Low-Rank Neural Network Structure Is Sufficient for Global Convergence: A Mean-Field Perspective

Janis (Heran) Aiad[1]* Haizhao Yang[2]
Shijun Zhang[3]
[1] École Polytechnique, École Normale Supérieure Paris-Saclay
[2] University of Maryland, Department of Mathematics; Department of Computer Science
[3] The Hong Kong Polytechnic University

Abstract

This work studies the training dynamics of low-rank neural networks with frozen random features in the mean-field regime. When the mean-field dynamics converges, the limit is shown to be a global minimizer; this holds for gradient-based training under standard independent and identically distributed initialization, despite low-rank constraints and nonconvex loss functions. By explicitly incorporating low-rank structure into the network architecture, a tractable mean-field evolution system is derived. Its well-posedness is established, and it is shown that with frozen random features L^0, frozen mixing matrices $L^{(\ell)}$, and only the channel weights w_ℓ trained, the universal approximation property is preserved while the learning dynamics are simplified. The analysis identifies a rank-channel feature learning mechanism, in which different low-rank channels specialize to distinct spatial locations and progressively capture higher-frequency components. This mechanism explains both the persistence of global convergence and the emergence of hierarchical frequency learning. Numerical experiments demonstrate that low-rank networks achieve faster convergence and higher accuracy on highly oscillatory targets, while using substantially fewer parameters than full-rank networks.

Keywords mean-field theory, low-rank neural networks, random features, convergence to global minimizer, non-convex optimization, function approximation

Contents

*This work, including all theoretical derivations, calculations, and manuscript, was completed by J.H-A. during a Visiting Assistant Researcher at University of Maryland.

1 Introduction

Neural networks have achieved remarkable success across a wide range of applications, including computer vision, natural language processing, scientific computing, and data-driven modeling. These advances have reshaped modern machine learning practice and enabled solutions to problems that were previously considered intractable. Despite this empirical success, theoretical understanding of neural network training remains limited. A central difficulty lies in the highly non-convex nature of the learning landscape. For most architectures and problem settings, the optimization process is poorly understood: it is generally unclear why gradient-based methods succeed in practice, how representations evolve during training, or under what conditions convergence can be guaranteed. As a result, much of current understanding relies on empirical observations rather than rigorous analysis.

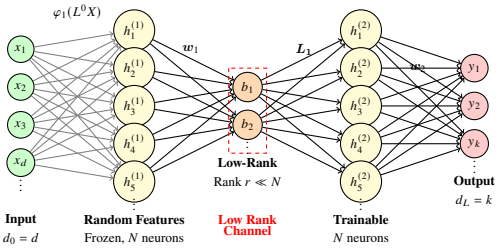


Figure 1: Architecture of a three-layer low-rank random feature network. The first layer consists of frozen random features, the second layer employs a low-rank mixing matrix L of rank $r \ll N$, and the third layer contains trainable weights w_2. The red dashed box highlights the low-rank bottleneck that reduces dimensionality from N to r channels.

However, full-rank models are heavily over-parameterized. The number of trainable parameters and the associated computational cost can be far larger than what is effectively needed to represent the target function. From both computational and modeling perspectives, it is natural to seek more structured and efficient alternatives. Low-rank neural networks offer such a possibility by reducing parameter redundancy and computation while retaining expressive power. As illustrated in Figure **??**, low-rank models achieve comparable or better training loss with 99% fewer parameters than full-rank networks.

This raises a fundamental question: *is low rank all we need for global convergence ?* Low-rank networks factorize weights as $W = LR^\top$ with $r \ll \min\{n, m\}$, substantially reducing parameters; when such factorizations replace full-rank matrices, can gradient-based training still converge to a global minimizer, or is full rank essential? It is unclear whether the favorable optimization and convergence properties of full-rank mean-field analyses extend to this setting. We therefore conduct a systematic theoretical investigation of low-rank neural networks from optimization and representation learning: This question is resolved in the present work, as summarized by the following informal theorem.

Theorem 1.1 ((informal) Convergence to a global minimizer for any depth with i.i.d. init.)**.** *For any depth $L \geq 2$, low-rank random feature networks with standard initialization and non-negative loss function, and if the weights in all layers $(w_1, \ldots, w_{L-1})$ converge as $t \to \infty$, then the limit is a global minimizer of the population loss.*

Theorem **??** is an informal statement of the main convergence result proved rigorously in Theorem **??**. Beyond global convergence, the analysis further establishes a hierarchical frequency learning phenomenon for low-rank neural networks, revealing how learned representations are organized across different low-rank channels. The main contributions of this work are summarized as follows.

- **Global convergence under low-rank constraints** (Theorem **??**). It is shown that when the dynamics of low-rank random feature neural networks converges, it converges only to global minimizers of the population loss, for any depth $L \geq 2$ under standard independent and identically distributed initialization. Unlike previous full-rank analyses, the result requires no special or ad-hoc initialization.
- **Mean-field feature learning** (Theorem **??**). A theoretical characterization of feature learning mechanism is established. A rigorously analyzed toy model demonstrates that each low-rank channel learns a spatially localized feature at a *distinct location*. The spatial features distribution (Figure **??**) is an important diagnostic for this mechanism and is strongly confirmed through numerical experiments.

On MNIST, low-rank networks with frozen random features reach ~97% test accuracy with 13k–31k trainable parameters, while matching a ReLU MLP at ~98% when all parameters are trained (Table **??**).

Table 1: MNIST: MLP baseline vs low-rank vs random features; LR = low-rank, RF = frozen random features.

Model	Rank	Trainable	Test acc (%)
MLP	–	669,706	98.39
RF-LR	5	7,695	93.72
RF-LR	10	10,260	96.24
RF-LR	15	12,825	96.97
RF-LR	25	17,955	97.00
RF-LR	50	30,780	97.01
LR Only	32	440,362	98.30

Numerical experiments confirm that this theoretical structure translates into practical benefits. On highly oscillatory function, low-rank networks achieve substantially faster convergence and significantly higher accuracy, reaching MSE $\sim 10^{-6}$, compared to approximately 10^{-5} for full rank networks. **At the same time, the number of trainable parameters is reduced by** 95%–99% These results suggest a fundamental insight into the loss landscape: appropriately structured low-rank parameterizations can improve optimization behavior, rather than merely restricting expressivity.

The remainder of the paper is organized as follows. Section **??** reviews related work. Section **??** presents the main theoretical results and proof ideas. Section **??** studies hierarchical frequency learning and channel specialization. Section **??** reports numerical experiments, and Section **??** concludes.

1.1 Related work

Mean-Field Theory for Neural Networks Mean-field theory rigorously analyzes neural network training dynamics in the large-width limit (typically $N \geq 1000$ neurons per layer). In this regime, the empirical distribution of parameters evolves deterministically, leading to powerful tools for understanding optimization and generalization. Foundational results by [**?**], [**?**], [**?**], and [**?**] established global convergence for two-layer networks under convex loss assumptions—crucially leveraging convexity to guarantee that all stationary points are global minimizers.

Progress beyond the convex case has accelerated. [**?**] proved global convergence for three-layer networks without assuming convex loss functions. Mean-field analyses of residual networks [**?**] and multilayer networks [**?**] further clarify when global optima arise from stationary points, but these results so far apply only to full-rank networks. Whether such convergence persists in computationally efficient, low-rank architectures remained open.

We address this gap: we train only the right factor R, keeping L frozen as random features—avoiding both RGD and the full-rank mean-field collapse to one parameter per intermediate layer under i.i.d. initialization [**?**]. By leveraging frozen random features, we prove that channel feature learning (in which r independent channels capture hierarchical frequency structure) leads to both global convergence and practical representational learning.

The present analysis extends the framework of Chizat and Bach [**?**] (Theorem 2: if the initial support spans $\mathbb{R}^{d+1}$ and Ψ is positively 2-homogeneous, any weak limit of the Wasserstein gradient flow is a global minimizer) to *low-rank, multi-layer* networks. We build on [**?**, **?**].

A key theoretical advance is the elimination of ad-hoc initialization. For standard fully-connected or convolutional networks, [**?**] prove that under i.i.d. initialization and constant initial biases the mean-field limit *collapses*: at each intermediate layer the weight dynamics reduces to a single deterministic translation parameter (independent of neuronal indices). To obtain global convergence they therefore require an ad-hoc initialization that avoids this degeneracy. We do *less* by way of initialization—we impose an architectural restriction (frozen random features and low-rank layers) and assume only standard i.i.d. initialization—and prove *more*: frozen random features ensure $\mathrm{supp}(L^0(C_1)) = \mathbb{R}^d$ throughout training, so the dynamics does not collapse and standard i.i.d. suffices; when the dynamics converges, the limit is a global minimizer for *arbitrary depth* $L \geq 2$. See Remark **??** and [**?**] for the full-rank setting.

Low-Rank and Random Features Our RF-LR architecture is based on [**?**]. Low-rank methods such as LoRA [**?**] are widely used; most theory focuses on expressivity rather than training dynamics. [**?**] study end-to-end low-rank training via reparameterization gradient descent (RGD); we train only post-activation parameters and freeze the other factor as random features. That choice yields convergence only to global minimizers and clarifies channel feature learning: each channel specializes to distinct spatial and frequency patterns.

Random features provide a bridge between neural networks and kernel methods. In the mean-field width limit, fixing the first layer as random features connects the network to kernel methods governed by the neural tangent kernel (NTK) [**?**], offering tractable analyses of optimization and generalization. Our architecture embraces this by freezing the first-layer

weights, so that low-rank mixing layers can learn task-specific representations atop a fixed random feature basis. Recent work establishes that low-rank structure suffices for the MLP NTK [**?**], while our results extend this to mean-field dynamics and global convergence. Together, these advances show that, both in NTK and mean-field regimes, low-rank networks can match the optimization guarantees of their full-rank counterparts.

2 Main Results and Proof Ideas

2.1 Low-Rank Random Feature Architecture

We consider a low-rank random feature (RF-LR) network architecture. Let $h^{(0)}(x) = x \in \mathbb{R}^{d_0}$. For layers $\ell = 1, \dots, L$:

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

where n_ℓ is the width of layer ℓ, $w_j^{(\ell)} \in \mathbb{R}^{d_\ell}$ are trainable coefficient vectors, $L_j^{(\ell)} \in \mathbb{R}^{d_{\ell-1}}$ and $b_j^{(\ell)} \in \mathbb{R}$ are frozen random features (i.i.d. uniform or Gaussian), and c is a trainable scalar output bias. The $1/n_\ell$ scaling ensures a well-defined mean-field width limit. In the mean-field width limit (typically $N \geq 1000$ neurons per layer), where N denotes the width of each layer, under the mean-field parameterization with i.i.d. initialization (uniform or Gaussian), the empirical distribution of weights converges to a deterministic measure-valued evolution.

Training policy: $w^{(\ell)}$ and c_ℓ are trained; $L^{(\ell)}$ and $b^{(\ell)}$ are frozen random draws. In the mean-field formulation, biases $b_j^{(\ell)}$ are encoded by augmenting the input data with a constant component (adding 1 to the input vector).

2.2 Mean-Field Forward Equations

We use the *neuronal embedding* framework of [**?**]: each neuron is indexed by a label C lying in a probability space (countable or uncountable), which we specify when sampling or constructing the network. In the finite-width case, C corresponds to discrete neuron indices; in the mean-field limit, C runs over the support of a measure, and expectations $_C[\cdot]$ replace empirical averages over neurons.

Definition. We define the mean-field forward pass layer by layer. The *frozen* first-layer feature map is $L^0(c_1) \in \mathbb{R}^{d_0}$; we specify the law of $L^0(C_1)$ when sampling the network. In layers $i \geq 2$, the low-rank structure is given by a *frozen* mixing matrix L (entries $L_{c_2,k}$, or $L^{(i)}_{c_i,k}$ for intermediate layers), also fixed random features drawn at initialization; we call it *mixing* in those layers. All other layer-wise quantities are defined exactly as follows.

- **Input ($i = 0$):** $h^0(X) = X \in \mathbb{R}^{d_0}$; no neuronal index.
- **Hidden ($i = 1, \dots, L-1$):** Index C_i. At $i = 1$: frozen $L^0(c_1)$; $H_1 = L^0(C_1)X$; $f_k^{(1)} =_{C_1} [w_1 \, \varphi_1(H_1)]$. For $i \geq 2$: $H_i = \sum_k L^{(i)}_{c_i,k} \, f_k^{(i)}$ with $f_k^{(i)} =_{C_{i-1}} [w_{i-1} \, \varphi_{i-1}(H_{i-1})]$. Activation $\varphi_i(H_i)$.
- **Output ($i = L$):** $\hat{y} =_{C_{L-1}} [w_{L-1} \, \varphi_{L-1}(H_{L-1})]$.

2.3 Mean-Field Backward Equations (ODEs)

The mean-field ODE system (backward equations) for the weights takes the form, for all layers $i = 1, \dots, L-1$:

$$\begin{aligned} \partial_t w_1(t, c_1, k) &= -\xi_1(t) \, _Z[d_L \, \varphi_1(L^0 X) \, B_k^{(2)}], \\ \partial_t w_i(t, c_i) &= -\xi_i(t) \, _Z[D_i \, \varphi_i(H_i(t, c_i; X, W))], \end{aligned} \tag{2}$$

where $_Z[\cdot]$ denotes expectation over $Z = (X, Y)$; $d_L(Z; W(t)) = \partial_{\hat{y}} \mathscr{L}(Y, \hat{y}(X; W(t)))$ is the loss derivative; $\xi_i(t) \geq 0$ are the learning-rate schedules. In (**??**), $D_{L-1} = d_L$; for $i \leq L-2$, D_i is the back-propagated pre-activation gradient at H_i [**?**]:

$$\begin{aligned} D_i &= \varphi_i'(H_i) \sum_k w_i \, _{C_{i+1}}[D_{i+1} \, L^{(i+1)}_{C_{i+1},k}] \\ D_{L-1} &= d_L \, w_{L-1} \, \varphi_{L-1}'(H_{L-1}). \end{aligned} \tag{3}$$

The channel-wise backpropagated signals are

$$B_k^{(\ell)}(t; X, W) =_{C_\ell} \Big[L^{(\ell)}_{C_\ell,k} \, \varphi_\ell'(H_\ell(t, C_\ell; X, W)) \, w_\ell(t, C_\ell)\Big].$$

2.4 Assumptions

The applicability of our results depends crucially on the following assumptions, which we state explicitly in Appendix **??** and are **satisfied by standard MLP architectures**

The main assumptions are: *Bounded Activations and Mixing* (Assumption **??**), *Sub-Gaussian Initialization* (Assumption **??**), *Data Distribution and Loss Regularity* (Assumption **??**), *Diversity of Random Features* (Assumption **??**), *Non-Degeneracy* (Assumption **??**), and *Training convergence to limit point* (Assumption **??**). The *Non-Degeneracy* assumption requires that the initial loss is better than the trivial zero predictor, ensuring the network learns a non-trivial solution. The *Convergence to Limit Point* assumption states that training reaches a limit point, which is a natural condition for analyzing convergence.

These assumptions are satisfied by Leaky ReLU or sigmoid (or tanh) networks on bounded data with standard initialization (Gaussian/Xavier): those activations have φ' bounded and bounded away from zero, and bounded inputs plus Gaussian/Xavier fulfill the remaining regularity, sub-Gaussian, and non-degeneracy conditions; for ReLU, the same holds with high probability in r (Appendix **??**).

