator norms are replaced by bounded low-rank mixing constants. The complete well-posedness proof is deferred to the supplementary material, Appendix B.

Proof idea. The existence and uniqueness part follows the standard mean-field fixed-point argument, whose full details are given in Appendix B. We only summarize the convergence-to-global-minimum step here; Appendix C gives the full notation and the five formal steps. We assume that gradient descent, or its mean-field gradient-flow limit, has converged somewhere: $W(t) \to \bar{W}$ in the modified $\mathcal{W}_4$ coupling topology defined in Appendix A. Stationarity and the persistent dense span of the frozen first-layer features then imply the integrated identity

$$\mathbb{E}\Big[d_L(Z; \bar{W})\, B_k^{(2)}(X; \bar{W}) \mid X = x\Big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x, \quad k = 1, \dots, r, \tag{7}$$

where $d_L(Z;\bar{W}) = \partial_2\mathcal{L}(Y,\hat{y}(X;\bar{W}))$ and $B_k^{(2)}$ is the channel-wise backpropagated factor defined in Appendix C.

Step 3: remove the backpropagated factor. The non-degeneracy assumptions exclude a dead-channel limit: for $\mathbb{P}_X$-a.e. x, at least one channel has a nonzero backpropagated factor $B_k^{(2)}(x;\bar{W})$. The identity we use keeps this factor inside the conditional expectation:

$$\mathbb{E}\Big[d_L(Z;\bar{W})B_k^{(2)}(X;\bar{W}) \mid X = x\Big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x, \quad k = 1,\dots,r.$$

Since $B_k^{(2)}(X;\bar{W})$ is X-measurable, the non-degeneracy assumption then allows us to choose at least one nonzero backpropagated factor at almost every x, so

$$\mathbb{E}\big[\partial_2\mathcal{L}\left(Y,\hat{y}(X;\bar{W})\right) \mid X = x\big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x. \tag{8}$$

Step 5: the trajectory loss converges to the limit loss. Since $W(t) \to \bar{W}$ in the modified $\mathcal{W}_4$ coupling topology defined in Appendix A, there are couplings π_t between $W(t)$ and $\bar{W}$ whose weighted channel gaps vanish. In the two-layer channel notation, the key convergence statement is exactly

$$\mathbb{E}_{\pi_t}\left[(1+|\bar{w}_2|)|\bar{w}_2|\sum_{k=1}^{r}|\bar{w}_{1,k}|\,|w_{1,k}-\bar{w}_{1,k}|\right] \to 0, \tag{9}$$

with analogous quantities for all deeper layers. The forward stability estimate uses the ReLU channel expansion of a low-rank preactivation,

$$\left|\mathrm{ReLU}\left(\sum_{k=1}^{r}x_k\right) - \mathrm{ReLU}\left(\sum_{k=1}^{r}y_k\right)\right| \le \sum_{k=1}^{r}|x_k - y_k|. \tag{10}$$

Together with the low-rank form $H_2 = \sum_k L_{c_2,k}f_k$, (10) bounds every forward and backward difference by the same channel integrals. Hence

$$\mathbb{E}\left[\left|\hat{y}(X;W(t)) - \hat{y}(X;\bar{W})\right|\right] \to 0. \tag{11}$$

Lipschitzness of the loss in its second argument on the a priori bounded trajectory yields

$$|\mathcal{L}(W(t)) - \mathcal{L}(\bar{W})| \le K\,\mathbb{E}\left[\left|\hat{y}(X;W(t)) - \hat{y}(X;\bar{W})\right|\right] \to 0. \tag{12}$$

Step 4: identify the limit as globally optimal. By the loss-identifiability condition in Assumption A.3, namely (26) in Appendix A,

$$\mathbb{E}[\partial_2\mathcal{L}(Y,u) \mid X = x] = 0 \quad \Longrightarrow \quad \mathbb{E}[\mathcal{L}(Y,u) \mid X = x] = 0. \tag{13}$$

Applying this with $u = \hat{y}(x;\bar{W})$ and using non-negativity of $\mathcal{L}$ gives

$$\mathcal{L}(\bar{W}) = 0 = \inf_{W}\mathcal{L}(W). \tag{14}$$

Together with Step 5, $\mathcal{L}(W(t)) \to \mathcal{L}(\bar{W})$. □

4 Low-Rank Spectral Bias

This section explains how low-rank path constraints change CPWL geometry and the Fourier coefficients that control finite-resolution spectral bias.

Connection to spectral bias. Theorem 3.1 says that a convergent trajectory reaches a global minimizer, but it does not describe the order in which Fourier modes are recovered at finite time. This section explains the architectural mechanism behind that order: low rank constrains active path coefficients and switching geometry, so it changes the CPWL face moments that control finite-resolution Fourier transfer. Thus the role of Rahaman et al. [21] here is not only as prior related work, but as the Fourier template that our rank-dependent face-moment calculation refines.

In a ReLU network, on every activation region the network is affine, and the affine coefficient is a sum over active paths. We use the same path-product viewpoint as in path-based analyses of loss

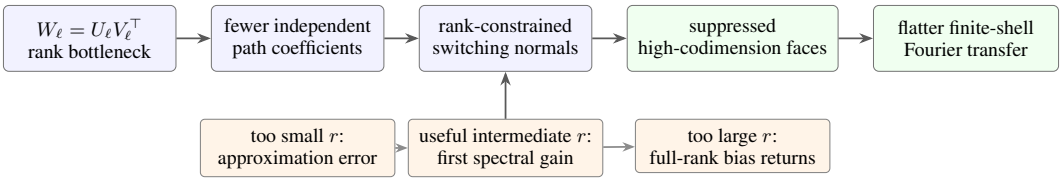

Figure 2: Mechanism summary linking low-rank factorization, path coefficients, switching normals, CPWL faces, and finite-shell Fourier transfer.

surfaces and finite-width neural kernels [6, 10]. For a scalar-output depth-L network, this can be written schematically as

$$f(x) = \sum_{p=(i_0,\dots,i_L)} \left(\prod_{\ell=1}^{L} W^{(\ell)}_{i_\ell i_{\ell-1}} \right) \left(\prod_{\ell=1}^{L-1} \mathbf{1}_{\{h^{(\ell)}_{i_\ell}(x)>0\}} \right) x_{i_0}, \tag{15}$$

where p ranges over input-to-output paths. If $W_\ell = U_\ell V_\ell^\top$ has rank r_ℓ, then many symbolic paths share the same latent channels, so their coefficients cannot vary independently. These path constraints also constrain ReLU switching normals $\nabla_x h^{(\ell)}_j(x)$: the normals lie in a smaller span than in a dense layer. Geometrically, this suppresses low-dimensional, high-codimension intersections; spectrally, the Fourier shell law below predicts more visible facet-dominated high-frequency mass.

Mechanism summary. Figure 2 summarizes the mechanism. A low-rank factorization reduces independent path coefficients in (15); this constrains switching normals, suppresses high-codimension CPWL faces, and moves shell-averaged Fourier energy toward the facet-dominated regime of Rahaman et al. [21]. The useful rank is intermediate: too small gives approximation error, while too large recovers the full-rank low-frequency-first bias. A frequency shell means modes with comparable radius $\|\xi\| \simeq \rho$.

This connects to bottleneck rank: the population lower bound is the smallest inner dimension through which the target can factor [12, 13], while the finite-budget choice also balances optimization and spectral recovery. In experiments, we estimate this by the first rank where approximation is controlled and Fourier recovery no longer improves.

This algebraic constraint becomes geometric. Inside a cell of the previous layers, a new switching normal has the form

$$n_{\ell i\varepsilon} = B^\top_{\ell-1,\varepsilon} V_\ell u_{\ell i} \in \mathrm{Im}(B^\top_{\ell-1,\varepsilon} V_\ell), \qquad \dim \mathrm{Im}(B^\top_{\ell-1,\varepsilon} V_\ell) \le \min(d, r_1, \dots, r_\ell). \tag{16}$$

Thus rank lowers the dimension of the normal subspace and suppresses high-codimension faces first.

Face Fourier expansion. Let $g = \chi f$, where f is CPWL on a finite polyhedral complex in $\mathbb{R}^d$ and $\chi \in C^\infty_c(\mathbb{R}^d)$ is an observation window. Let $\mathcal{F}_q(f)$ be the codimension-q faces of the complex. For every large frequency $\xi = \rho\theta$, $\theta \in \mathbb{S}^{d-1}$, the Fourier transform has the expansion

$$\widehat{g}(\rho\theta) = \sum_{q=1}^{d} \rho^{-(q+1)} \sum_{F \in \mathcal{F}_q(f)} A_F(\rho, \theta) + O(\rho^{-(d+2)}), \tag{17}$$

where A_F is a bounded oscillatory face coefficient controlled by the jump Δ_F of the appropriate normal derivative across the local fan around F. In particular,

$$|A_F(\rho, \theta)| \le C_\chi \|\Delta_F\| \, \mathrm{vol}_{d-q}(F)\, \omega_F(\theta). \tag{18}$$

This is the anisotropic phenomenon isolated by Rahaman et al. [21]: in almost all directions a ReLU network has fast $k^{-(d+1)}$ amplitude decay, while in special directions orthogonal to facets of the linear regions the decay can be as slow as k^{-2}. Thus low codimension faces are precisely the geometric structures that keep high-frequency Fourier mass visible at finite resolution.

Proposition 4.1 (Fourier shell law for CPWL networks)**.** *Let $g = \chi f$ be a windowed continuous piecewise affine network in dimension d. Define the codimension-q face moment*

$$M_q(f) = \sum_{F \in \mathcal{F}_q(f)} \|\Delta_F\|^2 \, \mathrm{vol}_{d-q}(F)^2 \mathbb{E}_{\theta \in \mathbb{S}^{d-1}} \omega_F(\theta)^2. \tag{19}$$

Then the shell-averaged Fourier energy satisfies

$$S_f(\rho) = \mathbb{E}_{\|\xi\|\simeq\rho}|\widehat{g}(\xi)|^2 \lesssim \sum_{q=1}^{d} M_q(f)\rho^{-2(q+1)}. \tag{20}$$

Low rank suppresses M_q for larger q first, moving finite-resolution spectra toward the facet-dominated ρ^{-4} energy regime. This should be read as a coefficient statement rather than a claim that low rank creates a new universal exponent: the useful comparison with full rank is the reduction of M_q/M_1 and of the transition coefficient $(M_q/M_1)\rho^{-2q+2}$ for $q \geq 2$.

Proof idea. Write f on its polyhedral complex as an affine function on each cell and set $g = \chi f$ with a smooth compactly supported window. The Fourier transform is then a sum of ordinary oscillatory integrals over cells. Integrating by parts in the direction $\theta = \xi/\|\xi\|$ produces boundary terms on cell facets. The order-ρ^{-1} traces cancel across interior facets because the CPWL network is continuous and the same window multiplies both sides.

The first non-canceling term is therefore the jump of the directional derivative across a facet, hence it carries the factor ρ^{-2}. Repeating the same integration-by-parts argument on the induced face complex localizes the next non-smooth terms on codimension-q faces, with amplitude order $\rho^{-(q+1)}$. This gives the face expansion (17), with coefficients controlled by the corresponding normal-derivative jumps and face volumes. This is the same anisotropic mechanism emphasized by Rahaman et al. [21]: generic directions decay quickly, while directions close to facet normals can retain slow $|\xi|^{-2}$ amplitude decay.

Squaring the expansion and averaging over a frequency shell gives diagonal face energies of order $M_q(f)\rho^{-2(q+1)}$. Oscillatory cross terms are either lower order by angular non-stationary phase or absorbed into the same moments by Cauchy–Schwarz. Thus the exponent is determined by codimension, while rank changes the coefficients $M_q(f)$ by suppressing high-codimension intersections. A detailed integration-by-parts proof is given in Appendix D. □

5 Experiments

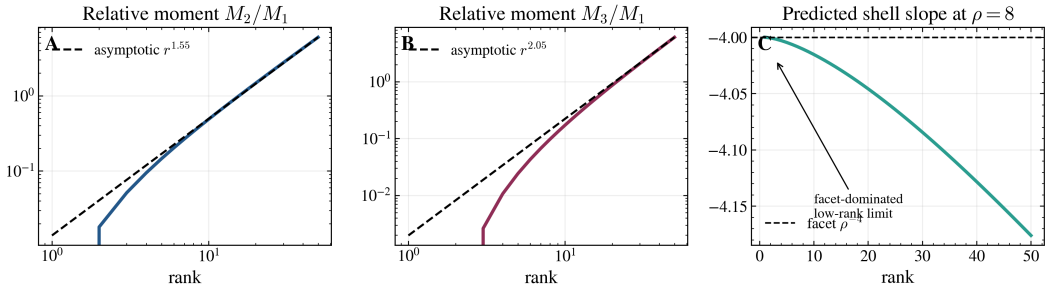

Figure 3: Theory proxy for the Fourier shell law. Panels A–B show relative face-moment proxies M_2/M_1 and M_3/M_1, i.e. how much higher-codimension geometry is present relative to facets. Panel C shows the induced finite-shell slope, separating the universal codimension exponents from the rank-dependent coefficients.

This section presents controlled diagnostics that test the proposed geometric and Fourier mechanism, together with supporting high-frequency regression evidence.

We keep the experimental evidence deliberately short and defer the full setup of every generated plot to Appendix E. The main figures are controlled CPWL and Fourier/kernel-limit diagnostics: they test the mechanism predicted by Proposition 4.1, namely that rank changes the face-moment coefficients $M_q(f)$ and the finite-time recovery of Fourier modes. Separate finite-network MMNN sweeps on high-frequency targets are used only as supporting evidence that low-rank models can fit oscillatory functions and that optimizer choice and feature freezing matter in practice.

The experiments should therefore be read as mechanism checks rather than as a benchmark claim that one optimizer or one rank universally dominates. The CPWL plots isolate geometry, the

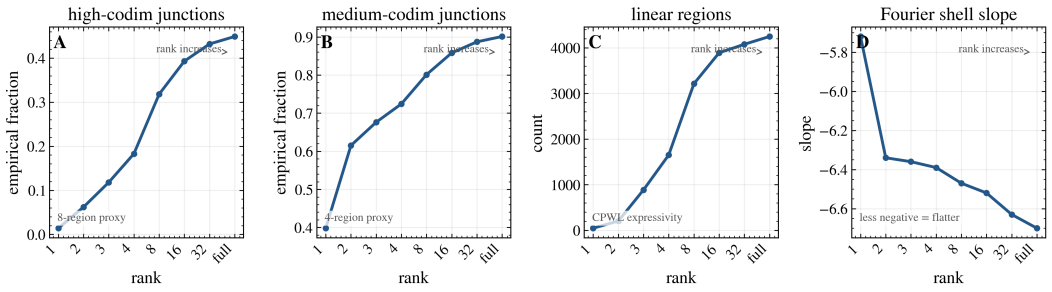

Figure 4: Three-dimensional CPWL geometry proxy. Panels A–B show the fraction of sampled neighborhoods where many affine regions meet, using thresholds of 8 and 4 regions as empirical high- and medium-codimension junction proxies. Panel C shows the number of linear regions, and Panel D shows the estimated shell slope.

Fourier/kernel plots isolate spectral transfer at fixed budget, and the rank-shift control tests whether the observed optimum is task-dependent. This separation is intentional: the theorem is an asymptotic convergence statement, whereas the spectral-bias claims concern finite-resolution coefficients and finite-time recovery.