2.5 Mean-Field ODEs are well posed

We establish well-posedness of the mean-field ODE system (**??**) by adapting the proof from [**?**] to account for the low-rank structure.

Theorem 2.1 (Well-posedness of mean-field ODEs)**.** *Under* Bounded Activations and Mixing *(Assumption* **??***),* Sub-Gaussian Initialization *(Assumption* **??***),* Data Distribution and Loss Regularity *(Assumption* **??***), and* Diversity of Random Features *(Assumption* **??***), there exists a unique solution to the mean-field ODE system* (**??**) *on* $t \in [0, \infty)$.

Proof sketch. We first states the bi-Lipschitz property of H_ℓ and $\varphi_\ell(H_\ell)$ in W (Lemma **??**, Appendix **??**): $|H_\ell(W') - H_\ell(W'')|$ and $|\varphi_\ell(H_\ell(W')) - \varphi_\ell(H_\ell(W''))|$ are bounded by $K\|L^{(\ell)}\|_{\infty,1} \le rK$ times weight differences. This is the simple but key adaptation to the low-rank case. *From this*, we adapt and introduce weight-space Orlicz sub-gaussian norms accounting for the r channels ($\max_{1\le k\le r}$ over channel index k, see Appendix **??**), update $K_0(t)$ with a factor $(1 + rK)^{1/2}$, establish sub-Gaussian a priori bounds with r-factors, and show the solution operator F is contractive in these spaces; Banach fixed point yields existence and uniqueness. Details in Appendix **??**. □

2.6 Global convergence

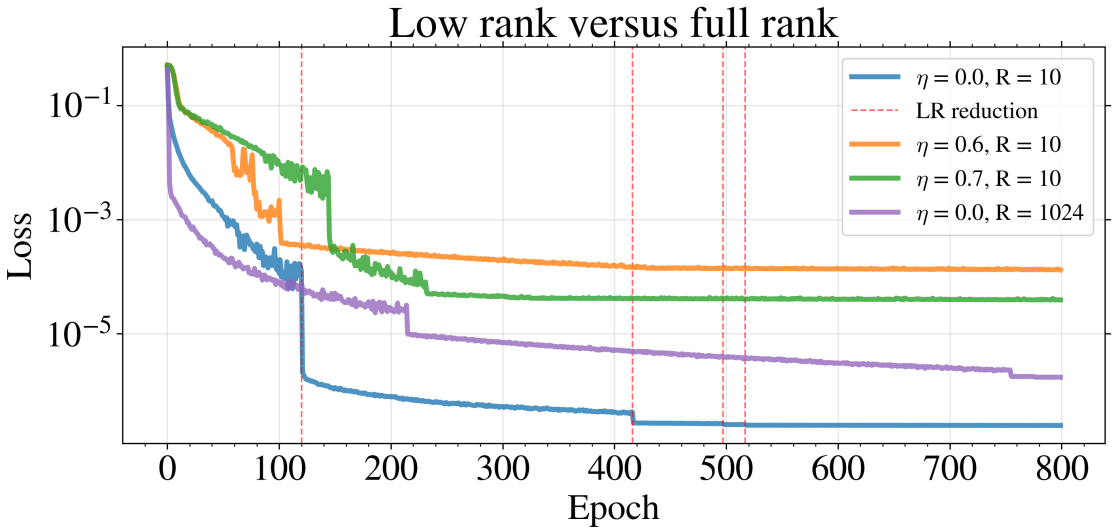
Frozen random features with full support $\mathrm{supp}(L^0(C_1)) = \mathbb{R}^d$ ensure that $\{\varphi_1(\langle L^0(c_1), \cdot\rangle) : c_1 \in \Omega_1\}$ has dense span in $L^2(\mathcal{P}_X)$ when φ_1 is non-polynomial (e.g. Leaky ReLU or sigmoid); $\varphi_1(af + b)$ for random a, b then forms a dense span. This property is maintained throughout training because $L^0(C_1)$ are frozen. For well-posedness (Theorem **??**), each $H_\ell(t, c_\ell; X, W)$ bi-Lipschitz in W is sufficient and equivalent to φ_ℓ Lipschitz; it is the only thing that matters. After defining norms, we update $K_0(t)$ with a factor $(1 + rK)^{1/2}$; the solution operator F, a priori bounds, and the contraction argument then proceed as in the full-rank framework [**?**].

Theorem 2.2 (RF-LR training only converges to global minimizers)**.** *Under all assumptions from* Bounded Activations and Mixing *(Assumption* **??***) through* Convergence to Limit Point *(Assumption* **??***), and the* loss condition *in* Data Distribution and Loss Regularity *(Assumption* **??***) (*$\partial_2\mathcal{L} = 0 \Rightarrow \mathcal{L} = 0$*), if the mean-field dynamics for low-rank random feature networks converges, then it is to a global minimizer of the population loss:* $\lim_{t\to\infty} W(t) = W^*$ *where* W^* *minimizes* $\mathscr{L}$*. For any depth* $L \ge 2$*, this holds with standard i.i.d. initialization.*

Equivalently, any limit point of the mean-field dynamics is a global minimizer of the population loss under the loss condition in Assumption **??**.

High-level proof idea (improving [?], Sec. 6.2.1). At a limit point $\bar{W}$, the gradient-flow ODE has zero time derivative. For the top layer (and, by backprop, at each layer), this yields $_Z[\text{upstream} \times \text{local}] = 0$ over the support of the layer's neuron measure. The crucial step: if the first-layer features $\{\varphi_1(\langle\theta, \cdot\rangle) : \theta \in \mathrm{supp}(\text{feature measure})\}$ have dense span in $L^2(\mathcal{P}_X)$, then one deduces $_Z[\partial_{\hat{y}}\mathscr{L}(Y, \hat{y}(X; \bar{W})) \mid X = x] = 0$ for $\mathcal{P}_X$-a.e. x. Under the loss condition in Assumption **??**, $\mathbb{E}[\partial_2\mathcal{L}(Y, u)|X = x] = 0$ implies $\mathbb{E}[\mathcal{L}(Y, u)|X = x] = 0$, so $\mathscr{L}(\bar{W}) = 0$ and $\bar{W}$ is a global minimizer.

In [**?**], the first-layer weights w_1 are *trained*, so $\mathrm{supp}(\bar{w}_1(U_1)) = \mathbb{R}^d$ at the limit may fail (e.g., solutions may become sparse or concentrated). Their proof therefore relies on a homotopy argument to show that $\mathrm{supp}(w_1(t, U_1)) = \mathbb{R}^d$ is preserved for all finite $t \ge 0$, so the limit retains enough richness.



rank $r = 10$) vs full-rank ($r = 1024$) on $y = \cos(2\pi x)$, $x \in [-1, 1]$, 3 layers, width 1024, $n = 5000$, batch=4, SGD lr 0.01 with rec

(momentum $\eta \in \{0, 0.3, 0.6, 0.7\}$) and one full-rank ($\eta = 0$). Red bars: SGD lr reductions for $\eta = 0$.

Figure 2: Low-rank (rank r=10) vs full-rank (r=1024) on $y = \cos(2\pi x)$, $x \in [-1, 1]$, 3 layers, width 1024, n=5000, batch=4, SGD lr 0.01 with red-bar decay. (momentum $\eta \in \{0, 0.3, 0.6, 0.7\}$) and one full-rank ($\eta$=0). Red bars: SGD lr reductions for η=0. Low-rank uses 99% fewer parameters and vanilla SGD performs best.

We *freeze* the feature maps $L^0(C_i)$ for all layers i; they are not trained, so $\mathrm{supp}(L^0(C_i)) = \mathbb{R}^d$ for every i and all t, and $\{\varphi_i(\langle L^0(c_i), \cdot\rangle) : c_i \in \Omega_i\}$ has dense span in $L^2(\mathcal{P}_{H_{i-1}})$ for every layer. Thus we can apply the global-optimality argument of [**?**] *without* the homotopy step: at a limit point, zero derivative plus dense span at each layer gives $_Z[\partial_{\hat{y}}\mathscr{L} \mid X = x] = 0$ a.e. and hence a global minimizer. The low-rank form ($H_i = \sum_k L_{c_i,k} f_k$, etc.) only changes upstream and local terms; the gradient-flow logic is unchanged. For $L \geq 3$, the layer-by-layer argument works by virtue of the frozen L^0. Unlike full-rank $L \geq 3$, which requires ad-hoc init to avoid collapse, we need only standard i.i.d. init (Remark **??**).

Remark 2.1 (Avoiding ad-hoc initialization). Full-rank $L \geq 3$ needs ad-hoc init in [**?**] (else intermediate layers collapse). Frozen L^0 keeps $\mathrm{supp}(L^0(C_1)) = \mathbb{R}^d$, so standard i.i.d. suffices and any limit is a global minimizer.

2.7 Quantitative Guarantees

We provide a quantitative approximation theorem that bounds the error between finite-width networks and the mean-field limit. The detailed proof are provided in Appendix **??**; here we summarize the key result.

Theorem 2.3 (Finite-width approximation error bound)**.** *Given a family* Init *of initialization laws and a tuple* $\{n_1, n_2\}$ *that is in the index set of* Init*, perform the coupling procedure for the low-rank architecture as described in Section* **??**. *Fix a terminal time* $T \in \epsilon\mathbb{N}_{\geq 0}$. *Under Assumptions* **??**, **??** *(see Appendix* **??***), and the low-rank structure with mixing matrix* L *satisfying* $\|L\|_{\infty,1} \leq rK$*, for* $\epsilon \leq 1$*, we have with probability at least* $1 - 2\delta$*,*

$$\mathscr{D}_T(W, \mathbf{W}) \leq C_{\exp} \cdot C_{width} \cdot C_{\log},$$

where $C_{\exp} = e^{K_T(1+rK)}$, $C_{width} = 1/\sqrt{n_{\min}} + \sqrt{\epsilon}$, $C_{\log} = \sqrt{\log(3(T+1)n_{\max}^2/\delta + e)}$, *with* $n_{\min} = \min\{n_1, n_2\}$, $n_{\max} = \max\{n_1, n_2\}$, $K_T = K(1 + T^K)$, *and the factor* $(1 + rK)$ *accounts for the low-rank structure through* $\|L\|_{\infty,1} \leq rK$.

Training in practice. The theorem (Appendix **??**) links $\mathbf{W}(\lfloor t/\epsilon \rfloor)$ to $W(t)$ with error $O(1/\sqrt{n_{\min}} + \sqrt{\epsilon})$, independent of d; proof in Appendix. The factor $(1 + rK)$ in the Grönwall constant $e^{K_T(1+rK)}$ arises in the ODE drift bounds (from

the r channels and $\|L\|_{\infty,1} \le rK$) and is then exponentiated by Grönwall; vs. full-rank [**?**] one has $e^{K_T(1+rK)}$ instead of e^{K_T}. Worst-case, the bound is exponential in rK; in practice, channel specialization often restricts to a subset and yields faster convergence (future work, Section **??**).

3 Feature Learning

The low-rank structure enables channel feature learning where different channels learn different spatial-frequency features. In particular, each channel learns a spike at a *different spatial value* (localization): e.g. channel k dominates near some x_k while others remain small there. Empirically, channels also separate by *frequency*—lower frequencies are captured first, higher ones progressively—so that the r channels jointly provide both spatial localization and frequency separation. Theorem **??** below provides a rigorous, conditional characterization of how channels establish and maintain dominance (when the stated hypothesis holds on an interval I), explaining the feature learning in practice.

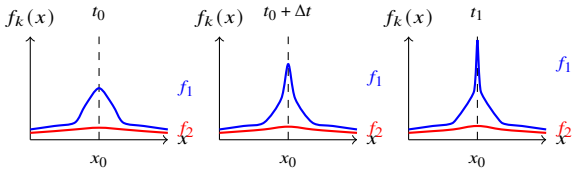


Figure 3: Channel spike learning (Theorem **??**). Functions $f_1(x)$ (blue) and $f_2(x)$ (red) at t_0, $t_0 + \Delta t$, and t_1. f_1 has a spike at x_0 that becomes more pronounced over time, while f_2 has no spike.

3.1 Mechanism of Spike Learning

The log-ratio growth in Theorem **??** arises from a self-reinforcing dominance loop. The evolution $\partial_t f_k(t,x) = -\xi_1(t)\, _{Z=(X,Y)}[d_L(Z; W(t))\, B_k(t; X)\, K_{\mu_0}(x, X)]$ is an integral over the data distribution, where $k \in \{1, \dots, r\}$ is the channel index and $K_{\mu_0}(x, X)$ is the kernel. When channel k dominates at x_0,

$$H_2 = \sum_{j=1}^{r} L_{c_2,j} f_j = \underbrace{L_{c_2,k} f_k}_{\text{large}} + \underbrace{\sum_{j \neq k} L_{c_2,j} f_j}_{\text{small}},$$

so the pre-activation and hence $\varphi_2'(H_2)$ (bounded away from zero under Assumption **??**) are effectively driven by that channel. The backpropagated signal $B_k(t; X) =_{C_2} [L_{C_2,k}\, w_2(t, C_2)\, \varphi_2'(H_2(t, C_2; X))]$ is a mixture over the second layer. The precise signs of $L_{c_2,k}$ or $w_2(t, c_2)$ need not matter: once a channel dominates, the activation derivative φ_2' modulates the contribution (e.g. by half-space for Sigmoid/ReLU/Leaky ReLU), w_2 evolves on that set via the mean-field w_2-ODE, and the dynamics create an *emergent* sign-coherence rather than requiring it a priori. If the activation (gate) is correlated with a channel, that correlation *amplifies*: the channel gets a larger B_k, hence a larger $\partial_t f_k$, so the channel and the gate become even more correlated. Under sign-coherence ($B_{k,1}$ has the same sign as f_k), the integral contribution from the dominant channel is larger, yielding a single self-reinforcing loop: larger dominance $\to$ larger $B_k \to$ larger $\partial_t f_k \to$ further amplified dominance. Theorem **??** shows that *when* the conditions (i)–(iii) in its hypothesis hold on an interval I, this loop implies $\partial_t R_{12} \ge 0$ and thus that dominance cannot be lost on I; see Appendix **??**.

3.2 Toy model with spike feature learning

The full mean-field system for discrete regression in Spatial-Fourier space is infinite dimensional; the two-point ($m = 2$) dynamics nonetheless capture the essential mechanism for feature learning. We state and prove this intuition for the *two-sided step* toy model: $m = 2$ support points $x^{(1)} = x_0 = -\delta$, $x^{(2)} = x_1 = +\delta$ ($\delta > 0$) with $y^{(1)} = +A$, $y^{(2)} = -A$. The target $y(x)$ is supported on $\{x_0, x_1\}$ with values $\pm A$ (Figure **??**). For finitely-supported data, the w_1-evolution is linear in feature space once residuals and backprop signals are fixed; one obtains closed-form f_k as a superposition of kernel bumps. Define $d_p(t) = d_L((x^{(p)}, y^{(p)}); W(t))$, $B_{k,p}(t) = B_k(t; x^{(p)})$, and

$$\Gamma_{k,p}(t) \equiv \int_0^t \xi_1(s)\, d_p(s)\, B_{k,p}(s)\, ds.$$

Then

$$f_k(t,x) = f_k(0,x) - \Gamma_{k,1}(t)\, K_{\mu_0}(x, x_0) - \Gamma_{k,2}(t)\, K_{\mu_0}(x, x_1). \tag{4}$$

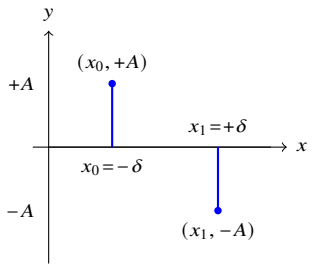


Figure 4: Two-sided step: $x_0 = -\delta, x_1 = +\delta$ with $y(x_0) = +A$, $y(x_1) = -A$.