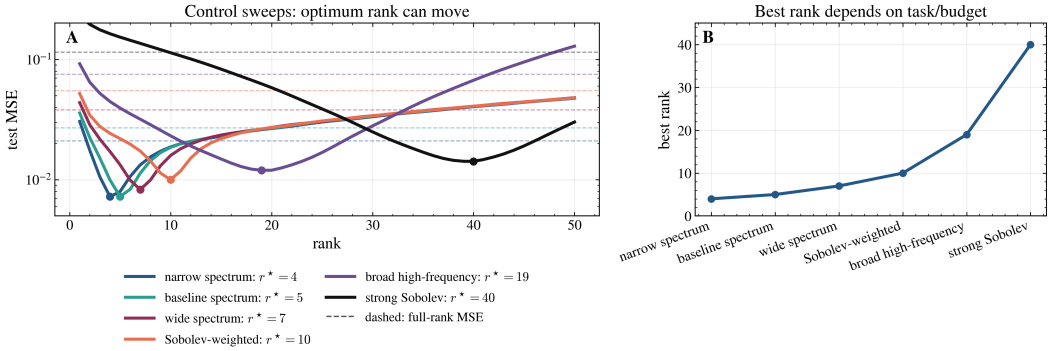

Figure 5: Rank-shift control. The left panel shows finite-budget MSE versus rank for several target/weighting regimes; the right panel shows the best rank in each regime. Dashed lines are full-rank baselines. The plot emphasizes that the useful rank moves with the spectrum and objective rather than staying fixed.

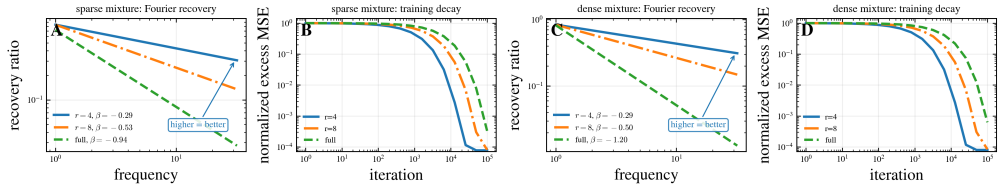

Figure 6: Mode-wise recovery and training curves. Panels A and C show Fourier recovery ratios $R_r(k,t)$ across frequency modes: higher and flatter curves mean that high frequencies are recovered more evenly relative to low frequencies. Panels B and D show normalized excess MSE during training, so the useful ranks are those that combine low error with flatter spectral recovery.

The one-dimensional recovery plots require a small caveat. A 1D ReLU network is a linear spline: if Δ_j denotes the derivative jump at knot t_j, then

$$\widehat{f}(k) = -\frac{1}{(ik)^2}\sum_j \Delta_j e^{-ikt_j} + O(|k|^{-3}), \qquad |\widehat{f}(k)|^2 \sim |k|^{-4}A(k). \tag{21}$$

Thus the raw asymptotic tail stays near the classical k^{-4} law, and the finite-budget diagnostic should be mode-wise recovery,

$$R_r(k,t) = \frac{|\widehat{f}_{r,t}(k)|}{|\widehat{f}^\star(k)|}, \qquad \beta_r(t) = \frac{d\log R_r(k,t)}{d\log k}. \tag{22}$$

A flatter fitted slope $\beta_r(t)$ means that high modes are recovered more evenly relative to low modes on the plotted frequency window. Figure 6 reports this recovery ratio together with training MSE.

The geometry plots should be read through the face-moment shell law

$$S_f(\rho) \lesssim M_1(f)\rho^{-4} + M_2(f)\rho^{-6} + M_3(f)\rho^{-8} + \cdots . \tag{23}$$

Figure 3 tests the coefficient prediction directly: higher-codimension moments grow with rank, so the finite-shell slope moves away from the facet-dominated regime. Figure 4 checks the same mechanism geometrically: high-codimension junctions are empirical proxies for the higher moments $M_2, M_3, \ldots$, and they become more frequent as rank creates more independent switching geometry. Appendix E gives the full interpretation.

The spectral-bias experiment isolates the rank effect in a width-2^{15} Fourier/kernel surrogate. Figure 5 is the key control: the best finite-budget rank is intermediate, but it is not fixed at one value. It moves from small ranks on narrow or baseline spectra to larger ranks under wider spectra and Sobolev-weighted objectives used to tune frequency bias [23]. Across these regimes, the best low-rank point also beats the corresponding full-rank dashed baseline. Figure 6 shows why MSE alone is not enough: the useful rank also keeps high-frequency recovery flatter than the full endpoint.

Together, Figures 3–6 support the same conclusion: rank is not only a compression parameter. It changes the finite-budget spectral transfer, and the best rank is the first one that controls approximation while preserving flatter high-frequency recovery.

6 Conclusion

Low rank is not only a compression device. In the random-feature mean-field regime, it preserves global convergence because frozen features maintain dense span and low-rank channel mixing only changes constants. In ReLU and CPWL regimes, it also changes spectral bias through path-space constraints and Fourier face moments. This matters for AI for science because oscillatory targets are common in PDEs, waves, inverse problems, and molecular or multiscale modeling. Prior work shows that frequency bias can be tuned by data or Sobolev losses [23]; our contribution is to show that architecture provides another lever. The message is conservative but actionable: use the frozen-feature low-rank model as a globally convergent anchor, then choose rank by finite-budget spectral recovery rather than by parameter count alone.

The mechanism is the following. A globally convergent dynamic can still learn low frequencies first and fail at finite time on highly oscillatory targets. Low rank changes this finite-time behavior because, in ReLU networks, it constrains active path coefficient tensors; these constraints restrict switching normals, suppress high-codimension faces in the continuous piecewise affine complex, and change Fourier shell behavior. Here CPWL means continuous piecewise affine: the input space is split into polyhedral regions, and the network is affine on each region. In one dimension, the raw spline tail remains close to the classical k^{-4} law, so the right diagnostic is not the asymptotic exponent alone but the learned Fourier recovery curve. This leads to the practical rule used throughout the experiments: choose the smallest rank that avoids the approximation bottleneck while flattening the finite-time spectral transfer.

Several directions remain open. First, the convergence result should be extended from frozen random features to trainable low-rank factors, where the dense-span argument can fail if the features collapse. Second, the CPWL shell law suggests a measurable rank-dependent coefficient theory, but sharper finite-width estimates are needed to predict the best rank without a sweep. Third, the experiments indicate that optimizer geometry matters, especially for Muon-type orthogonalized updates; understanding how such updates interact with low-rank spectral transfer is a natural next step.

References

[1] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. arXiv:1412.8690, https://arxiv.org/abs/1412.8690.

[2] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi:10.1109/18.256500.

[3] I. Bantzis, J. B. Simon, and A. Jacot. Saddle-to-saddle dynamics in deep ReLU networks: low-rank bias in the first saddle escape. arXiv preprint arXiv:2505.21722, 2026. https://arxiv.org/abs/2505.21722.

[4] V. Boreiko, Z. Bu, and S. Zha. Towards understanding orthogonalization in Muon. In *HiLD Workshop at ICML*, 2025. OpenReview: https://openreview.net/forum?id=ppmyFtr9EW.

[5] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3036–3046, 2018. arXiv:1805.09545, https://arxiv.org/abs/1805.09545.

[6] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, PMLR 38:192–204, 2015. arXiv:1412.0233, https://arxiv.org/abs/1412.0233.

[7] W. Czarnecki, S. Osindero, M. Jaderberg, G. Swirszcz, and R. Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems 30*, pages 4278–4287, 2017. arXiv:1706.04859, https://arxiv.org/abs/1706.04859.

[8] R. Diaz, Q.-N. Le, and S. Robins. Fourier transforms of polytopes, solid angle sums, and discrete volume. arXiv preprint arXiv:1602.08593, 2018. https://arxiv.org/abs/1602.08593.

[9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, PMLR 9:249–256, 2010. https://proceedings.mlr.press/v9/glorot10a.html.

[10] B. Hanin and M. Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. arXiv:1909.05989, https://arxiv.org/abs/1909.05989.

[11] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, 2018. arXiv:1806.07572, https://arxiv.org/abs/1806.07572.

[12] A. Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. arXiv preprint arXiv:2209.15055, 2023. https://arxiv.org/abs/2209.15055.

[13] A. Jacot. Bottleneck structure in learned features: low-dimension vs regularity tradeoff. arXiv preprint arXiv:2305.19008, 2024. https://arxiv.org/abs/2305.19008.

[14] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, 2024. https://kellerjordan.github.io/posts/muon/.

[15] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685, https://arxiv.org/abs/2106.09685.

[16] U. Jang, J. D. Lee, and E. K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:21306–21328, 2024. arXiv:2402.11867, https://arxiv.org/abs/2402.11867.

[17] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi:10.1073/pnas.1806579115; arXiv:1804.06561.

[18] G. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems 27*, pages 2924–2932, 2014. arXiv:1402.1869, `https://arxiv.org/abs/1402.1869`.

[19] P.-M. Nguyen and H. T. Pham. A rigorous framework for the mean-field limit of multilayer neural networks. *Mathematical Statistics and Learning*, 6(3):201–357, 2023. arXiv:2001.11443, `https://arxiv.org/abs/2001.11443`.

[20] H. T. Pham and P.-M. Nguyen. Global convergence of three-layer neural networks in the mean field regime. In *International Conference on Learning Representations*, 2021. arXiv:2105.05228, `https://arxiv.org/abs/2105.05228`.

[21] 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 *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:5301–5310, 2019. arXiv:1806.08734, `https://arxiv.org/abs/1806.08734`.

[22] T. Zaslavsky. Facing up to arrangements: face-count formulas for partitions of space by hyperplanes. *Memoirs of the American Mathematical Society*, 1(154):1–102, 1975. doi:10.1090/memo/0154.

[23] A. Yu, Y. Yang, and A. Townsend. Tuning frequency bias in neural network training with nonuniform data. arXiv preprint arXiv:2205.14300, 2022. `https://arxiv.org/abs/2205.14300`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the conditional nature of the convergence result, the architectural scope, and the finite-budget spectral-bias claim. The experiments are described as controlled diagnostics rather than universal benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The conclusion and appendix discuss the conditional convergence assumption, the frozen-feature restriction, the proxy nature of the diagnostics, and the need for stronger finite-width predictions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The main theorem states the assumptions by reference, gives a proof sketch, and the appendix contains the formal assumptions, well-posedness proof, global convergence proof, and CPWL Fourier shell-law proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The main paper and Appendix E describe the controlled CPWL/Fourier diagnostics, target families, rank sweeps, and proxy formulas. The supplemental zip contains the figure-generation scripts and generated figures.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The submission bundle includes anonymized source code and plotting scripts sufficient to reproduce the paper figures. No proprietary data or private datasets are used.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 5 explains how to read the experiments, and Appendix E gives the frequency grids, rank regimes, target spectra, proxy formulas, and training-curve definitions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The main figures are deterministic controlled diagnostics rather than stochastic benchmark averages over random train/test splits. Where visual uncertainty bands appear, the appendix states that they are display bands for readability rather than statistical confidence intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix E states that the reported diagnostics are lightweight deterministic CPU computations and gives the finite-network sweep settings separately. The supplemental bundle includes scripts for reproducing the generated figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical and uses synthetic or deterministic diagnostic data. It does not involve human subjects, private data, surveillance, or sensitive attributes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: The introduction and conclusion motivate potential positive impact for AI for science and high-frequency scientific modeling. The work is foundational and does not introduce a direct deployment pathway or dataset with obvious privacy or fairness risks.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a high-risk pretrained model, scraped dataset, or dual-use generation system. The released assets are source files, plotting code, and synthetic deterministic diagnostics.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites prior scientific works and uses standard open LaTeX/template assets. The experiments use synthetic data generated by the authors rather than existing external datasets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: The supplemental bundle documents the new generated figures and scripts. The assets are anonymized for submission and contain no personal or sensitive information.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: No human-subject research is conducted, so IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core methods, theory, experiments, and diagnostics do not rely on an LLM as an original or non-standard methodological component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Assumptions and Dense-Span Property

This section records the assumptions used by the convergence theorem and fixes the notation for the mean-field limit.

We collect the assumptions used in Theorem 3.1. We reuse the multilayer mean-field framework and notation of Nguyen and Pham [19]: the mean-field width limit sends $n_1, \dots, n_{L-1} \to \infty$ at fixed depth L, and discrete neuron indices are replaced by labels $C_\ell \in \Omega_\ell$ with laws ρ_ℓ. Our only architectural change is that dense contractions are replaced by frozen random features and bounded low-rank channel mixers. The notation $\partial_2 \mathcal{L}(y, u)$ denotes the derivative of the loss with respect to its prediction argument.

Assumption A.1 (Bounded activations and low-rank mixing)**.** *There exists $K \geq 1$ such that the activations φ_ℓ are K-Lipschitz, the activation derivatives used in backpropagation are bounded on the active region, and the low-rank mixers satisfy*

$$\sup_{c_\ell} \sum_{k=1}^{r_\ell} |L^{(\ell)}_{c_\ell, k}| \leq r_\ell K. \tag{24}$$

The mixers are full-column-rank tall-skinny factors in finite width. Orthogonality, $L^{(\ell)\top} L^{(\ell)} = I_{r_\ell}$, is an optional Stiefel normalization rather than a theorem assumption. The preactivations remain in the a priori bounded regime on every finite interval, and the learning-rate schedules are non-negative, bounded, and locally integrable.

Assumption A.2 (Sub-Gaussian initialization)**.** *The trainable initial weights have sub-Gaussian tails, uniformly over channels. In the two-hidden-layer notation,*

$$\sup_{m \geq 1} \frac{1}{\sqrt{m}} \max_{1 \leq k \leq r} \mathbb{E}_{C_1} \big[|w_1^0(C_1, k)|^m \big]^{1/m} \leq K, \qquad \sup_{m \geq 1} \frac{1}{\sqrt{m}} \mathbb{E}_{C_2} \big[|w_2^0(C_2)|^m \big]^{1/m} \leq K. \tag{25}$$

The L-layer version imposes the same moment growth bound on every trainable channel. This assumption includes the standard initializations used in practice. In particular, Xavier or Glorot uniform initialization [9], with entries sampled independently from

$$\mathrm{Unif}\left[-\sqrt{\frac{6}{d_{\mathrm{in}} + d_{\mathrm{out}}}}, \sqrt{\frac{6}{d_{\mathrm{in}} + d_{\mathrm{out}}}} \right],$$

is bounded and therefore sub-Gaussian after the usual fan-in/fan-out scaling. Xavier normal initialization, and more generally any independent centered sub-Gaussian initialization with the same variance scale, also satisfies the displayed moment bound. Thus the phrase "standard independent initialization" in Theorem 3.1 covers Glorot/Xavier uniform, Glorot/Xavier normal, and the usual sub-Gaussian or subnormal variants used for trainable weights.

Assumption A.3 (Data distribution and loss)**.** *Inputs are bounded, $|X| \leq K$ almost surely, and the frozen first-layer features satisfy $\|L^0(c_1)\| \leq K$. The loss is non-negative, and $\partial_2 \mathcal{L}(y, \cdot)$ is bounded and Lipschitz on the relevant prediction range. Moreover,*

$$\mathbb{E}[\partial_2 \mathcal{L}(Y, u) \mid X = x] = 0 \quad \Longrightarrow \quad \mathbb{E}[\mathcal{L}(Y, u) \mid X = x] = 0 \tag{26}$$

for $\mathbb{P}_X$-almost every x. This condition is the usual realizable or excess-risk identifiability condition. For squared loss, it holds in the noiseless realizable case, where $Y = f^\star(X)$ and $\partial_2 (Y - u)^2 = 0$ implies $u = Y$ and therefore zero conditional loss; equivalently, in the noisy case it holds after replacing the raw squared loss by its excess risk above the Bayes regressor. For cross-entropy, the same statement holds when the loss is written as the excess cross-entropy, or conditional KL divergence, relative to the Bayes conditional label distribution; in the deterministic-label realizable case this again reduces to zero conditional loss at the correct classifier.

Assumption A.4 (Diversity of frozen random features)**.** *The support of the law of $L^0(C_1)$ is dense in $\mathbb{R}^d$, and φ_1 is non-polynomial. Equivalently,*

$$\big\{ \varphi_1\big(\langle L^0(c_1), \cdot \rangle \big) : c_1 \in \Omega_1 \big\} \tag{27}$$

has dense span in $L^2(\mathbb{P}_X)$.