The spike shape is determined by K_{μ_0}; all learning dynamics reduce to the *scalar* coefficients $\Gamma_{k,1}(t)$, $\Gamma_{k,2}(t)$. The kernel K_{μ_0} (NNGP, [**?**]) satisfies $K_{\mu_0}(x_p, x_p) = K_0 > 0$ for $p \in \{0, 1\}$ (same by symmetry). The off-diagonal $K_{\mu_0}(x_0, x_1) = K_{\mu_0}(-\delta, +\delta)$ is *positive* and *fastly decaying* in δ: $0 < K_{\mu_0}(x_0, x_1) \le \psi(\delta)$ for some $\psi(\delta)$ with $\psi(\delta) \to 0$ rapidly as $\delta \to \infty$. Thus the cross-term is small for separated points and, being positive, reinforces the leading local term and yields even better positivity in the log-ratio dynamics.

The full evolution has *two terms* (local plus non-local):

$$\partial_t f_k(t, x_p) = -\xi_1(t)\, K_{\mu_0}(x_p, x_p)\, d_p(t)\, B_{k,p+1}(t) + E_p(t), \tag{5}$$

where the non-local remainder $E_p(t)$ satisfies $|E_p(t)| \le C'\, \psi(\delta)$ with $\psi(\delta)$ fastly decaying in δ. We state the theorem for x_0 only; the result is symmetrical at x_1 (Appendix **??**).

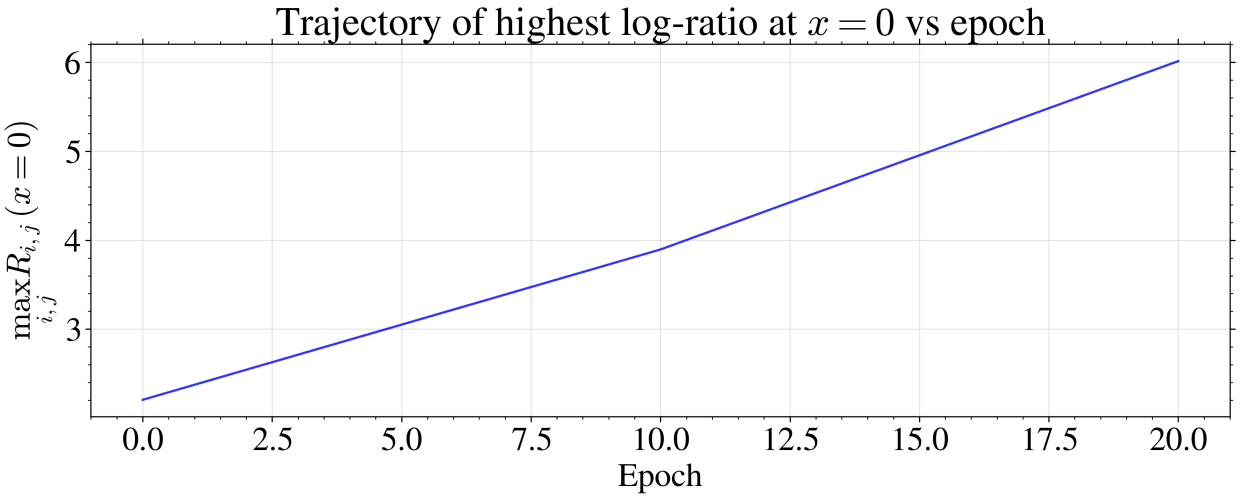


Figure 5: Trajectory of $\max_{i,j} R_{i,j}(x = 0)$ vs. epoch (layer-3 channels). Channel specialization at $x = 0$ as in Theorem **??**. Setup: $n = 1024$, $r = 15$, $\cos(8\pi x)$, 20 epochs.

Theorem 3.1 (Two-sided step: log-ratio growth at x_0 (conditional))**.** ***Setting:*** *The two-sided step with $r = 2$ channels and a 3-layer RF-LR; the dynamics are* (**??**) *and the full evolution* (**??**)*. The result is* ***conditional****:* if *the three conditions in (Hypothesis) hold on an interval $I \subseteq [0, \infty)$,* then *the stated conclusion holds on I.*

Definition (log-ratio criterion). *At x_0: $d_0(t)$ is the residual and $B_{k,1}(t) = B_k(t; x_0)$ the backprop signal; at x_1: $d_1(t)$ and $B_{k,2}(t) = B_k(t; x_1)$. Define*

$$R_{12}(t, x_0) = \log \frac{|f_1(t, x_0)|}{|f_2(t, x_0)|}.$$

And under the hypothesis that backpropagated signals are stable ***Hypothesis (at x_0, for all $t \in I$):*** *(i) $-d_0(t) \ge 0$; (ii) $B_{1,1}(t)$ has the same sign as $f_1(t, x_0)$; (iii) there exists $\rho_0 \in [0, 1)$ such that $|B_{2,1}(t)| \le \rho_0 \frac{|f_2(t,x_0)|}{|f_1(t,x_0)|} |B_{1,1}(t)|$.*

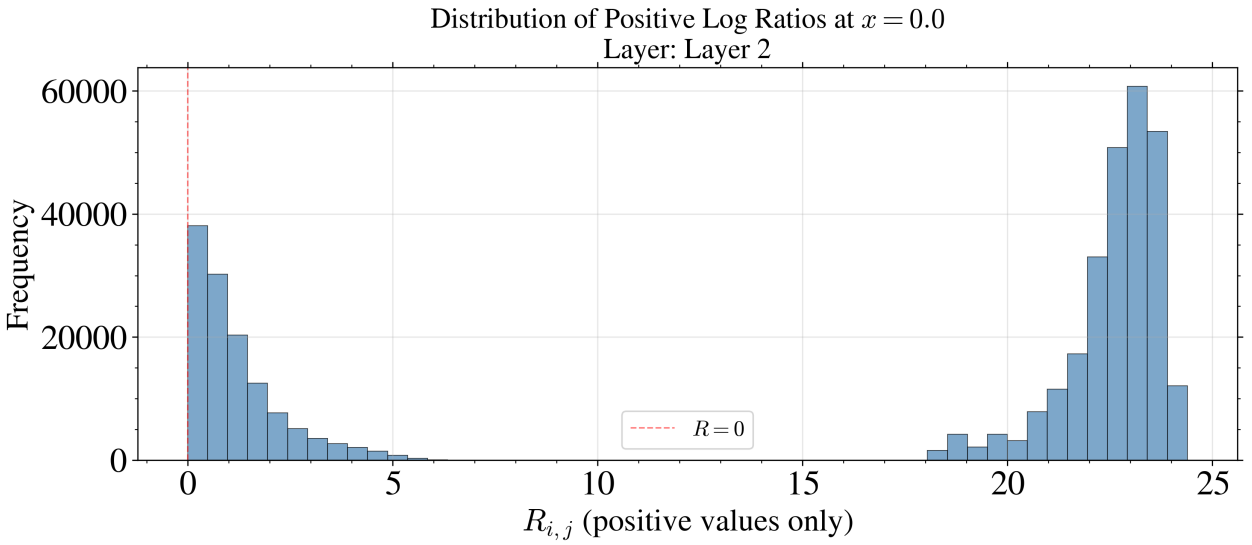


Figure 6: Log-ratios $R_{i,j}$ at $x = 0$ over training (all activation pairs; layer-2 H_k).

Conclusion: *For $t \in I$,*

$$\partial_t R_{12}(t, x_0) = (1 - \rho_0)\, \xi_1(t)\, K_{\mu_0}(x_0, x_0)\, (-d_0(t))\, \frac{|B_{1,1}(t)|}{|f_1(t, x_0)|} + \varepsilon_0(t),$$

with $|\varepsilon_0(t)| \le C'' \psi(\delta)$ for $\psi(\delta)$ fastly decaying in δ (from E_p). Hence $\partial_t R_{12}(t, x_0) \ge 0$ whenever the leading term dominates $|\varepsilon_0|$; since $K_{\mu_0}(x_0, x_1) > 0$, the off-diagonal coupling contributes with a favorable sign and reinforces the leading term, yielding even better positivity. Strict dominance of channel 1 at x_0 cannot be lost on I and is amplified whenever $|B_{1,1}|$ is not too small. The result is symmetrical at x_1 (channel 2 dominates there).

Remark 3.1 (On the hypothesis). The theorem does *not* assert that (i)–(iii) hold for the canonical two-sided step ($y = \pm A$ at $x = \pm\delta$) from generic initial conditions; it only establishes that *when* they hold on I, the conclusion follows. Verifying (i)–(iii) from the ODEs for specific initial conditions and A, δ is outside the scope of this result. See Appendix **??** for a discussion of when (i)–(iii) hold in practical settings.

Proof sketch. The proof is deferred to Appendix **??** and only uses eq:dtfk-full-delta and the log-ratio calculus at x_0. □

4 Numerical Results

font=small Our experiments focus on 1-dimensional data for three reasons: (i) large sample sizes ensure that convergence of $_Z[\cdot]$ is achieved, so optimization dynamics are not confounded by overfitting or sample complexity; (ii) the frequency-dependent target $f(x) = \cos(f_1 \pi x^2) - 0.8 \cos(f_2 \pi x^2)$ on $[-1, 1]$ provides a controlled setting to study hierarchical frequency learning, with a direct link to the two-sided step and Theorem **??**; (iii) 1D allows clear interpretability of channel specialization (spatial location and frequency) and of the log-ratio evolution.

4.1 Log-ratio growth and spike features

We validate our theoretical predictions on frequency-dependent function approximation. Assumption **??** (Appendix **??**) requires φ_2' bounded away from zero and thus excludes ReLU; in practice this can be relaxed with high probability, exponentially in r (Appendix **??**). We train 3-layer low-rank ReLU networks on $f(x) = \cos(8\pi x)$ with $n = 1024$ neurons, $r = 15$ channels, and $N = 5000$ samples. At $x = 0$ we measure layer-2 low-rank channels f_k, log-ratios $R_{i,j} = \log|f_i| - \log|f_j|$, and pre-activations H_k (with H_2 revealing the half-space separation that drives spike learning). We track $\max_{i,j} R_{i,j}(x = 0)$ during training (Figure **??**) and the distribution over all pairs (Figure **??**); the sustained

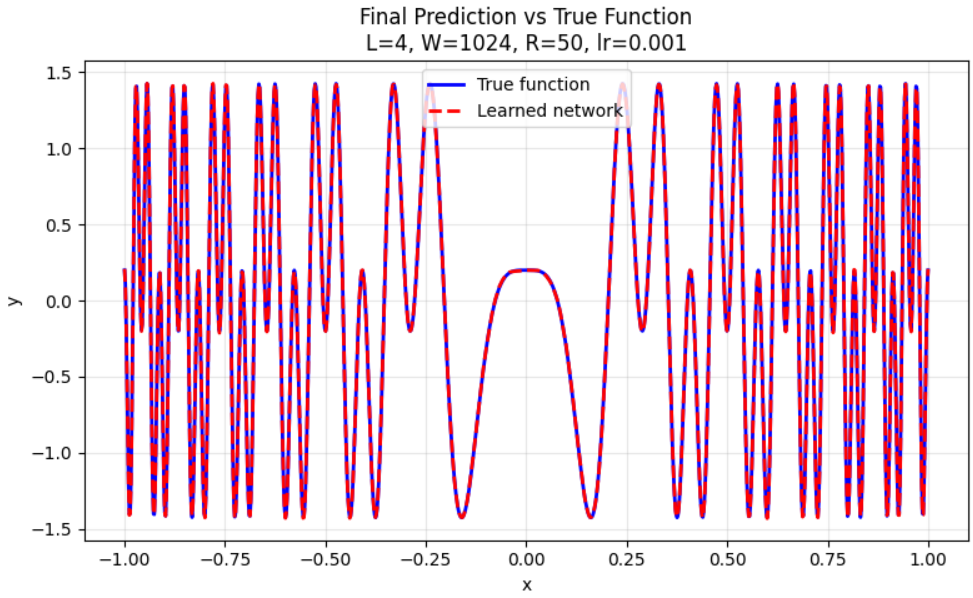


Figure 7: Final prediction vs. target $f(x) = \cos(36\pi x^2) - 0.8\cos(12\pi x^2)$. 4-layer, n=1024, r=50, N=1000, RF-LR MSEs are within 10^{-8} to $\sim 10^{-4}$

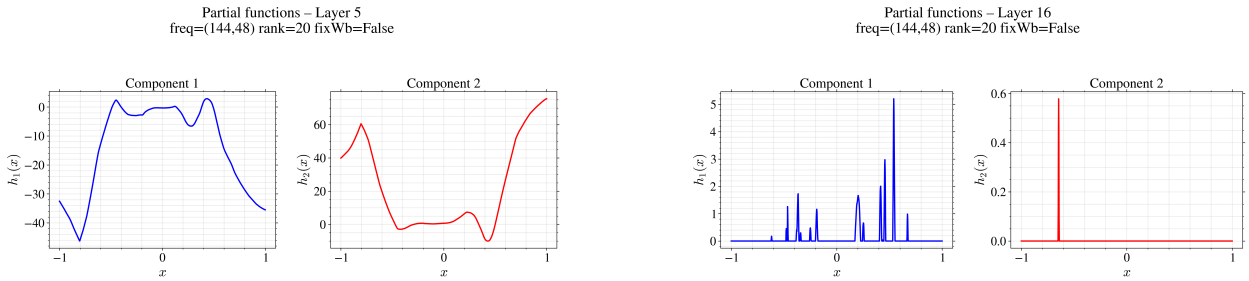


Figure 8: Asymmetry when RF is removed (a) (b) Asym. Layer 7. (c) Asym. Layer 16; r=20. (b)–(c) ; $f_1 = 144$, $f_2 = 48$.

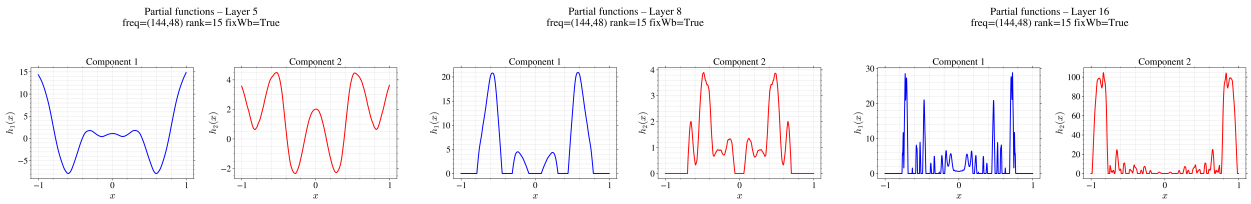


Figure 9: (symmetric) with bigger spikes. (a) $f_1^{(5)}$. (b) ReLU(H_8) (c) ReLU(H_{16}).