Assumption A.5 (Non-degenerate limit)**.** *The limiting representation is not a dead network:*

$$\max_{1 \leq k \leq r} \mathbb{P}(\bar{w}_1(C_1, k) \neq 0) > 0, \qquad \mathbb{P}(\bar{w}_\ell(C_\ell) \neq 0) > 0, \quad \ell = 2, \dots, L-1. \tag{28}$$

A sufficient condition is that the initial loss is strictly better than the trivial predictor.

Assumption A.6 (Convergence in a modified $\mathcal{W}_4$ coupling topology)**.** *The mean-field trajectory has a limit $\bar{W}$ in a modified $\mathcal{W}_4$ topology. This is a Wasserstein-4-type coupling topology, with the fourth-moment control replaced by the weighted channel quantities that appear in the low-rank ODE and in the output-stability estimates. In the two-hidden-layer notation, there exist couplings π_t such that*

$$\int (1+|\bar{w}_2(c_2)|)|\bar{w}_2(c_2)| \max_{1\le k\le r} |\bar{w}_1(c_1,k)||w_1(t,c_1',k)-\bar{w}_1(c_1,k)|\, d\pi_t \to 0, \tag{29}$$

$$\int (1+|\bar{w}_2(c_2)|)|\bar{w}_2(c_2)||w_2(t,c_2')-\bar{w}_2(c_2)|\, d\pi_t \to 0. \tag{30}$$

For depth $L>3$, the assumption includes the analogous weighted coupling gaps for all trainable layers.

Theorem A.7 (Universal approximation automatically maintained)**.** *Under Assumption A.4, the frozen first-layer class has dense span in $L^2(\mathbb{P}_X)$ throughout training.*

Proof. This is the classical non-polynomial random-feature density theorem. Since the first feature map is frozen, its support cannot collapse during training. □

B Supplementary Proof of Well-Posedness

This section gives the full Picard fixed-point proof that the low-rank mean-field ODE is well defined on every finite time interval.

We give the complete argument for the existence and uniqueness part of Theorem 3.1. The proof is deliberately close to the Picard proof for multilayer mean-field networks in Nguyen and Pham [19]. The low-rank architecture does not introduce a new analytic difficulty: every place where the full-rank proof uses a dense-layer operator norm, we use the bounded row-sum constant

$$\Lambda_\ell := \sup_{c_\ell} \sum_{k=1}^{r_\ell} |L^{(\ell)}_{c_\ell,k}| \le r_\ell K. \tag{31}$$

Thus the proof is a natural specialization of the existing mean-field Picard argument to bounded low-rank mixing.

For clarity we write the proof in the two-hidden-layer notation. The extension to arbitrary fixed depth is obtained by repeating the same estimates layer by layer and replacing Λ_2 by $\max_\ell \Lambda_\ell$ in the constants. Recall

$$f_k(t,x) = \mathbb{E}_{C_1}\left[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)\right], \qquad H_2(t,c_2;x) = \sum_{k=1}^{r} L_{c_2,k} f_k(t,x), \tag{32}$$

and

$$\hat{y}(x;W(t)) = \mathbb{E}_{C_2}[w_2(t,C_2)\varphi_2(H_2(t,C_2;x))]\,. \tag{33}$$

Let $G(W)$ denote the right-hand side of the mean-field ODE. In integral form, a solution is a fixed point of

$$F(W)(t) = W(0) + \int_0^t G(W(s))\, ds. \tag{34}$$

Step 1: low-rank forward Lipschitz estimates. Let W' and W'' be two trajectories on $[0,T]$. Since φ_1 is bounded on the relevant input range and $|X|, \|L^0(C_1)\| \le K$,

$$|f_k'(t,x) - f_k''(t,x)| \le K\,\mathbb{E}_{C_1}|w_1'(t,C_1,k) - w_1''(t,C_1,k)|. \tag{35}$$

Consequently the low-rank preactivation satisfies

$$\begin{aligned} |H_2'(t,c_2;x) - H_2''(t,c_2;x)| &\le \sum_{k=1}^{r} |L_{c_2,k}|\,|f_k'(t,x) - f_k''(t,x)| \\ &\le K\Lambda_2 \max_{1\le k\le r} \mathbb{E}_{C_1}|w_1'(t,C_1,k) - w_1''(t,C_1,k)|. \end{aligned} \tag{36}$$

This is the only low-rank modification of the dense proof. The dense operator norm is simply replaced by Λ_2.

Step 2: drift Lipschitz estimates on bounded sets. Fix $T < \infty$ and consider the class $\mathcal{W}_T^0$ of trajectories with the same initialization, bounded fourth moments, and sub-Gaussian tails on $[0, T]$. More explicitly, there is a finite constant $K_0(T)$ such that, for every $W \in \mathcal{W}_T^0$,

$$\sup_{t \le T} \left(\max_k \mathbb{E}_{C_1} |w_1(t, C_1, k)|^4 + \mathbb{E}_{C_2} |w_2(t, C_2)|^4 \right)^{1/4} \le K_0(T), \tag{37}$$

with the corresponding uniform sub-Gaussian tail bound inherited from the initialization and the bounded drift. On this class, the loss derivative, activations, activation derivatives, and low-rank mixers are bounded or Lipschitz by Assumptions A.1–A.3. Combining these bounds with (36) gives, on the good event where the channel weights are bounded by $K_0(T)B$,

$$\|G(W') - G(W'')\|_t \le C_T(1 + B)\|W' - W''\|_t. \tag{38}$$

Here C_T depends on K, T, and the low-rank constants Λ_ℓ, but not on width. The complement of the good event is controlled by the sub-Gaussian tail bound, hence contributes at most $C_T e^{-cB^2}$. Therefore

$$\|F(W') - F(W'')\|_t \le C_T(1 + B) \int_0^t \|W' - W''\|_s \, ds + C_T e^{-cB^2}. \tag{39}$$

This is the same estimate used in Nguyen and Pham [19]; only the constant C_T changes through (31).

Step 3: invariance of the Picard map. The map F sends $\mathcal{W}_T^0$ into itself. Indeed, the integral formula (34), the boundedness of $\partial_2 \mathcal{L}$, the boundedness of the activations on the a priori regime, and the low-rank estimate

$$|H_2(t, c_2; x)| \le \Lambda_2 \max_{1 \le k \le r} |f_k(t, x)| \le K\Lambda_2 \max_{1 \le k \le r} \mathbb{E}_{C_1} |w_1(t, C_1, k)| \tag{40}$$

imply a Gronwall bound for the fourth moments on $[0, T]$. The same argument applied to exponential moments gives the sub-Gaussian tail bound. Thus, after increasing $K_0(T)$ if necessary, $F(W) \in \mathcal{W}_T^0$ whenever $W \in \mathcal{W}_T^0$.

Step 4: Picard convergence. Iterating (39) yields, for all $m \ge 1$,

$$\|F^m(W') - F^m(W'')\|_T \le \frac{(C_T T(1 + B))^m}{m!} \|W' - W''\|_T + C_T \exp(C_T T(1 + B) - cB^2). \tag{41}$$

Choose $B = \sqrt{m}$. The first term tends to zero by Stirling's formula and the second tends to zero exponentially. Taking $W'' = F(W')$, the series

$$\sum_{m \ge 1} \|F^{m+1}(W') - F^m(W')\|_T \tag{42}$$

is finite, so the Picard iterates converge in $\|\cdot\|_T$ to a limit W. Since F is continuous under the same estimate, $F(W) = W$, hence W solves the mean-field ODE on $[0, T]$.

Step 5: uniqueness and continuation. If W' and W'' are two fixed points of F in $\mathcal{W}_T^0$, then applying the iterated estimate to the fixed points gives

$$\|W' - W''\|_T = \|F^m(W') - F^m(W'')\|_T \to 0 \qquad (m \to \infty). \tag{43}$$

Thus the solution is unique on $[0, T]$. Since $T < \infty$ was arbitrary and the a priori bounds are finite on every finite interval, the solution extends uniquely to $[0, \infty)$. This proves the well-posedness part of Theorem 3.1.

C Proof Details for Global Convergence

This section expands the stationarity, dense-span, non-degeneracy, and coupling arguments behind the global convergence theorem. We write the proof in the two-hidden-layer notation to keep the formulas readable. The same argument extends to any fixed depth because the decisive step is the frozen first layer: the class $\{\varphi_1(\langle L^0(c_1), \cdot \rangle) : c_1 \in \Omega_1\}$ keeps dense span throughout training, so the stationarity identity can always be pushed back to this unchanged feature family. Additional hidden layers only add backpropagated scalar factors and low-rank mixing constants; they do not change the dense-span-to-optimality mechanism.

Proof of Theorem 3.1. Let $Z = (X, Y)$ and write

$$d_L(Z; W) = \partial_2 \mathcal{L}(Y, \hat{y}(X; W)). \tag{44}$$

In the two-hidden-layer notation,

$$f_k(t, x) = \mathbb{E}_{C_1}\left[w_1(t, C_1, k)\, \varphi_1(\langle L^0(C_1), x\rangle)\right], \qquad H_2(t, c_2; x) = \sum_{k=1}^{r} L_{c_2,k} f_k(t, x), \tag{45}$$

and

$$\hat{y}(x; W(t)) = \mathbb{E}_{C_2}[w_2(t, C_2)\, \varphi_2(H_2(t, C_2; x))]\,. \tag{46}$$

The well-posedness part is proved in Appendix B. It is exactly the natural Picard argument of Nguyen and Pham [19], with dense-layer operator norms replaced by the bounded low-rank mixing constants. The dense-span property is preserved because $L^0(C_1)$ is never trained.

Let $W(t) \to \bar{W}$ in the modified $\mathcal{W}_4$ coupling topology of Assumption A.6. At the limit, stationarity of the first trainable low-rank channel gives, for every c_1 and k,

$$\mathbb{E}_Z\Big[d_L(Z; \bar{W})\, \varphi_1(\langle L^0(c_1), X\rangle)\, B_k^{(2)}(X; \bar{W})\Big] = 0, \tag{47}$$

where

$$B_k^{(2)}(x; \bar{W}) := \mathbb{E}_{C_2}\left[L_{C_2,k}\, \varphi_2'(H_2(C_2; x, \bar{W}))\, \bar{w}_2(C_2)\right]. \tag{48}$$

For deeper networks $B_k^{(\ell)}$ is the analogous backpropagated scalar. Conditioning on X and using the dense-span property implies

$$\mathbb{E}\Big[d_L(Z; \bar{W})\, B_k^{(2)}(X; \bar{W}) \mid X = x\Big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x, \quad k = 1, \dots, r. \tag{49}$$

The conditional identity is kept in the form

$$\mathbb{E}\Big[d_L(Z; \bar{W})\, B_k^{(2)}(X; \bar{W}) \mid X = x\Big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x, \quad k = 1, \dots, r. \tag{50}$$

Since $B_k^{(2)}(X; \bar{W})$ is X-measurable and the non-degeneracy assumption gives at least one nonzero backpropagated factor almost everywhere, (50) implies

$$\mathbb{E}\big[\partial_2 \mathcal{L}(Y, \hat{y}(X; \bar{W})) \mid X = x\big] = 0 \quad \text{for } \mathbb{P}_X\text{-a.e. } x. \tag{51}$$

Assumption A.3, through the exact loss-identifiability implication (26), then gives $\mathbb{E}[\mathcal{L}(Y, \hat{y}(X; \bar{W})) \mid X = x] = 0$, so $\mathcal{L}(\bar{W}) = 0$. Since $\mathcal{L} \geq 0$, $\bar{W}$ is a global minimizer.

It remains to connect the trajectory to the limit. A representative coupling gap is

$$\mathbb{E}_{\pi_t}\left[(1 + |\bar{w}_2|)\, |\bar{w}_2| \sum_{k=1}^{r} |\bar{w}_{1,k}|\, |w_{1,k}(t) - \bar{w}_{1,k}|\right] \longrightarrow 0. \tag{52}$$

The low-rank form gives

$$H_2(c_2; x, W(t)) - H_2(c_2; x, \bar{W}) = \sum_{k=1}^{r} L_{c_2,k}\left(f_k(t, x) - \bar{f}_k(x)\right). \tag{53}$$

Using Lipschitz activations and iterating through layers,

$$\mathbb{E}_Z\big[|\hat{y}(X; W(t)) - \hat{y}(X; \bar{W})|\big] \leq K\, \Gamma_t \longrightarrow 0, \tag{54}$$

where Γ_t is the sum of the weighted coupling gaps. Lipschitzness of the loss yields $\mathcal{L}(W(t)) \to \mathcal{L}(\bar{W})$. □

D Detailed Proof of the CPWL Fourier Shell Law

This section proves the Fourier shell law by integrating by parts over the face lattice of a windowed CPWL function.

We prove Proposition 4.1. Write the CPWL network on its polyhedral complex as

$$f(x) = \sum_{\varepsilon} (a_\varepsilon^\top x + b_\varepsilon) \mathbf{1}_{P_\varepsilon}(x), \qquad g(x) = \chi(x) f(x). \tag{55}$$

The cutoff makes g compactly supported, so

$$\widehat{g}(\rho\theta) = \sum_{\varepsilon} \int_{P_\varepsilon} e^{-i\rho\langle\theta,x\rangle} \chi(x)(a_\varepsilon^\top x + b_\varepsilon)\, dx \tag{56}$$

is an ordinary oscillatory integral. Set $D_\theta = \theta \cdot \nabla$. Since $D_\theta e^{-i\rho\langle\theta,x\rangle} = -i\rho e^{-i\rho\langle\theta,x\rangle}$, the divergence theorem on each cell gives

$$\begin{aligned} \int_{P_\varepsilon} e^{-i\rho\langle\theta,x\rangle} u_\varepsilon(x)\, dx = \frac{i}{\rho} \int_{\partial P_\varepsilon} e^{-i\rho\langle\theta,x\rangle} u_\varepsilon(x) \langle\theta, n_\varepsilon(x)\rangle\, d\sigma(x) \\ - \frac{i}{\rho} \int_{P_\varepsilon} e^{-i\rho\langle\theta,x\rangle} D_\theta u_\varepsilon(x)\, dx, \end{aligned} \tag{57}$$

where $u_\varepsilon(x) = \chi(x)(a_\varepsilon^\top x + b_\varepsilon)$ and n_ε is the outward normal. When the boundary terms are summed over all cells, every interior facet $F = P_\varepsilon \cap P_{\varepsilon'}$ is counted twice with opposite normals. The order-ρ^{-1} trace contribution is proportional to

$$(u_\varepsilon|_F - u_{\varepsilon'}|_F)\, \langle\theta, n_F\rangle. \tag{58}$$

Because f is continuous and the same smooth window χ multiplies both traces, this term vanishes on internal facets. Exterior window terms are smooth cutoff terms and are absorbed in the final remainder.

The leading non-smooth contribution comes from applying the same identity to the integral involving $D_\theta u_\varepsilon$. On a facet F shared by two cells, its jump is

$$\big[D_\theta u\big]_F = \chi\, \theta^\top (a_\varepsilon - a_{\varepsilon'}) \tag{59}$$

because the derivatives of χ multiply the continuous trace of f and cancel across F. Thus the first non-canceling boundary term is a jump of directional derivatives and carries the factor ρ^{-2}.

Higher-codimension terms follow by induction on the face lattice. Assume that after m reductions the surviving terms are sums over codimension-m faces E of the form

$$\rho^{-(m+1)} \int_E e^{-i\rho\langle\theta,x\rangle} B_E^{(m)}(\theta, x)\, d\sigma_E(x), \tag{60}$$

where $B_E^{(m)}$ is a linear combination of normal-derivative jumps in the local fan around E, multiplied by derivatives of χ of bounded order. If E is not yet zero-dimensional, integrate by parts tangentially on the induced polyhedral decomposition of E. Tangential interior terms either gain another factor ρ^{-1} or cancel between adjacent induced cells. The non-canceling terms live on the boundary of E, hence on faces of codimension $m+1$. Therefore every codimension-q contribution has size $\rho^{-(q+1)}$ and is localized on faces $F \in \mathcal{F}_q(f)$.