Table 2: Hyperparameters for all figures in this section. Symmetry runs (Figs. **??**, **??**, **??**) and Fig. **??**: Adam [**?**], batch 100, lr 0.001, γ=0.9 every 100 steps; test 4936 for symmetry. Log-ratio (Figs. **??**, **??**): SGD, lr 0.01, batch 160. Target (f_1, f_2) denotes $f(x) = \cos(f_1 \pi x^2) - 0.8\cos(f_2 \pi x^2)$ on $[-1, 1]$.

Figure	L	n	r	Target	N_{train}	Epochs	batch	RF
??, **??**	3	1024	15	$\cos(8\pi x)$	5k	10k	160	–
??	4	1024	8	(36,12)	2k	5k	100	True
??	8	1024	15	(144,48)	4k	10k	100	True
??	8	1024	20	(144,48)	4k	10k	100	False
??	4	1024	50	(36,12)	1k	1k	100	False

growth indicates that one channel increasingly dominates at $x = 0$, consistent with Theorem **??**. The mean-field weight distribution over training is shown in Figure **??**, and spike-like specialization in Figure **??**(a). Experimental setup and further details are in Appendix **??**.

4.2 Channel and activations learn symmetric spikes

We examine symmetry preservation by RF-LR on highly oscillating targets. Under batched optimization, markedly non-symmetrical features can be learned for symmetric targets. Specifically, all our targets are symmetric about $x = 0$ since both terms are even in x. When trained with SGD, low-rank networks maintain this symmetry, whereas full-rank networks show asymmetric structures. Channel feature learning in low-rank networks thus preserves geometric properties of the target, likely due to the implicit regularization of the low-rank constraint.

5 Conclusion

We have shown that when the mean-field dynamics converges, the limit is a global minimizer; this *persists* under low-rank constraints for mean-field networks of depth $L \geq 2$. Recent work has established convergence to a global minimizer for full-rank three-layer networks without assuming loss convexity, but it remained open whether these guarantees hold for more layers. Our key insight is that the low-rank and random features structure are minimal to maintain the universal approximation property throughout training, which preserves the conditions needed for the limit to be a global minimizer when training converges.

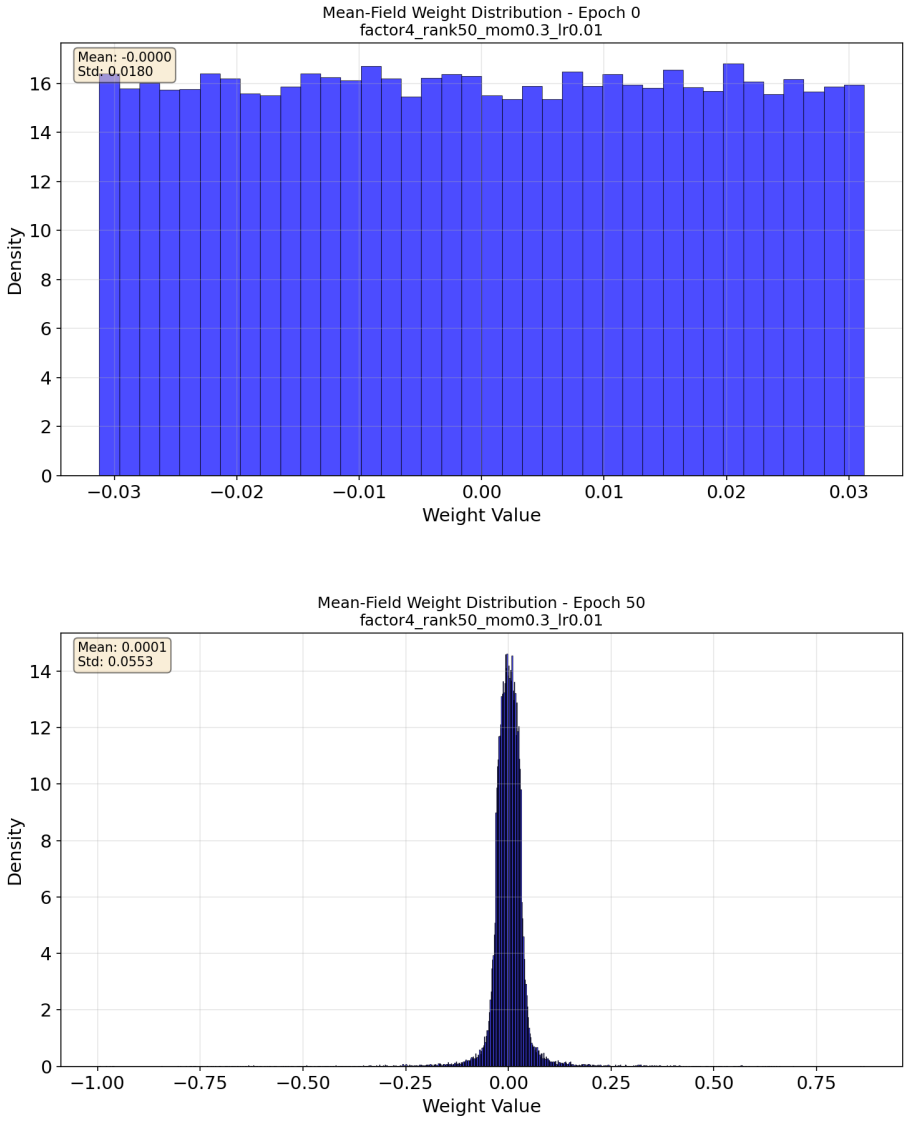


Figure 10: w_2 density at epoch 0 (top) and symmetric at 50 (bottom).

A Appendix Overview: Our Results and Their Correspondence to Nguyen et al.

This section lists all lemmas, theorems, and corollaries in our appendix and, for each, states the corresponding result in [**?**] of which it is an adaptation. Our proofs follow their structure and are modified to account for the low-rank mixing $H_\ell = \sum_{k=1}^{r} L^{(\ell)}_{c_\ell,k} f^{(\ell)}_k$, the $\|L^{(\ell)}\|_{\infty,1} \le rK$ bounds, and the $\max_{1\le k\le r}$ over channels in norms.

Assumptions. Our Assumptions **??**–**??** (Appendix **??**) and **??** (Appendix **??**) adapt the forward, backward, init, lrSchedule, and neuronal-embedding assumptions and the initialization/regularity framework in [**?**]. The low-rank mixing bound in **??** and their diversity-of-random-features assumption are specific to our RF-LR setting.

Theorem **??** and Appendices **??** and **??** are original to this work.

Reference: Nguyen et al. In [**?**], thm/lem/prop/cor share one counter. **Main text** (proofs in appendix): Lemma 8 (bounds MF a priori), Lemma 10 (difference MF). **Appendix:** Theorems 43–45; Lemmas 46–49 (Lipschitz forward/backward MF, general); Lemmas 50–53 (square Hoeffding; initialization compare; bounds NN a priori; a priori

Table 3: Our appendix lemmas/theorems/corollary and the Nguyen et al. [**?**] result each adapts.

Ours	Type	One-line statement	Adaptation of (Nguyen et al.)
??	Lemma	Bi-Lipschitz of H_ℓ, $\varphi_\ell(H_\ell)$ in W with $\|L^{(\ell)}\|_{\infty,1} \le rK$.	Lemmas 46, 48 (Lipschitz forward MF, -general)
??	Lemma	A priori $W_t \le K_0(t)$ with $(1 + rK)^{1/2}$ factor.	Lemma 8 (bounds MF a priori)
??	Lemma	$F(\mathcal{W}_T^0) \subseteq \mathcal{W}_T^0$.	Invariance in proof of Theorem 7 (existence ODE)
??	Lemma	Solution operator F contractive in $\|\cdot\|_t$.	Lemma 10 (difference MF)
??	Lemma	Sub-Gaussian bounds for $\sum_k a_k U_k$ (low-rank sums).	Lemma 51, Theorem 45 (initialization compare, iid-hilbert-higher-moment)
??	Lemma	$\|H_2\|$, $\|f_k\|$ bounds scale with $\|L\|_{\infty,1}$.	Lemma 48 (Lipschitz forward MF - general)
??	Lemma	Moment bounds for $H_2 = \sum_k L_{c_2,k} f_k$.	Lemma 52 (bounds NN a priori)
??	Lemma	$\mathscr{D}_T(W, \tilde{W}) \le \cdots e^{K_T(1+rK)}$.	Proposition 22 (particle coupling - bounded)
??	Lemma	$\mathscr{D}_T(\tilde{W}, \mathbf{W}) \le \cdots e^{K_T(1+rK)}$.	Proposition 23 (gradient descent - bounded)
??	Theorem	Dense span $\{\varphi_1(\langle L^0(c_1), \cdot\rangle)\}$ maintained (frozen L^0).	Assumption (diversity); no direct lemma (specific to frozen RF)
??	Theorem	$\mathscr{D}_T(W, \mathbf{W}) \le C_{\text{exp}} \cdot C_{\text{width}} \cdot C_{\log}$.	Corollary 17 and full quantitative framework
??	Corollary	$\|\mathbb{E}_Z[\psi(Y, \hat{\mathbf{y}})] - \mathbb{E}_Z[\psi(Y, \hat{y})]\| \le \cdots$.	Corollary 17 (gradient descent quality)

MF time difference); Theorem 54 (iid dynamics-full); Lemma 55 (full-support-2); Propositions 22, 23 (particle coupling bounded; gradient descent bounded); Corollary 17 (gradient descent quality).

B Notation and Neuronal Embedding Framework

We keep the notation and neuronal embedding framework of [**?**]. In the mean-field framework, neurons are indexed by continuous random variables rather than discrete indices. This *neuronal embedding* approach treats each neuron as a sample from a probability measure, enabling rigorous analysis in the infinite-width limit.

Neuronal indices: We use $C_1 \in \mathcal{C}_1$ and $C_2 \in \mathcal{C}_2$ to denote random variables indexing neurons in the first and second layers, respectively. These are drawn from probability measures μ_1 and μ_2 on spaces $\mathcal{C}_1$ and $\mathcal{C}_2$. In the finite-width case, $C_1(j_1)$ and $C_2(j_2)$ correspond to countable or uncountable neuron indices $j_1 \in \{1, \ldots, n_1\}$ and $j_2 \in \{1, \ldots, n_2\}$.

Weight functions: The weights are functions of time and neuronal indices:

- $w_1(t, C_1, k) \in \mathbb{R}$: weight for channel $k \in \{1, \ldots, r\}$ of neuron C_1 in the first layer at time t.
- $w_2(t, C_2) \in \mathbb{R}$: weight for neuron C_2 in the second layer at time t.
- $W(t) = (w_1(t, \cdot, \cdot), w_2(t, \cdot))$: the full weight configuration at time t.

Feature maps and activations:

- $L^0(C_1) \in \mathbb{R}^{d_0}$: frozen random feature vector for neuron C_1 (drawn i.i.d. from a Gaussian measure); we specify the law of $L^0(C_1)$ when sampling the network.
- $\varphi_1 : \mathbb{R} \to \mathbb{R}$: activation function for the first layer (e.g., Leaky ReLU or sigmoid).
- $\varphi_2 : \mathbb{R} \to \mathbb{R}$: activation function for the second layer.
- $H_2(t, c_2; X, W(t)) = \sum_{k=1}^r L_{c_2,k}\, f_k(t; X, W(t))$: second-layer pre-activation, where $f_k(t; X, W(t)) =_{C_1} [w_1(t, C_1, k)\, \varphi_1(L^0(C_1)X)]$ are the channel-wise partial functions.

Backpropagation signal (all layers): The gradient $\partial\mathscr{L}/\partial w$ at each layer is (upstream backpropagated signal) $\times$ (local derivative). Deriving $\partial_t w$ from $-\xi\, {}_Z[\cdots]$ at each layer yields:

- **Top layer (layer $L-1$):** The backpropagated signal from the loss is $D_{L-1} = d_L = \partial_{\hat{y}}\mathscr{L}(Y, \hat{y}(X; W(t)))$; it is used in the w_{L-1}-ODE.
- **Channel-k signal from layer ℓ to the layer below** ($\ell = 2, \ldots, L-1$): $B_k^{(\ell)}(t; X, W) = {}_{C_\ell}[L^{(\ell)}_{C_\ell,k}\, \varphi'_\ell(H_\ell(t, C_\ell; X, W))\, w_\ell(t, C_\ell)]$, $k \in \{1, \ldots, r\}$. This aggregates gradient information from layer ℓ through the mixing matrix $L^{(\ell)}$ and the activation derivative φ'_ℓ; for Leaky ReLU, $\varphi'_\ell(u) = 1\{u > 0\} + \alpha 1\{u \le 0\}$ with $\alpha \in (0, 1)$. In the 3-layer case, $B_k := B_k^{(2)}$.

Loss and learning rates:

- $d_L(Z; W(t)) = \partial_{\hat{y}}\mathscr{L}(Y, \hat{y}(X; W(t)))$: loss derivative with respect to the network output, where $Z = (X, Y)$ is a data sample and $\hat{y}(X; W(t))$ is the network output.
- $\xi_1(t), \xi_2(t) \ge 0$: learning rate schedules for the first and second layers, respectively.

Kernel function:

- $K_{\mu_0}(x, X)$: kernel that measures similarity between input locations x and X, induced by the untrained first-layer feature measure μ_0 (the pushforward of the initial first-layer weights). It is defined by

$$K_{\mu_0}(x, x') \equiv \int_d \varphi_1(\theta x)\, \varphi_1(\theta x')\, \mu_0(d\theta).$$

  For Leaky ReLU or sigmoid φ_1 and μ_0 the pushforward of i.i.d. Gaussian or uniform-on-sphere first-layer weights, K_{μ_0} coincides with the first-layer NNGP kernel [**?**]. This kernel appears in the evolution equation for the partial functions $f_k(t, x)$.

Expectations: ${}_Z[\cdot]$ denotes expectation over the data distribution, ${}_{C_1}[\cdot]$ over the first-layer neuron measure, and ${}_{C_2}[\cdot]$ over the second-layer neuron measure.

C Assumptions

This section contains the complete statement of all assumptions used in our theoretical analysis. These assumptions are referenced in the main text with their names in italics.

Assumption C.1 (Bounded Activations and Mixing)**.** There exists a constant $K \ge 1$ such that:

- **Activation functions**: φ_1 and φ_2 are K-Lipschitz; $\|\varphi'_2\|_\infty \le K$; and φ'_2 is **bounded away from zero**, i.e. $\inf_u |\varphi'_2(u)| \ge 1/K$. For well-posedness, $\varphi_\ell(H_\ell)$ must be bounded in the analysis; this holds for **sigmoid** and **tanh** ($\|\varphi_\ell\|_\infty \le K$); for **Leaky ReLU** $\varphi(u) = \max(u, \alpha u)$ with $\alpha \in (0, 1)$, φ is unbounded but $\varphi(H_\ell)$ is bounded when pre-activations H_ℓ are (as in our a priori ODE bounds). **ReLU is excluded** because φ' vanishes on $(-\infty, 0]$. For ReLU and a high-probability relaxation in practice, see Appendix **??**.
- **Low-rank mixing matrix**: The mixing matrix entries $L_{c_2,k}$ are random variables (e.g., Uniform) with $\sup_{c_2,k} |L_{c_2,k}| \le K$ almost surely, and $k \mapsto L_{c_2,k}$ is measurable for each c_2. This implies $\|L\|_{\infty,1} \equiv \sup_{c_2} \sum_{k=1}^r |L_{c_2,k}| \le rK$ almost surely.