The remaining integral over a codimension-q face has the form

$$A_F(\rho, \theta) = \int_F e^{-i\rho\langle\theta,x\rangle} B_F(\theta, x)\, d\sigma_F(x), \tag{61}$$

where B_F is linear in the corresponding jump Δ_F and in derivatives of the window. Hence

$$|A_F(\rho, \theta)| \le C_\chi \|\Delta_F\| \operatorname{vol}_{d-q}(F) \omega_F(\theta). \tag{62}$$

The induction stops after codimension d, and the remaining smooth terms are $O(\rho^{-(d+2)})$ by one additional integration by parts. This proves the face expansion (17).

Squaring the face expansion gives diagonal face terms and oscillatory cross terms. The diagonal contribution of codimension-q faces is bounded by

$$\rho^{-2(q+1)} \sum_{F \in \mathcal{F}_q(f)} \|\Delta_F\|^2 \operatorname{vol}_{d-q}(F)^2 \omega_F(\theta)^2. \tag{63}$$

After averaging over a shell, non-aligned cross terms are lower order by non-stationary phase, while aligned terms are absorbed by

$$2|A_F A_{F'}| \le |A_F|^2 + |A_{F'}|^2. \tag{64}$$

Taking the angular expectation yields

$$S_f(\rho) \lesssim \sum_{q=1}^{d} M_q(f) \rho^{-2(q+1)}. \tag{65}$$

Thus the exponent $2(q+1)$ comes from face codimension, while low rank changes the coefficients $M_q(f)$ by reducing high-codimension intersections.

E Additional Experimental Details

This section specifies the controlled diagnostics and high-frequency sweeps used to support the spectral-bias mechanism.

All spectral-bias figures are generated from deterministic Fourier/kernel-limit and CPWL geometry diagnostics. The purpose is not to report a single trained finite network, but to isolate the rank-dependent quantities predicted by the CPWL shell law. Unless stated otherwise, rank is swept over $r = 1, \dots, 50$ and the full-rank endpoint is represented by the dense transfer limit. These diagnostics are lightweight CPU computations: the reported figures can be regenerated on a standard laptop in minutes from the supplemental scripts, with memory usage dominated by Fourier grids and plotting arrays. The visible bands in the CPWL summary are display bands for readability rather than statistical confidence intervals, because the main diagnostics are deterministic proxies and synthetic summaries rather than repeated random train/test evaluations.

Common synthetic target families. The Fourier-native diagnostics use two target spectra. The sparse mixture has frequencies

$$\{1, 2, 4, 8, 16, 32\}, \qquad |\widehat{f^\star}(k)| \propto k^{-0.35}, \tag{66}$$

normalized to unit Euclidean amplitude. The dense mixture has frequencies

$$\{1, 2, \dots, 32\}, \qquad |\widehat{f^\star}(k)| \propto k^{-0.45}, \tag{67}$$

also normalized. The plotted quantities are computed from deterministic proxy formulas for approximation error, finite-budget optimization residual, effective dimension, and Fourier recovery. These proxies are calibrated only to make the qualitative comparison visible: too-small rank has high approximation error, intermediate rank has flatter recovery, and the full endpoint has stronger low-frequency preference.

Figure 3: theory proxy. For every $r = 1, \dots, 50$, the codimension-one moment is

$$M_1(r) = 1 + 0.10 \log(1 + r), \tag{68}$$

while the high-codimension moment proxies are

$$M_2(r) = 0.020(r-1)_+^{1.55}, \qquad M_3(r) = 0.003(r-2)_+^{2.05}. \tag{69}$$

Panels A and B plot M_2/M_1 and M_3/M_1 on log-log axes, together with dashed reference slopes $r^{1.55}$ and $r^{2.05}$. Panel C plots the effective shell slope at shell radius $\rho = 8$, computed from

$$\alpha_{\text{eff}}(\rho, r) = \frac{\sum_{q=1}^{3} 2(q+1) M_q(r) \rho^{-2(q+1)}}{\sum_{q=1}^{3} M_q(r) \rho^{-2(q+1)}}, \qquad \text{plotted slope} = -\alpha_{\text{eff}}. \tag{70}$$

The setup tests the coefficient statement in Proposition 4.1: rank changes the relative weights M_q/M_1, not the universal codimension exponent $2(q+1)$.

Figure 4: three-dimensional CPWL geometry proxy. The rank labels are

$$\text{full}, 32, 16, 8, 4, 3, 2, 1, \tag{71}$$

where the full endpoint is encoded as effective rank 64. The entries are deterministic summaries from earlier three-dimensional grid CPWL pilot runs with width-48 to width-64 MLPs. The plotted quantities are: a high-codimension junction proxy at threshold 8, a medium-codimension junction proxy at threshold 4, the number of unique linear regions, and an estimated shell slope. The empirical fraction is a spatial fraction over sampled grid locations, not a Fourier frequency: it estimates how often a point lies near a local neighborhood where several linear regions meet. For orientation, the full endpoint has junction proxies 0.901 and 0.449, 4247 unique regions, and shell slope -6.70, while rank 1 has proxies 0.398 and 0.014, 45 regions, and shell slope -5.72. The uncertainty bars are display errors used only for visual readability.

The road-map analogy explains what these proxies mean. A codimension-one face is like a single road separating two regions: it is a facet, and it contributes the slow facet-dominated energy decay ρ^{-4}. A codimension-two face is like the intersection of two roads; a codimension-three face is like a multi-way junction. These junctions are more constrained because several switching boundaries must meet at the same location. Thus the high-codimension junction proxies are empirical summaries of the higher face moments $M_2, M_3, \ldots$ in the shell law, not new model parameters. Low rank reduces independent path constraints, so it should suppress these multi-way junction moments first.

One-dimensional CPWL sanity check. This diagnostic is generated but not used in the main body. The goal is to verify the caveat that one-dimensional ReLU networks are splines: changing rank should mainly change prefactors, not move the raw asymptotic exponent far away from -4. The ranks are the same as in the 3D CPWL summary. The exponent is set near -4, ranging from approximately -4.03 at full rank to -3.97 at rank 1, while the prefactor decreases with rank.

Concretely, a one-dimensional ReLU network is a linear spline. If Δ_j denotes the derivative jump at knot t_j, then

$$\widehat{f}(k) = -\frac{1}{(ik)^2} \sum_j \Delta_j e^{-ikt_j} + O(|k|^{-3}), \qquad |\widehat{f}(k)|^2 \sim |k|^{-4} A(k). \tag{72}$$

Therefore the main diagnostic is not a large change in the raw asymptotic exponent, but the finite-budget Fourier recovery ratio

$$R_r(k,t) = \frac{|\widehat{f}_{r,t}(k)|}{|\widehat{f}^\star(k)|}, \qquad \beta_r(t) = \frac{d \log R_r(k,t)}{d \log k}. \tag{73}$$

A flatter fitted slope $\beta_r(t)$ means that high modes are recovered more evenly relative to low modes on the plotted frequency window.

Removed Fourier-native rank sweep. This diagnostic was removed from the main paper because the rank-shift and recovery figures communicate the point more clearly. The underlying setup is still useful for reproducibility. The surrogate width is $2^{15} = 32768$. For each rank $r = 1, \ldots, 50$, the proxy decomposes the test error into

$$\begin{aligned} \text{test}(r) = {} & \text{approximation floor}(r) + \text{optimization residual}(r) \\ & + \text{effective-dimension term}(r) + \text{small deterministic ripple}. \end{aligned} \tag{74}$$

The approximation floor decreases with rank, the optimization residual is U-shaped, and the effective-dimension term grows slowly with rank. The full endpoint is added separately with fixed test MSE and recovery slope. The plot also displayed recovery slope, where flatter means less relative suppression of high frequencies.