Assumption C.2 (Sub-Gaussian Initialization)**.** The initial weights satisfy sub-Gaussian moment bounds:

- For the first-layer weights: $\sup_{m\ge1} \frac{1}{\sqrt{m}} \max_{1\le k\le r} {}_{C_1}\left[|w_1^0(C_1, k)|^m\right]^{1/m} \le K$ for some $K > 0$.
- For the second-layer weights: $\sup_{m\ge1} \frac{1}{\sqrt{m}} {}_{C_2}\left[|w_2^0(C_2)|^m\right]^{1/m} \le K$.
- Equivalently, in terms of ψ_2 norms: $[\![w_1(0)]\!]_\psi < \infty$ and $[\![w_2(0)]\!]_\psi < \infty$, where $[\![\cdot]\!]_\psi$ denotes the ψ_2 norm controlling moment growth.

This ensures that the initial weight distributions have controlled tail behavior, which is essential for the well-posedness argument.

Assumption C.3 (Data Distribution and Loss Regularity)**.**

- **Bounded inputs**: $|X| \le K$ with probability 1, and the feature map satisfies $\|L^0(c_1)\| \le K$ for all $c_1 \in \Omega_1$.
- **Loss**: $\partial_2\mathcal{L}(y,\cdot)$ is K-bounded and K-Lipschitz for all y in the support of $\mathcal{P}$ (for well-posedness and continuity of $\mathscr{L}$). **Loss condition:** $\partial_2\mathcal{L}(y,\hat{y}) = 0$ implies $\mathcal{L}(y,\hat{y}) = 0$. For non-negative $\mathcal{L}$, whenever $\mathbb{E}[\partial_2\mathcal{L}(Y,u)|X = x] = 0$ we then have $\mathbb{E}[\mathcal{L}(Y,u)|X = x] = 0$, so the first-order condition identifies global minimizers. Examples: MSE ($\partial_2\mathcal{L} = 0$ iff $\hat{y} = y$, and $\mathcal{L}(y,y) = 0$); many classification losses.

Assumption C.4 (Diversity of Random Features)**.** The support of ρ^1 (the measure on Ω_1 indexing the first-layer random features) is ^{d} (or dense in ^{d}). This ensures that the random features $\{\varphi_1(L^0(c_1)\cdot) : c_1 \in \Omega_1\}$ have dense span in $L^2(\mathcal{P}_X)$, which is essential for the universal approximation property.

Assumption C.5 (Non-Degeneracy)**.** To ensure that the limit point $(\bar{w}_1,\dots,\bar{w}_L)$ is non-degenerate (e.g., $\max_{1\le k\le r}(\bar{w}_1(C_1,k) \neq 0) > 0$ for the first layer and $(\bar{w}_\ell(C_\ell) \neq 0) > 0$ for $\ell = 2,\dots,L$), we require one of the following:

1. The initial loss satisfies $\mathscr{L}(w_1^0,\dots,w_L^0) <_Z [\mathcal{L}(Y,\varphi_L(0))]$. Then by the gradient flow property, the limit point must have non-zero mass for each of $\bar{w}_1,\dots,\bar{w}_L$ (e.g. $\max_{1\le k\le r}(\bar{w}_1(C_1,k) \neq 0) > 0$ and $(\bar{w}_\ell(C_\ell) \neq 0) > 0$ for $\ell = 2,\dots,L$), so $(\bar{w}_1,\dots,\bar{w}_L)$ is non-degenerate. This condition requires that the initial network performs better than the trivial predictor $\hat{y} = 0$, which is satisfied for most reasonable initializations (e.g., small random weights) with high probability.

Theorem C.1 (Universal approximation automatically maintained)**.** *The learning trajectory automatically maintains the universal approximation property of the function class represented by the first layer's neurons throughout training. Specifically, if supp$(L^0(C_1)) = \mathbb{R}^d$ and φ_1 is Leaky ReLU or sigmoid (or any non-polynomial activation), then the function class $\{\varphi_1(L^0(c_1)\cdot) : c_1 \in \Omega_1\}$ has dense span in $L^2(\mathcal{P}_X)$ throughout training.*

This follows from the fact that since $L^0(C_1)$ are frozen random features with full support, and Leaky ReLU or sigmoid is non-polynomial, the dense span property is automatically maintained: $\varphi_1(af + b)$ for random a, b always forms a dense span. This is the key property that, combined with low-rank structure and the loss condition in Assumption **??***, enables convergence to a global minimizer.*

Assumption C.6 (Convergence to Limit Point)**.** There exist functions $\bar{w}_1 : \Omega_1 \times \{1,\dots,r\} \to$ and $\bar{w}_2 : \Omega_2 \to$ such that as $t \to \infty$, there exists a coupling π_t of $\rho^1 \times \rho^2$ and itself such that:

$$\int (1 + |\bar{w}_2(c_2)|)|\bar{w}_2(c_2)| \max_{1\le k\le r} |\bar{w}_1(c_1,k)|\,|w_1^*(t,c_1',k) - \bar{w}_1(c_1,k)|\,d\pi_t(c_1,c_2,c_1',c_2') \to 0, \tag{6}$$

$$\int (1 + |\bar{w}_2(c_2)|)|\bar{w}_2(c_2)|\,|w_2^*(t,c_2') - \bar{w}_2(c_2)|\,d\pi_t(c_1,c_2,c_1',c_2') \to 0, \tag{7}$$

where $W^*(t) = (w_1^*(t,\cdot,\cdot), w_2^*(t,\cdot))$ is the solution to the mean-field ODEs (**??**). This assumption ensures that the training dynamics converge to a well-defined limit point in a Wasserstein-like sense.

Remark C.1 (On the assumptions). Universal approximation is automatically maintained (Theorem **??**), which is the crucial difference from previous work. We do not assume loss convexity; the function class maintains its approximation power throughout training. The frozen random features in the first layer with full support automatically ensure dense span: since $\varphi_1(af + b)$ for random a, b and non-polynomial φ_1 (e.g. Leaky ReLU or sigmoid) always forms a dense span, this property is automatically maintained rather than assumed, providing a rich fixed basis independent of the low-rank structure in subsequent layers.

The *Convergence to Limit Point* assumption is typically verified by showing that the loss decreases along trajectories (gradient flow property), establishing compactness of the trajectory set, and using stability arguments such as

LaSalle's invariance principle. The low-rank structure enters through the $\max_k$ over r channels in (**??**), but the overall structure remains similar to the full-rank case.

D Technical Lemmas and Proofs for Well-Posedness

The lemmas below and the techniques underlying the assumptions in the previous section are from the rigorous mean-field framework of [**?**]. We adapt them to the low-rank case by accounting for the r channels and the mixing matrix L (e.g., through $\|L\|_{\infty,1} \le rK$).

D.1 Key New Trick: Bi-Lipschitz Property

Before proving well-posedness, we establish a key technical lemma that will be used throughout the proof.

Lemma D.1 (Bi-Lipschitz property of H_ℓ, $\ell = 2, \dots, L-1$)**.** *For any $W' = (w'_1, \dots, w'_{L-1})$ and $W'' = (w''_1, \dots, w''_{L-1})$ (in the 3-layer case $W = (w_1, w_2)$) satisfying the regularity assumptions, we have the following for each $\ell = 2, \dots, L-1$.*

***Layer** $\ell = 2$.*

$$\left|H_2(t, c_2; X, W') - H_2(t, c_2; X, W'')\right| \le K \, \|L^{(2)}\|_{\infty,1} \max_{1 \le k \le r} {}_{C_1}\left[|w'_1(t, C_1, k) - w''_1(t, C_1, k)|\right],$$

where $\|L^{(\ell)}\|_{\infty,1} \equiv \sup_{c_\ell} \sum_{k=1}^r |L^{(\ell)}_{c_\ell,k}| \le rK$ under the entrywise bound ($L^{(2)} = L$). If φ_2 is Leaky ReLU or Lipschitz, then

$$\left|\varphi_2(H_2(t, c_2; X, W')) - \varphi_2(H_2(t, c_2; X, W''))\right| \le K \, \|L^{(2)}\|_{\infty,1} \max_{1 \le k \le r} {}_{C_1}\left[|w'_1(t, C_1, k) - w''_1(t, C_1, k)|\right].$$

***Layers** $\ell = 3, \dots, L-1$.*

$$\left|H_\ell(t, c_\ell; X, W') - H_\ell(t, c_\ell; X, W'')\right| \le K \, \|L^{(\ell)}\|_{\infty,1} \left({}_{C_{\ell-1}}\left[|w'_{\ell-1}(t, C_{\ell-1}) - w''_{\ell-1}(t, C_{\ell-1})|\right] + w''_{\ell-1,t} \cdot \mathscr{B}_{\ell-1}(W', W'') \right),$$

where $\mathscr{B}_{\ell-1}(W', W'')$ is the RHS of the bi-Lipschitz inequality for $H_{\ell-1}$ above. If φ_ℓ is Leaky ReLU or Lipschitz, the same bound holds for $|\varphi_\ell(H_\ell(t, c_\ell; X, W')) - \varphi_\ell(H_\ell(t, c_\ell; X, W''))|$.

Proof. **Layer $\ell = 2$.** By definition $H_2(t, c_2; X, W) = \sum_{k=1}^r L_{c_2,k} \, f_k(t; X, W)$ with $f_k(t; X, W) =_{C_1} [w_1(t, C_1, k) \, \varphi_1(L^0(C_1)X)]$. Then

$$\begin{aligned} H_2(t, c_2; X, W') - H_2(t, c_2; X, W'') &= \sum_{k=1}^r L_{c_2,k} \big(f_k(t; X, W') - f_k(t; X, W'')\big) \\ &= \sum_{k=1}^r L_{c_2,k} \, {}_{C_1}\left[(w'_1(t, C_1, k) - w''_1(t, C_1, k)) \, \varphi_1(L^0(C_1)X)\right]. \end{aligned}$$

Taking absolute values and using boundedness of φ_1:

$$\begin{aligned} \left|H_2(t, c_2; X, W') - H_2(t, c_2; X, W'')\right| &\le \sum_{k=1}^r |L_{c_2,k}| \, {}_{C_1}\left[|w'_1(t, C_1, k) - w''_1(t, C_1, k)|\right] |\varphi_1(L^0(C_1)X)| \\ &\le K \sum_{k=1}^r |L_{c_2,k}| \max_{1 \le k \le r} {}_{C_1}\left[|w'_1(t, C_1, k) - w''_1(t, C_1, k)|\right] \\ &\le K \, \|L^{(2)}\|_{\infty,1} \max_{1 \le k \le r} {}_{C_1}\left[|w'_1(t, C_1, k) - w''_1(t, C_1, k)|\right]. \end{aligned}$$

The $\varphi_2(H_2)$ part follows from the Lipschitz property of φ_2.

Layers $\ell \ge 3$. We have $H_\ell(t, c_\ell; X, W) = \sum_{k=1}^r L^{(\ell)}_{c_\ell,k} \, f^{(\ell)}_k(t; X, W)$ with $f^{(\ell)}_k(t; X, W) =_{C_{\ell-1}} [w_{\ell-1}(t, C_{\ell-1}, k) \, \varphi_{\ell-1}(H_{\ell-1}(t, C_{\ell-1}; X, W))]$ (and $w_{\ell-1}(t, C_{\ell-1}, k) \equiv w_{\ell-1}(t, C_{\ell-1})$ when the layer has no channel index). Then

$$f^{(\ell)}_k(W') - f^{(\ell)}_k(W'') =_{C_{\ell-1}} \left[(w'_{\ell-1} - w''_{\ell-1}) \, \varphi_{\ell-1}(H_{\ell-1}(W')) + w''_{\ell-1} \left(\varphi_{\ell-1}(H_{\ell-1}(W')) - \varphi_{\ell-1}(H_{\ell-1}(W''))\right)\right].$$

Using $|\varphi_{\ell-1}| \le K$, $|\varphi_{\ell-1}(a) - \varphi_{\ell-1}(b)| \le K|a-b|$, and the inductive bound for $|H_{\ell-1}(W') - H_{\ell-1}(W'')|$:

$$\begin{aligned}|f_k^{(\ell)}(W') - f_k^{(\ell)}(W'')| &\le K\, C_{\ell-1}\big[|w'_{\ell-1} - w''_{\ell-1}|\big] + K\, C_{\ell-1}\big[|w''_{\ell-1}|\,|H_{\ell-1}(W') - H_{\ell-1}(W'')|\big] \\ &\le K\, C_{\ell-1}\big[|w'_{\ell-1} - w''_{\ell-1}|\big] + K\, w''_{\ell-1 t}\, \mathscr{B}_{\ell-1}(W', W'').\end{aligned}$$

Thus $|H_\ell(W') - H_\ell(W'')| \le \|L^{(\ell)}\|_{\infty,1} \max_k |f_k^{(\ell)}(W') - f_k^{(\ell)}(W'')|$ yields the claim. The $\varphi_\ell(H_\ell)$ part follows from the Lipschitz property of φ_ℓ. □

From here, the proof follows as usual. For each $\ell = 2, \dots, L-1$, $H_\ell(t, c_\ell; X, W)$ bi-Lipschitz in W is *sufficient and equivalent to* φ_ℓ Lipschitz; that is the only thing that matters. The remaining is simple: after defining norms, we update $K_0(t)$ with a factor $(1 + rK)^{1/2}$; the solution operator F, a priori bounds, and the contraction argument then proceed as in the full-rank framework [**?**].

D.2 Norms and Spaces

We equip the mean-field parameters with several norms. For the low-rank case: $w_{1t} = \max_{1\le k\le r} C_1[\sup_{s\le t} |w_1(s, C_1, k)|^{50}]^{1/50}$, $w_{2t} =_{C_2} [\sup_{s\le t} |w_2(s, C_2)|^{50}]^{1/50}$, $W_t = \max(w_{1t}, w_{2t})$. L^2-type: $\|w_1\|_t = \max_{1\le k\le r} C_1[\sup_{s\le t} |w_1(s, C_1, k)|^2]^{1/2}$, $\|w_2\|_t =_{C_2} [\sup_{s\le t} |w_2(s, C_2)|^2]^{1/2}$, $\|W\|_t = \max(\|w_1\|_t, \|w_2\|_t)$. ψ_2-type: $[\![w_1]\!]_{\psi,t} = \sqrt{50}\, \sup_{m\ge 1} \frac{1}{\sqrt{m}} \max_{1\le k\le r} C_1[\sup_{s\le t} |w_1(s, C_1, k)|^m]^{1/m}$, $[\![w_2]\!]_{\psi,t} = \sqrt{50}\, \sup_{m\ge 1} \frac{1}{\sqrt{m}} C_2[\sup_{s\le t} |w_2(s, C_2)|^m]^{1/m}$, $[\![W]\!]_{\psi,t} = \max([\![w_1]\!]_{\psi,t}, [\![w_2]\!]_{\psi,t})$. The factor $\sqrt{50}$ ensures $[\![W]\!]_{\psi,t} \ge W_t$ and $[\![W]\!]_{\psi,t} \ge \|W\|_t$. Random variables: $\mathsf{max}_t^w(W) = \max_{1\le k\le r} \sup_{s\le t} |w_1(s, C_1, k)|$, $\mathsf{max}_t^{w_2}(W) = \sup_{s\le t} |w_2(s, C_2)|$. Distance for W, W':

$$\|W - W'\|_t = \max(\|w_1 - w'_1\|_t, \|w_2 - w'_2\|_t), \tag{8}$$

with $\|w_1 - w'_1\|_t = \max_{1\le k\le r} C_1[\sup_{s\le t} |w_1(s, C_1, k) - w'_1(s, C_1, k)|^2]^{1/2}$ and $\|w_2 - w'_2\|_t =_{C_2} [\sup_{s\le t} |w_2(s, C_2) - w'_2(s, C_2)|^2]^{1/2}$.

D.3 Solution Operator and Fixed Point Formulation

Denote by $\mathfrak{W}_T$ the space of mean-field parameters W with $\|W\|_T < \infty$. Given $T \ge 0$ and $W(0)$, we define F mapping $W' \in \mathfrak{W}_T$ to $F(W')(t) = \{F_1(W')(t, \cdot, \cdot), F_2(W')(t, \cdot)\}$, where $F_1(W')(t, c_1, k) = w_1(0, c_1, k) - \int_0^t \xi_1(s)_Z[d_L(Z; W'(s))\, \varphi_1(L^0(c_1)X)\, B_k(s; X, W'(s))]ds$ and $F_2(W')(t, c_2) = w_2(0, c_2) - \int_0^t \xi_2(s)_Z[d_L(Z; W'(s))\, \varphi_2(H_2(s, c_2; X, W'(s)))]ds$, with $B_k(s; X, W') =_{C_2} [L_{C_2,k}\, \varphi_2'(H_2(s, C_2; X, W'))\, w_2'(s, C_2)]$. At initialization $F(W')(0, \cdot, \cdot) = W(0)$; the time integrals use W'. A solution on $[0, T]$ is $W \in \mathfrak{W}_T$ with $F(W) = W$. We say W is a solution on $[0, \infty)$ if its restriction to $[0, T]$ is a solution for all $T > 0$.

D.4 A Priori Bounds

Lemma D.2 (Weight bounds)**.** *Under Assumptions* **??** *and* **??**, *given an initialization $W(0)$, a solution W to the mean-field ODEs, if it exists, must satisfy that for any $t \in [0, \infty)$:*

$$W_t \le K_0(t),$$

where $K_0(t)$ is a non-decreasing function of the form

$$K_0(t) = (1 + rK)^{1/2}\, K^\kappa (1 + t^\kappa)(1 + W_0^\kappa),$$

for some constant $\kappa > 0$ depending on K and r.

A similar result holds for the ψ_2 norm. Under the same assumptions, for any $t \in [0, \infty)$, there exists $K_0(t) \ge 1$ of the form

$$K_0(t) = (1 + rK)^{1/2}\, K^\kappa (1 + t^\kappa)(1 + [\![W]\!]_{\psi,0}^\kappa),$$

such that a solution W, if it exists, must satisfy $W_t \le [\![W]\!]_{\psi,t} \le K_0(t)$ for any $t \in [0, \infty)$. Furthermore, by assuming $[\![W]\!]_{\psi,0} < \infty$, for any $B \ge 0$:

$$\big(\max\big(\mathsf{max}_t^w(W), \mathsf{max}_t^{w_2}(W)\big) \ge K_0(t)B\big) \le Ce^{-K_1 B^2},$$

for some universal constants $C, K_1 > 0$.

Proof. We do not provide this proof here because of conciseness to prove sub-gaussian bounds for trivial lipschiz variable. The proof follows from Grönwall-type arguments applied to the ODEs (**??**). The key steps are:

1. Use the boundedness and Lipschitz properties of φ_1, φ_2, φ_2', and $\partial_2 \mathcal{L}$.
2. Control H_2 using the low-rank structure: $|H_2| \le \|L\|_{\infty,1} \max_k |f_k| \le rK \max_{k\, C_1} [|w_1(C_1, k)|]$.
3. Control B_k using the structure: $|B_k| \le K_{C_2}[|L_{C_2,k}|\,|w_2|] \le rK_{C_2}^2[|w_2|]$.
4. Apply Minkowski's inequality and Grönwall's lemma to obtain polynomial growth in t.
5. The sub-Gaussian tail bound follows from the ψ_2 control and a union bound.

The constant κ depends on K and r through the factor $\|L\|_{\infty,1} \le rK$, but remains polynomial rather than exponential in r. □

These a priori bounds lead us to consider the following spaces, given an initialization $W(0)$ and an arbitrary terminal time $T > 0$:

- The space $\mathcal{W}_T$ of mean-field parameters $W' = \{W'(t)\}_{t \le T}$ such that $W'_T \le K_0(T)$.
- The space $\mathcal{W}_T^0 \subset \mathcal{W}_T$ of mean-field parameters $W' \in \mathcal{W}_T$ such that:

$$[\![W']\!]_{\psi,T} \le K_0(T),$$
$$\left(\max\left(\mathsf{max}_T^w(W'), \mathsf{max}_T^{w_2}(W')\right) \ge K_0(T)B\right) \le Ce^{-K_1B^2} \quad \forall B \ge 0,$$

  and $W'(0) = W(0)$ (so all elements in $\mathcal{W}_T^0$ share the same initialization).
  It is easy to see that $\mathcal{W}_T^0 \subset \mathcal{W}_T$ since $W'_T \le [\![W']\!]_{\psi,T}$.

We equip these spaces with the metric $(W', W'') \mapsto \|W' - W''\|_T$. By Lemma **??**, any solution W to the mean-field ODEs, if it exists, must belong to $\mathcal{W}_T^0$.

Lemma D.3 (Solution operator maps bounded sets to bounded sets)**.** *Under Assumptions* **??** *and* **??**, *for any $W' \in \mathcal{W}_T^0$, we have $F(W') \in \mathcal{W}_T^0$.*

Proof. The proof follows the same argument as Lemma **??**, using the integral form defining F and the bounded/Lipschitz properties of the drifts. The key is that applying bounded/Lipschitz drifts to W' preserves the moment and tail bounds with constants controlled by $K_0(T)$. □

D.5 Difference Estimate

Lemma D.4 (Solution operator is contractive)**.** *For a given $B \ge 0$, consider two collections of mean-field parameters $W', W'' \in \mathcal{W}_T$ such that:*

$$\left(\max\left(\mathsf{max}_T^w(W'), \mathsf{max}_T^{w_2}(W')\right) \ge K_0(T)B\right) \le Ce^{-K_1B^2},$$
$$\left(\max\left(\mathsf{max}_T^w(W''), \mathsf{max}_T^{w_2}(W'')\right) \ge K_0(T)B\right) \le Ce^{-K_1B^2}.$$

Under Assumptions **??**–**??**, *for any $t \le T$:*

$$\|F(W') - F(W'')\|_t \le (KK_0(T))^4 \int_0^t \left((1+B)\|W' - W''\|_s + \sqrt{2}e^{-K_1B^2/2}\right) ds,$$

where the constant depends on K and r through $\|L\|_{\infty,1} \le rK$.

Proof. The proof uses a good/bad event decomposition:

1. On the good event $\{\max(\mathsf{max}_T^w, \mathsf{max}_T^{w_2}) \le K_0(T)B\}$, the drifts are Lipschitz in W with constant $(1+B)K_0(T)$.

2. The difference $|H_2(W') - H_2(W'')|$ is controlled using Lemma **??**:

$$|H_2(W') - H_2(W'')| \le K\|L\|_{\infty,1} \max_k {}_{C_1}[|w_1'(C_1,k) - w_1''(C_1,k)|] \le rK^2 \max_k \|w_1' - w_1''\|_t.$$

3. The difference $|B_k(W') - B_k(W'')|$ is controlled similarly, giving a factor $(1+B)$.

4. The bad event contributes an exponentially small remainder $e^{-K_1B^2/2}$.

5. Integrating in time and using Minkowski's inequality yields the result.

□

D.6 Complete Proof of Theorem ??

Proof of Theorem **??**. We perform a Picard-type iteration argument. Consider an arbitrary finite $T \ge 0$ and $W', W'' \in \mathcal{W}_T^0$. From Lemma **??**:

$$\begin{aligned}\|F(W') - F(W'')\|_t &\le (KK_0(T))^4 \left((1+B)\int_0^t \|W' - W''\|_s ds + T\sqrt{2}e^{-K_1B^2/2}\right)\\ &\equiv k_1(1+B)\int_0^t \|W' - W''\|_s ds + k_2 e^{-k_3B^2},\end{aligned}$$

for any $B > 0$, where $k_1 = (KK_0(T))^4$, $k_2 = (KK_0(T))^4 T\sqrt{2}$, and $k_3 = K_1/2$.

By Lemma **??**, F maps $\mathcal{W}_T^0$ to $\mathcal{W}_T^0$. We can iterate this inequality to obtain:

$$\begin{aligned}\|F^{(m)}(W') - F^{(m)}(W'')\|_T &\le k_1(1+B)\int_0^T \|F^{(m-1)}(W') - F^{(m-1)}(W'')\|_{T_2} dT_2 + k_2 e^{-k_3B^2}\\ &\le k_1^2(1+B)^2 \int_0^T \int_0^{T_2} \|F^{(m-2)}(W') - F^{(m-2)}(W'')\|_{T_3} \mathbb{I}(T_2 \le T) dT_3 dT_2\\ &\quad + k_2 \sum_{\ell=1}^{2} \frac{(Tk_1(1+B))^{\ell-1}}{\ell!} e^{-k_3B^2}\\ &\quad \cdots\\ &\le \frac{1}{m!} T^m k_1^m (1+B)^m \|W' - W''\|_T + k_2 e^{Tk_1(1+B) - k_3B^2}\\ &\le \frac{1}{m!} T^m k_1^m (1+\sqrt{m})^m \|W' - W''\|_T + k_2 e^{Tk_1(1+\sqrt{m}) - k_3 m},\end{aligned}$$

where we choose $B = \sqrt{m}$ in the last display. Note that since $W_0 < \infty$, $K_0(T)$ and hence k_1, k_2 are finite for finite T.

By substituting $W'' = F(W')$, we obtain:

$$\sum_{m=1}^{\infty} \|F^{(m+1)}(W') - F^{(m)}(W')\|_T = \sum_{m=1}^{\infty} \|F^{(m)}(W'') - F^{(m)}(W')\|_T < \infty.$$

Hence as $m \to \infty$, $F^{(m)}(W')$ converges in $\|\cdot\|_T$ to a limit $W \in \mathfrak{W}_T$, which is a fixed point of F. By Lemma **??**, W belongs to $\mathcal{W}_T^0$.

The uniqueness of the fixed point comes from the above estimate, since if W' and W'' are fixed points of F, then they are both in $\mathcal{W}_T^0$, and:

$$\|W' - W''\|_T = \|F^{(m)}(W') - F^{(m)}(W'')\|_T \le \frac{1}{m!} T^m k_1^m (1+\sqrt{m})^m \|W' - W''\|_T + k_2 e^{Tk_1(1+\sqrt{m}) - k_3 m},$$

and one can take m arbitrarily large. This proves that the solution exists and is unique on $t \in [0,T]$. Since T is arbitrary, we have existence and uniqueness of the solution to the mean-field ODEs on the time interval $[0,\infty)$. □

E Lemma and proofs of convergence

E.1 Channel Mixing and Low-Rank Structure

The low-rank channel mixing structure enters through:

$$H_2(t, c_2; X, W) = \sum_{k=1}^{r} L_{c_2,k}\, f_k(t; X, W),$$

where $f_k(t; X, W) =_{C_1} [w_1(t, C_1, k)\, \varphi_1(L^0(C_1)X)]$ are the r partial functions.

Under the entrywise bound $\sup_{c_2,k} |L_{c_2,k}| \le K$, we have:

$$\|L\|_{\infty,1} \le rK, \qquad \|L\|_{\infty,2} \le \sqrt{r}\, K.$$

Lemma E.1 (Sub-Gaussian bounds for low-rank sums)**.** *Let $(U_k)_{k=1}^r$ be real random variables on a common probability space and $(a_k)_{k=1}^r \in^r$ deterministic. Then for every $m \ge 1$,*

$$\Big[\Big|\sum_{k=1}^{r} a_k U_k\Big|^m\Big]^{1/m} \le \sum_{k=1}^{r} |a_k|\, [|U_k|^m]^{1/m} \le \Big(\sum_{k=1}^{r} |a_k|\Big) \max_{1\le k\le r} [|U_k|^m]^{1/m}.$$

Consequently, any "ψ_2-type" seminorm defined by $\sup_{m\ge1} m^{-1/2}[|\cdot|^m]^{1/m}$ satisfies

$$\sup_{m\ge1} \frac{1}{\sqrt{m}} \Big[\Big|\sum_{k=1}^{r} a_k U_k\Big|^m\Big]^{1/m} \le \Big(\sum_{k=1}^{r} |a_k|\Big) \max_{1\le k\le r} \sup_{m\ge1} \frac{1}{\sqrt{m}} [|U_k|^m]^{1/m}.$$

If one prefers an ℓ_2 version, then also $\sum_{k=1}^r |a_k| \le \sqrt{r}\,(\sum_k a_k^2)^{1/2}$ yields a $\sqrt{r}$ factor.

Lemma E.2 (Forward propagation bounds scale with mixing matrix norm)**.** *Assume φ_1 is bounded by K. Then for every $t \ge 0$ and every $c_2 \in \Omega_2$,*

$$\sup_{s\le t} \big|H_2(s, c_2; X, W)\big| \le \Big(\sum_{k=1}^{r} |L_{c_2,k}|\Big) \max_{1\le k\le r} \sup_{s\le t} |f_k(s; X, W)| \le \|L\|_{\infty,1} \max_{1\le k\le r} \sup_{s\le t} |f_k(s; X, W)|.$$

Moreover,

$$\sup_{s\le t} |f_k(s; X, W)| \le_{C_1} \Big[\sup_{s\le t} |w_1(s, C_1, k)|\, \big|\varphi_1(L^0(C_1)X)\big|\Big] \le K\, {}_{C_1}\Big[\sup_{s\le t} |w_1(s, C_1, k)|\Big].$$

Combining these inequalities and taking moments yields, for every $m \ge 1$,

$$X\Big[\sup_{s\le t} \sup_{c_2} |H_2(s, c_2; X, W)|^m\Big]^{1/m} \le K\, \|L\|_{\infty,1} \max_{1\le k\le r} {}_{C_1}\Big[\sup_{s\le t} |w_1(s, C_1, k)|^m\Big]^{1/m}.$$

In particular, in the ψ_2-type calibration,

$$\sqrt{50} \sup_{m\ge1} \frac{1}{\sqrt{m}}\, X\Big[\sup_{s\le t} \sup_{c_2} |H_2(s, c_2; X, W)|^m\Big]^{1/m} \le K\, \|L\|_{\infty,1}\, [\![w_1]\!]_{\psi,t}.$$

E.2 Proof of Theorem ?? (global minimizer at limit, any depth)

This subsection adapts the core argument of [**?**], Sec. 6.3 (Proof of Theorem 34), to our low-rank setting for any depth $L \ge 2$. We omit the homotopy argument (their Lemma 37): the first-layer feature map $L^0(C_1)$ is *frozen*, so $\mathrm{supp}(L^0(C_1)) =^d$ at init and at all t, and dense span follows without a homotopy. The main-text

High-level proof idea and [**?**] Sec. 6.2.1 summarize the idea; we give the formal steps.