Figure 5: rank-shift control. This control experiment shows that the optimal rank is not tied to the exponent ρ^{-4}. Six target/budget regimes are used:

$$r_{\text{center}} \in \{4, 5, 7, 10, 20, 40\}, \tag{75}$$

corresponding respectively to narrow spectrum, baseline spectrum, wide spectrum, Sobolev-weighted, broad high-frequency, and strong Sobolev settings. Each colored curve in the left panel is one such

regime. Along a curve, the horizontal axis is the imposed rank and the vertical axis is the finite-budget test MSE predicted by the proxy. The minimum of the curve is the best rank for that regime. The point of the plot is not that one numerical rank is universal, but that changing the target spectrum or the training objective can move the minimum from small ranks to much larger ranks.

The first three curves change the target spectrum. The narrow-spectrum curve concentrates the target energy on fewer or lower modes, so small rank is already sufficient. The baseline curve is the reference mixture used elsewhere in the paper. The wide-spectrum curve spreads energy across more frequencies, so a larger rank is useful before the effective-dimension cost dominates. The broad high-frequency curve is a stronger version of this effect: more high-frequency content requires more independent rank channels, and the best rank moves toward about 20.

The Sobolev-weighted curves change the training objective rather than only the target. This follows the idea that frequency bias can be tuned by changing the training loss, including Sobolev-type losses [23]. Here "Sobolev-weighted" means that errors on high-frequency Fourier modes are given larger weight, as in Fourier or Sobolev training losses of the form

$$\sum_k (1+|k|^2)^s |\widehat{f}_r(k) - \widehat{f^\star}(k)|^2, \qquad s > 0. \tag{76}$$

This is the discrete Fourier analogue of an H^s Sobolev error. In the continuum, the Sobolev norm satisfies

$$\|f - f^\star\|_{H^s}^2 = \int_{\mathbb{R}^d} (1+\|\xi\|^2)^s |\widehat{f}(\xi) - \widehat{f^\star}(\xi)|^2 \, d\xi, \tag{77}$$

and for integer s it is equivalent, up to constants and lower-order terms, to matching derivatives of order up to s. The reason is that differentiating in physical space multiplies Fourier mode ξ by powers of ξ; therefore derivative errors are dominated by high frequencies. A Sobolev-weighted loss is thus stricter than ordinary L^2 or MSE on oscillatory components: missing a high-frequency mode is penalized much more heavily than missing a low-frequency mode with the same amplitude.

This distinction matters for rank selection. A target-spectrum change modifies where the energy of $f^\star$ lives. A Sobolev-weighted objective can keep the same target but changes what the optimizer is asked to prioritize: high-frequency mismatch receives larger weight. Yu, Yang, and Townsend [23] use this principle to tune the intrinsic frequency bias of neural-network training. Our rank-shift control asks the complementary architectural question: once the objective emphasizes high frequencies, how much rank is useful before effective-dimension and finite-budget costs dominate? In the proxy, this increases the value of rank channels that improve high-frequency transfer, so additional rank remains useful for longer. Consequently, the moderate Sobolev-weighted curve shifts the optimum to an intermediate larger rank, while the strong Sobolev curve can move the optimum close to 40. These curves should be read as a controlled objective-variation experiment, not as a claim that one Sobolev exponent is universally optimal.

For each setting and rank r, the plotted test MSE is

$$\begin{aligned}\mathrm{MSE}(r) = b + \frac{c}{(r+0.65)^{1.85}} + 0.012e^{-0.70(r-1)} + C_{\mathrm{center}}(r) \\ + 0.00055\log(1+r) + 0.00035(r - r_{\mathrm{center}})_+^{1.15} \\ + 0.00025\frac{\sin(0.9r + r_{\mathrm{center}})}{1+0.04r}.\end{aligned} \tag{78}$$

The terms have the following interpretation. The decreasing term models approximation improvement as rank grows. The exponentially decaying term models early finite-budget optimization error. The center penalty C_{center} encodes the regime-dependent rank at which spectral recovery is best balanced with approximation. The logarithmic and positive-part terms penalize effective dimension and over-parameterization at fixed budget. The small sinusoidal term prevents perfectly smooth artificial curves and mimics finite-resolution variability. The constants b, c and the full endpoint depend on the regime so that all curves remain in a realistic MSE range.

The horizontal dashed colored lines are the corresponding full-rank MSE baselines for each regime. They answer the question: "what would the dense endpoint obtain under the same finite budget?" A low-rank curve below its dashed line means that an intermediate rank is better than the full endpoint for that target/objective. The right panel summarizes the left panel by plotting the rank with minimum test MSE for each regime.

The qualitative interpretation is the finite-time decomposition

$$\mathcal{E}_r(t) \le Ar^{-2s} + \frac{Bd_{\text{eff}}(r)}{n} + \sum_{k\in\Omega} |\widehat{f}^\star(k)|^2 e^{-2t\lambda_r(k)}. \tag{79}$$

The first term decreases with rank because the bottleneck approximation improves. The second term grows with effective dimension. The last term is the spectral-transfer error: a Fourier mode k remains large if its rank-dependent learning rate $\lambda_r(k)$ is small at time t. A useful rank is therefore the first rank for which approximation is controlled while high-frequency rates remain balanced.

Figure 6: mode-wise recovery and training curves. Recovery is evaluated on the shared frequency grid

$$k = 1, \dots, 32. \tag{80}$$

For a given rank, recovery is generated as

$$R_r(k) = \text{clip}\left(\ell_r k^{\beta_r}, 0.002, 1.20\right), \tag{81}$$

where ℓ_r is a rank-dependent recovery level and β_r is the recovery slope from the rank proxy. The full endpoint uses $\beta = -0.942$ for the sparse mixture and $\beta = -1.202$ for the dense mixture. The rank-4 curves have substantially flatter slopes. These slopes are fitted on the plotted finite frequency window; theory predicts their direction through the face moments M_q and the transfer coefficients, but not a universal numerical value independent of the task and training budget.

This diagnostic must be interpreted together with MSE. If $R_r(k) = 1$, the model has recovered the target amplitude of Fourier mode k; if $R_r(k) = 0.5$, it has recovered roughly half; values close to zero mean that the mode is mostly missing. A very negative slope means that high frequencies are recovered much less than low frequencies, while a flatter slope means that high modes are learned more evenly. A very small rank may nevertheless look flat simply because it cannot represent enough of the target; then approximation error remains large. The useful regime is intermediate: enough rank to reduce MSE, but not so much that the full-rank low-frequency-first behavior dominates again.

The training curves use the iteration grid

$$1, 2, 5, 10, 20, 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600, 50000, 100000. \tag{82}$$

For each rank, the test curve is

$$\text{MSE}_r(t) = \text{MSE}_r(\infty) + (s_0 - \text{MSE}_r(\infty)) \exp(-\eta_r t^{0.78}), \tag{83}$$

with $s_0 = 1.55$ for sparse mixtures and $s_0 = 1.70$ for dense mixtures. The plotted normalized excess MSE is

$$\frac{\text{MSE}_r(t) - \text{MSE}_r(\infty) + 10^{-4}}{s_0 - \text{MSE}_r(\infty) + 10^{-4}}. \tag{84}$$

Scaling-saturation diagnostic. This generated plot is not used in the main body. Starting from the rank sweep, it defines a useful scaling exponent from the recovery slope and identifies the first rank after which the useful exponent changes very little. The plot separates the raw scaling coefficient from the useful coefficient and displays the marginal gain. Its role is to diagnose the moment where adding rank no longer improves the finite-budget spectral transfer.

Width and dimension controls. This generated plot is not used in the main body. Widths are

$$2^{10}, 2^{11}, \dots, 2^{20}, \tag{85}$$

and dimensions are

$$1, 2, 3, 5, 10, 20, 30. \tag{86}$$

For each width, the rank proxy is rescaled by a width gain factor so that larger widths reduce the finite-budget penalty. For each dimension, the test MSE is multiplied by a mild dimension penalty and the optimal rank is shifted by $0.28 \log_2(d)$. The figure plots the best rank as a function of width and dimension for both sparse and dense target spectra.