Dense span without homotopy. $L^0(C_1)$ is not trained. By Assumption **??** and Theorem **??**, $\{\varphi_1(L^0(c_1)\cdot) : c_1 \in \Omega_1\}$ has dense span in $L^2(\mathcal{P}_X)$ at all $t \ge 0$ and at the limit.

Zero derivative at the limit. Let $\bar{W} = (\bar{w}_1, \ldots, \bar{w}_{L-1})$ be a limit point under Assumption **??** (and its natural extension to L layers with couplings for all w_ℓ). At the limit, $\partial_t w_\ell = 0$ for all ℓ. From (**??**), for the top layer $\ell = L-1$:

$$_Z\big[d_L(Z;\bar{W})\,\varphi_{L-1}(H_{L-1}(c_{L-1};X,\bar{W}))\big] = 0, \qquad \forall c_{L-1}.$$

By backpropagation, the first-layer ODE yields

$$_Z\big[d_L(Z;\bar{W})\,\varphi_1(L^0(c_1)X)\,B_k^{(2)}(X;\bar{W})\big] = 0, \qquad c_1 \in \mathrm{supp}(\rho^1),\ k \in \{1,\ldots,r\},$$

where $B_k^{(2)}(X;\bar{W}) =_{C_2} [L_{C_2,k}\,\varphi_2'(H_2(C_2;X,\bar{W}))\,\bar{w}_2(C_2)]$ (and for $L>3$, $B_k^{(\ell)}$ involves $H_\ell, \bar{w}_\ell, L^{(\ell)}$ in the same way). Since $B_k^{(2)}$ depends only on X and $\bar{W}$, we have

$$_X\big[_{Y|X}[d_L(Z;\bar{W}) \mid X]\,\varphi_1(L^0(c_1)X)\,B_k^{(2)}(X;\bar{W})\big] = 0 \quad \forall c_1, k.$$

The function $x \mapsto_{Y|X} [d_L \mid X = x]\,B_k^{(2)}(x;\bar{W})$ thus has zero inner product in $L^2(\mathcal{P}_X)$ with $\varphi_1(L^0(c_1)\cdot)$ for all c_1. By dense span, $_{Y|X}[d_L(Z;\bar{W}) \mid X = x]\,B_k^{(2)}(x;\bar{W}) = 0$ for $\mathcal{P}_X$-a.e. x.

From integrated identity to $\partial_{\hat{y}}\mathscr{L}$ a.e. We have $_{Y|X}[d_L(Z;\bar{W}) \mid X = x]\,B_k^{(2)}(x;\bar{W}) = 0$ for $\mathcal{P}_X$-a.e. x. Under Assumption **??**, $\max_{1\le k\le r}(\bar{w}_1(C_1,k) \neq 0) > 0$ and $(\bar{w}_\ell(C_\ell) \neq 0) > 0$ for $\ell = 2,\ldots,L-1$; by Assumption **??**, $\varphi_2'(H_2(c_2;x,\bar{W})) \neq 0$ for $\mathcal{P}_X$-a.e. x and ρ^2-a.e. c_2. Hence for $\mathcal{P}_X$-a.e. x and almost every c_2, the factor in $B_k^{(2)}$ involving φ_2' is non-zero, so $B_k^{(2)}(x;\bar{W})$ is non-zero on a set of positive $\mathcal{P}_X$-measure. On that set, $_{Y|X}[d_L(Z;\bar{W}) \mid X = x] = 0$, and thus $_Z[\partial_{\hat{y}}\mathscr{L}(Y,\hat{y}(X;\bar{W})) \mid X = x] = 0$ for $\mathcal{P}_X$-a.e. x.

On ReLU and relaxing the activation-derivative assumption. Assumption **??** requires φ_2' bounded away from zero and excludes ReLU. At convergence, we are at a stationary point: $\partial_t w_\ell = 0$ for all ℓ, so the ODE right-hand sides vanish and *all backpropagated signals contribute zero to the gradient*—training has converged because the gradient is zero. The step above uses $\varphi_2'(H_2) \neq 0$ to deduce $B_k^{(2)}(x;\bar{W}) \neq 0$ on a set of positive measure and thus $[d_L \mid X = x] = 0$ a.e. The only problematic case for ReLU is when the limit $\bar{W}$ is such that $B_k^{(2)} = 0$ $\mathcal{P}_X$-a.e. purely because $\varphi_2'(H_2) = 0$ everywhere on the relevant support (for ReLU, $H_2 \le 0$). That is convergence to a *degenerate* point where the gradient vanishes solely because all backprop through the activation derivative are zero. This event is exponentially rare in the rank r: it requires the pre-activation configuration (over r channels and the layer width) to lie in a degenerate set, and the probability of landing there is of order $e^{-O(r)}$.

Optimality (Assumption ??). We have $[\partial_2\mathcal{L}(Y,\hat{y}(X;\bar{W})) \mid X = x] = 0$ for $\mathcal{P}_X$-a.e. x. By the loss condition, $\mathbb{E}[\partial_2\mathcal{L}(Y,u)|X = x] = 0$ implies $\mathbb{E}[\mathcal{L}(Y,u)|X = x] = 0$. Thus $[\mathcal{L}(Y,\hat{y}(X;\bar{W})) \mid X = x] = 0$ for $\mathcal{P}_X$-a.e. x, so $\mathscr{L}(\bar{W}) = 0$. For non-negative $\mathcal{L}$, $\bar{W}$ is a global minimizer.

$\mathscr{L}(W(t)) \to \mathscr{L}(\bar{W})$. By Assumption **??**, the couplings π_t and the Wasserstein-like integrals (**??**)–(**??**) (and their L-layer analogues) tend to 0. The output difference $|\hat{y}(X;W(t)) - \hat{y}(X;\bar{W})|$ is bounded by a K-multiple of those integrals (via the low-rank structure: $H_2 = \sum_k L_{c_2,k} f_k$, $B_k^{(\ell)}$, and the regularity of φ_ℓ, $\partial_2\mathcal{L}$). Thus $\mathscr{L}(W(t)) - \mathscr{L}(\bar{W}) =_Z [\mathcal{L}(Y,\hat{y}(X;W(t))) - \mathcal{L}(Y,\hat{y}(X;\bar{W}))]$ is bounded by $K_Z[|\hat{y}(X;W(t)) - \hat{y}(X;\bar{W})|] \to 0$ as $t \to \infty$.

F Detailed Proof Sketches

F.1 Well-Posedness: Picard Iteration Details

The key technical result is that the low-rank structure only multiplies constants by r:

Lemma F.1 (Moment bounds for low-rank sums)**.** *For the low-rank sum $H_2(t,c_2;X,W) = \sum_{k=1}^r L_{c_2,k}\,f_k(t;X,W)$, we have*

$$_X\Big[\sup_{s\le t}\sup_{c_2}|H_2(s,c_2;X,W)|^m\Big]^{1/m} \le K\,\|L\|_{\infty,1}\max_{1\le k\le r}{}_{C_1}\Big[\sup_{s\le t}|w_1(s,C_1,k)|^m\Big]^{1/m},$$

where $\|L\|_{\infty,1} \le rK$ under entrywise bounds.

This lemma shows that the forward propagation bounds are multiplied by at most rK, but the structure of the Picard iteration remains unchanged. The contraction mapping argument proceeds as in the full-rank case, with constants depending on $\|L\|_{\infty,1}$.

G Quantitative Approximation by the Mean-Field Limit for Low-Rank Networks

This appendix provides a rigorous quantitative bound on the approximation error between finite-width low-rank networks and their mean-field limit. The key result shows that the approximation error scales as $O(1/\sqrt{n_{\min}} + \sqrt{\epsilon})$ with explicit constants, where $n_{\min}$ is the minimum width across layers and ϵ is the learning rate step size. This bound holds for any n_1 and n_2, independent of the data dimension d, similar to the full-rank case [**?**, **?**]. The bound suggests that widths $n_1, n_2 \approx 1000$ are typically sufficient to observe mean-field behaviors, as empirically validated in [**?**] for high-dimensional real-world data.

G.1 Main Result: Approximation by the MF Limit

Assumption G.1 (Initialization for Low-Rank Networks)**.** We assume that ess-sup $\max_{1\le k\le r} |w_1^0(C_1, k)| \le K$ and ess-sup$|w_2^0(C_2)| \le K$, where w_1^0 and w_2^0 are the initial weights as described in the low-rank architecture setup.

Theorem G.1 (Finite-width approximation error bound)**.** *Given a family* Init *of initialization laws and a tuple* $\{n_1, n_2\}$ *that is in the index set of* Init*, perform the coupling procedure for the low-rank architecture as described in Section* **??**. *Fix a terminal time* $T \in \epsilon\mathbb{N}_{\ge 0}$. *Under Assumptions* **??**, **??**, *and the low-rank structure with mixing matrix* L *satisfying* $\|L\|_{\infty,1} \le rK$, *for* $\epsilon \le 1$, *we have with probability at least* $1 - 2\delta$,

$$\mathscr{D}_T(W, \mathbf{W}) \le C_{\exp} \cdot C_{width} \cdot C_{\log},$$

where $C_{\exp} = e^{K_T(1+rK)}$, $C_{width} = 1/\sqrt{n_{\min}} + \sqrt{\epsilon}$, $C_{\log} = \sqrt{\log(3(T+1)n_{\max}^2/\delta + e)}$, *with* $n_{\min} = \min\{n_1, n_2\}$, $n_{\max} = \max\{n_1, n_2\}$, $K_T = K(1 + T^K)$, *and the factor* $(1 + rK)$ *accounts for the low-rank structure through* $\|L\|_{\infty,1} \le rK$.

The theorem gives a connection between $\mathbf{W}(\lfloor t/\epsilon \rfloor)$ (the discrete-time finite-width network) and $W(t)$ (the continuous-time mean-field limit). The key difference from the full-rank case [**?**] is the multiplicative factor $(1 + rK)$ in the exponential constant. In the full-rank setting, the exponential factor is e^{K_T}, while our low-rank architecture introduces an additional $(1 + rK)$ factor that accounts for the r independent channels evolving through the mixing matrix L. This factor reflects the channel feature learning structure: the r channels evolve independently through the mixing matrix L, each contributing a factor that reflects the multi-channel nature of the learning dynamics.

Corollary G.1 (Test function approximation quality)**.** *Under the same setting as Theorem* **??**, *consider any test function* $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *which is* K*-Lipschitz in the second variable uniformly in the first variable (an example of* ψ *is the loss* $\mathscr{L}$*). For any* $\delta > 0$, *with probability at least* $1 - 3\delta$,

$$\sup_{t\le T} \left|\mathbb{E}_Z\left[\psi\left(Y, \hat{\mathbf{y}}\left(X; \mathbf{W}\left(\lfloor t/\epsilon \rfloor\right)\right)\right)\right] - \mathbb{E}_Z\left[\psi\left(Y, \hat{y}\left(X; W\left(t\right)\right)\right)\right]\right| \le e^{2K_T(1+rK)} \left(\frac{1}{\sqrt{n_{\min}}} + \sqrt{\epsilon}\right) \log^{1/2}\left(\frac{3(T+1)n_{\max}^2}{\delta} + e\right),$$

where $\hat{\mathbf{y}}(X; \mathbf{W})$ *and* $\hat{y}(X; W)$ *denote the outputs of the finite-width and mean-field networks respectively.*

Proof sketch. The proof follows the same structure as in the full-rank case [**?**]. Since ψ is K-Lipschitz in the second variable, we have:

$$|\psi(Y, \hat{\mathbf{y}}(X; \mathbf{W})) - \psi(Y, \hat{y}(X; W))| \le K|\hat{\mathbf{y}}(X; \mathbf{W}) - \hat{y}(X; W)|.$$

The difference $|\hat{\mathbf{y}}(X; \mathbf{W}) - \hat{y}(X; W)|$ can be bounded by the distance $\mathscr{D}_T(W, \mathbf{W})$ using the low-rank structure of the network. Specifically, for the low-rank architecture, the output difference involves:

$$\begin{aligned}
|\hat{\mathbf{y}}(X; \mathbf{W}) - \hat{y}(X; W)| \le \sum_{k=1}^{r} |L_{C_2,k}| \Bigg| \frac{1}{n_1} \sum_{j_1=1}^{n_1} \mathbf{w}_1(\lfloor t/\epsilon \rfloor, j_1, k) \varphi_1(\langle L^0(C_1(j_1)), X\rangle) \\
- \mathbb{E}_{C_1}[w_1(t, C_1, k)\varphi_1(\langle L^0(C_1), X\rangle)] \Bigg| + (\text{similar terms for } w_2) \\
\le rK\mathscr{D}_T(W, \mathbf{W}) + K\mathscr{D}_T(W, \mathbf{W}) = K(1 + rK)\mathscr{D}_T(W, \mathbf{W}),
\end{aligned}$$

where we used $\|L\|_{\infty,1} \le rK$. Taking expectation over Z and applying Theorem **??** completes the proof. □

Comparison with the full-rank Case. The main difference between our low-rank result and the full-rank case [**?**] is the appearance of the factor $(1 + rK)$ in the exponential constant. In the full-rank case, the bound is:

$$\mathscr{D}_T(W, \mathbf{W}) \leq e^{K_T} \left(\frac{1}{\sqrt{n_{\min}}} + \sqrt{\epsilon} \right) \log^{1/2} \left(\frac{3(T+1)n_{\max}^2}{\delta} + e \right),$$

whereas in our low-rank case, we have:

$$\mathscr{D}_T(W, \mathbf{W}) \leq e^{K_T(1+rK)} \left(\frac{1}{\sqrt{n_{\min}}} + \sqrt{\epsilon} \right) \log^{1/2} \left(\frac{3(T+1)n_{\max}^2}{\delta} + e \right).$$

The factor $(1 + rK)$ arises from:

- The low-rank structure of H_2, which involves a sum over r channels: $H_2(t, c_2; X, W) = \sum_{k=1}^{r} L_{c_2,k}\, f_k(t; X, W)$.
- The mixing matrix bounds: $\|L\|_{\infty,1} = \sup_{c_2} \sum_{k=1}^{r} |L_{c_2,k}| \leq rK$.
- The backpropagated signal B_k which aggregates over n_2 neurons with mixing coefficients $L_{C_2,k}$.

However, the fundamental structure of the proof remains the same: we still decompose the error into particle coupling and gradient descent discretization, and the scaling with $n_{\min}$ and ϵ is identical. The low-rank structure only affects the exponential constant, not the polynomial scaling. This suggests that the mean-field approximation quality is preserved under low-rank constraints, with the trade-off being a potentially larger (but still finite) exponential constant.

Remark G.1 (Deterministic mixing matrix and hierarchical learning). The mixing matrix L can be chosen deterministically (e.g., on a grid) rather than randomly, since the proof only requires the boundedness condition $\|L\|_{\infty,1} \leq rK$. There is no advantage to maximizing the entries of L; fixing L deterministically with appropriate structure can enable hierarchical learning, where different channels k specialize to different frequency components or scales through the backpropagated signal B_k. This design choice allows for structured learning dynamics while maintaining the same theoretical guarantees.

G.2 Particle ODEs for Low-Rank Networks

We construct auxiliary trajectories, which we call the *particle ODEs* for the low-rank case. These are continuous-time trajectories of finitely many neurons, averaged over the data distribution, adapted to the low-rank structure:

$$\begin{aligned}
\frac{\partial}{\partial t}\tilde{w}_2(t, j_2) &= -\xi_2(t)\mathbb{E}_Z \left[d_L(Z; \tilde{W}(t))\, \varphi_2(\tilde{H}_2(t, j_2; X, \tilde{W}(t))) \right], \\
\frac{\partial}{\partial t}\tilde{w}_1(t, j_1, k) &= -\xi_1(t)\mathbb{E}_Z \left[d_L(Z; \tilde{W}(t))\, \varphi_1(\langle L^0(C_1(j_1)), X \rangle)\, \tilde{B}_k(t; X, \tilde{W}(t)) \right],
\end{aligned}$$

where $j_1 = 1, \ldots, n_1$, $j_2 = 1, \ldots, n_2$, $k = 1, \ldots, r$, $\tilde{W}(t) = (\tilde{w}_1(t, \cdot, \cdot), \tilde{w}_2(t, \cdot))$, and $t \in \mathbb{R}_{\geq 0}$.

The second-layer output for the particle ODEs is:

$$\tilde{H}_2(t, j_2; X, \tilde{W}) = \sum_{k=1}^{r} L_{C_2(j_2),k}\, \frac{1}{n_1} \sum_{j_1=1}^{n_1} \tilde{w}_1(t, j_1, k)\, \varphi_1(\langle L^0(C_1(j_1)), X \rangle),$$

and the backpropagated signal is:

$$\tilde{B}_k(t; X, \tilde{W}) = \frac{1}{n_2} \sum_{j_2=1}^{n_2} L_{C_2(j_2),k}\, \varphi_2'(\tilde{H}_2(t, j_2; X, \tilde{W}))\, \tilde{w}_2(t, j_2).$$

We specify the initialization $\tilde{W}(0)$: $\tilde{w}_1(0, j_1, k) = w_1^0(C_1(j_1), k)$ and $\tilde{w}_2(0, j_2) = w_2^0(C_2(j_2))$. That is, it shares the same initialization with the neural network $\mathbf{W}(0)$, and hence is coupled with the neural network and the MF ODEs.

The existence and uniqueness of the solution to the particle ODEs follows from the same proof as in Theorem **??**, adapted to account for the low-rank structure. We equip $\tilde{W}(t)$ with the norm:

$$\tilde{W}_T = \max \left\{ \max_{j_1 \leq n_1, 1 \leq k \leq r} \sup_{t \leq T} |\tilde{w}_1(t, j_1, k)|, \max_{j_2 \leq n_2} \sup_{t \leq T} |\tilde{w}_2(t, j_2)| \right\}.$$

We define the distance measures:

$$\mathscr{D}_T(W,\tilde{W}) = \sup\Big\{|w_1(t,C_1(j_1),k) - \tilde{w}_1(t,C_1(j_1),k)|, |w_2(t,C_2(j_2)) - \tilde{w}_2(t,C_2(j_2))| : \\ t \le T, j_1 \le n_1, j_2 \le n_2, 1 \le k \le r\Big\},$$

$$\mathscr{D}_T(\tilde{W},\mathbf{W}) = \sup\Big\{|\mathbf{w}_1(\lfloor t/\epsilon \rfloor, j_1, k) - \tilde{w}_1(t,C_1(j_1),k)|, \\ |\mathbf{w}_2(\lfloor t/\epsilon \rfloor, j_2) - \tilde{w}_2(t,C_2(j_2))| : t \le T, j_1 \le n_1, j_2 \le n_2, 1 \le k \le r\Big\}.$$

G.3 Key Lemmas

Lemma G.1 (Particle coupling bound)**.** *Under the same setting as Theorem* **??***, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\mathscr{D}_T(W,\tilde{W}) \le \frac{1}{\sqrt{n_{\min}}} \log^{1/2}\left(\frac{3(T+1)n_{\max}^2}{\delta} + e\right) e^{K_T(1+rK)},$$

in which $n_{\min} = \min\{n_1,n_2\}$, $n_{\max} = \max\{n_1,n_2\}$, and $K_T = K(1+T^K)$.

Lemma G.2 (Gradient descent discretization error)**.** *Under the same setting as Theorem* **??***, for any $\delta > 0$ and $\epsilon \le 1$, with probability at least $1-\delta$,*

$$\mathscr{D}_T(\tilde{W},\mathbf{W}) \le \sqrt{\epsilon \log\left(\frac{2n_1n_2r}{\delta} + e\right)} e^{K_T(1+rK)},$$

in which $K_T = K(1+T^K)$.

Proof of Theorem **??**. Using the triangle inequality:

$$\mathscr{D}_T(W,\mathbf{W}) \le \mathscr{D}_T(W,\tilde{W}) + \mathscr{D}_T(\tilde{W},\mathbf{W}),$$

the result follows immediately from Lemmas **??** and **??**, noting that the $\log(2n_1n_2r/\delta)$ term can be absorbed into the $\log(3(T+1)n_{\max}^2/\delta + e)$ term up to constants. □

G.4 Proof of Lemma ??

Proof. In the following, let K_t denote a generic positive constant that may change from line to line and takes the form $K_t = K(1+t^K)$, such that $K_t \ge 1$ and $K_t \le K_T$ for all $t \le T$. We first note that at initialization, $\mathscr{D}_0(W,\tilde{W}) = 0$. Since $W_0 \le K$, we have $W_T \le K_T$ by the a priori bounds. Furthermore, it is easy to see that $\tilde{W}_0 \le W_0 \le K$ almost surely. By the same argument, $\tilde{W}_T \le K_T$ almost surely.

We decompose the proof into several steps.

Step 1: Main bound with low-rank structure. Let us define, for brevity, the differences specific to the low-rank architecture:

$$q_2(t,x,j_2,c_2) = \tilde{H}_2(t,j_2;x,\tilde{W}(t)) - H_2(t,c_2;x,W(t)),$$
$$q_{B,k}(t,x) = \tilde{B}_k(t;x,\tilde{W}(t)) - B_k(t;x,W(t)),$$

where $H_2(t,c_2;x,W) = \sum_{k=1}^r L_{c_2,k}\,\mathbb{E}_{C_1}[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)]$ and $B_k(t;x,W) = \mathbb{E}_{C_2}[L_{C_2,k}\,\varphi_2'(H_2(t,C_2;x,W))\,w_2(t,C_2)]$.

Consider $t \ge 0$. We first bound the difference in the updates between W and $\tilde{W}$.

Bound for w_2 and $\tilde{w}_2$: By Assumption **??** and the definition of the ODEs:

$$\begin{aligned}
&\left|\frac{\partial}{\partial t}\tilde{w}_2(t,j_2) - \frac{\partial}{\partial t}w_2(t,C_2(j_2))\right| \\
&\quad= \left|\xi_2(t)\mathbb{E}_Z\left[d_L(Z;\tilde{W}(t))\,\varphi_2(\tilde{H}_2(t,j_2;X,\tilde{W}(t))) - d_L(Z;W(t))\,\varphi_2(H_2(t,C_2(j_2);X,W(t)))\right]\right| \\
&\quad\le K\mathbb{E}_Z\left[|d_L(Z;\tilde{W}(t)) - d_L(Z;W(t))|\,|\varphi_2(\tilde{H}_2(t,j_2;X,\tilde{W}(t)))|\right] \\
&\qquad + K\mathbb{E}_Z\left[|d_L(Z;W(t))|\,|\varphi_2(\tilde{H}_2(t,j_2;X,\tilde{W}(t))) - \varphi_2(H_2(t,C_2(j_2);X,W(t)))|\right] \\
&\quad\le K_t\mathbb{E}_Z\left[|q_2(t,X,j_2,C_2(j_2))|\right] + K_t\mathscr{D}_t(W,\tilde{W}),
\end{aligned}$$

where we used that d_L is Lipschitz in W, φ_2 is Lipschitz, and H_2 differences are controlled.

Bound for w_1 and $\tilde{w}_1$: For the low-rank case, we have $k = 1,\dots,r$ channels. By the definition of the ODEs:

$$\begin{aligned}
&\left|\frac{\partial}{\partial t}\tilde{w}_1(t,j_1,k) - \frac{\partial}{\partial t}w_1(t,C_1(j_1),k)\right| \\
&\quad= \left|\xi_1(t)\mathbb{E}_Z\left[d_L(Z;\tilde{W}(t))\,\varphi_1(\langle L^0(C_1(j_1)),X\rangle)\,\tilde{B}_k(t;X,\tilde{W}(t))\right.\right. \\
&\qquad\left.\left. -d_L(Z;W(t))\,\varphi_1(\langle L^0(C_1(j_1)),X\rangle)\,B_k(t;X,W(t))\right]\right| \\
&\quad\le K\mathbb{E}_Z\left[|d_L(Z;\tilde{W}(t)) - d_L(Z;W(t))|\,|\tilde{B}_k(t;X,\tilde{W}(t))|\right] \\
&\qquad + K\mathbb{E}_Z\left[|d_L(Z;W(t))|\,|q_{B,k}(t,X)|\right] \\
&\quad\le K_t\mathbb{E}_Z\left[|q_{B,k}(t,X)|\right] + K_t\mathscr{D}_t(W,\tilde{W}),
\end{aligned}$$

where the expectation over j_2 in $\tilde{B}_k$ will be handled via concentration.

Step 2: Decomposition of q_2 with low-rank structure. The key difference from the full-rank case is that H_2 involves a sum over r channels. We decompose:

$$\begin{aligned}
|q_2(t,x,j_2,c_2)| &= \left|\sum_{k=1}^{r} L_{C_2(j_2),k}\left(\frac{1}{n_1}\sum_{j_1=1}^{n_1}\tilde{w}_1(t,j_1,k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle)\right.\right. \\
&\qquad\left.\left. -\mathbb{E}_{C_1}[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)]\right)\right| \\
&\le \sum_{k=1}^{r}|L_{C_2(j_2),k}|\left|\frac{1}{n_1}\sum_{j_1=1}^{n_1}\tilde{w}_1(t,j_1,k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle)\right. \\
&\qquad\left. -\mathbb{E}_{C_1}[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)]\right|.
\end{aligned}$$

We further decompose each term in the sum:

$$\begin{aligned}
&\left|\frac{1}{n_1}\sum_{j_1=1}^{n_1}\tilde{w}_1(t,j_1,k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle) - \mathbb{E}_{C_1}[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)]\right| \\
&\quad\le \max_{j_1\le n_1}\left|\tilde{w}_1(t,j_1,k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle) - w_1(t,C_1(j_1),k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle)\right| \\
&\qquad + \left|\frac{1}{n_1}\sum_{j_1=1}^{n_1}w_1(t,C_1(j_1),k)\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle) - \mathbb{E}_{C_1}[w_1(t,C_1,k)\,\varphi_1(\langle L^0(C_1),x\rangle)]\right| \\
&\quad\equiv Q_{2,1,k}(x,j_2) + Q_{2,2,k}(x,j_2).
\end{aligned}$$

Therefore:

$$|q_2(t,x,j_2,c_2)| \le \sum_{k=1}^{r}|L_{C_2(j_2),k}|(Q_{2,1,k}(x,j_2)+Q_{2,2,k}(x,j_2)) \le rK\max_{1\le k\le r}(Q_{2,1,k}(x,j_2)+Q_{2,2,k}(x,j_2)),$$

where we used $\|L\|_{\infty,1}\le rK$.

Step 3: Decomposition of $q_{B,k}$ with low-rank structure. For the backpropagated signal difference:

$$|q_{B,k}(t,x)| = \left| \frac{1}{n_2} \sum_{j_2=1}^{n_2} L_{C_2(j_2),k}\, \varphi_2'(\tilde{H}_2(t,j_2;x,\tilde{W}))\, \tilde{w}_2(t,j_2) \right. \\ \left. -\mathbb{E}_{C_2}[L_{C_2,k}\, \varphi_2'(H_2(t,C_2;x,W))\, w_2(t,C_2)] \right|.$$

We decompose:

$$\begin{aligned} |q_{B,k}(t,x)| \le & \max_{j_2 \le n_2} \left| L_{C_2(j_2),k}\, \varphi_2'(\tilde{H}_2(t,j_2;x,\tilde{W}))\, \tilde{w}_2(t,j_2) \right. \\ & \left. - L_{C_2(j_2),k}\, \varphi_2'(H_2(t,C_2(j_2);x,W))\, w_2(t,C_2(j_2)) \right| \\ & + \left| \frac{1}{n_2} \sum_{j_2=1}^{n_2} L_{C_2(j_2),k}\, \varphi_2'(H_2(t,C_2(j_2);x,W))\, w_2(t,C_2(j_2)) \right. \\ & \left. - \mathbb{E}_{C_2}[L_{C_2,k}\, \varphi_2'(H_2(t,C_2;x,W))\, w_2(t,C_2)] \right| \\ \equiv & \, Q_{B,1,k}(x) + Q_{B,2,k}(x). \end{aligned}$$

For $Q_{B,1,k}$, using Assumption **??** and the fact that $|L_{C_2(j_2),k}| \le K$:

$$\begin{aligned} \mathbb{E}_Z[Q_{B,1,k}(X)] \le & K \max_{j_2 \le n_2} \left(|\tilde{w}_2(t,j_2) - w_2(t,C_2(j_2))| \right. \\ & \left. + |w_2(t,C_2(j_2))|\, \mathbb{E}_Z[|\tilde{H}_2(t,j_2;X,\tilde{W}) - H_2(t,C_2(j_2);X,W)|] \right) \\ \le & \, K_t \left(\mathscr{D}_t(W,\tilde{W}) + \max_{j_2 \le n_2} \mathbb{E}_Z[|q_2(t,X,j_2,C_2(j_2))|] \right). \end{aligned}$$

Step 4: Concentration bounds adapted to low-rank. For $Q_{2,1,k}$, we have:

$$\begin{aligned} \max_{j_2 \le n_2} \mathbb{E}_Z[Q_{2,1,k}(X,j_2)] \le & K \max_{j_1 \le n_1, 1 \le k \le r} |\tilde{w}_1(t,j_1,k) - w_1(t,C_1(j_1),k)| \\ \le & \, K_t \mathscr{D}_t(W,\tilde{W}). \end{aligned}$$

For $Q_{2,2,k}$, we apply concentration. Let us write:

$$Z_{1,k}(x,c_1) = w_1(t,c_1,k)\, \varphi_1(\langle L^0(c_1), x \rangle).$$

Recall that $\{C_1(j_1)\}_{j_1=1}^{n_1}$ are i.i.d. We have:

$$\mathbb{E}[Z_{1,k}(X,C_1(j_1))|X] = \mathbb{E}_{C_1}[Z_{1,k}(X,C_1)],$$

and $\{Z_{1,k}(X,C_1(j_1))\}_{j_1=1}^{n_1}$ are independent conditional on X. By Assumption **??**, $|Z_{1,k}(X,C_1(j_1))| \le K_t$ almost surely. Then by concentration inequalities:

$$\mathbb{P}\left(\mathbb{E}_Z[Q_{2,2,k}(X,j_2)] \ge K_t \gamma_{2,k} \right) \le \frac{1}{\gamma_{2,k}} \exp\left(-\frac{n_1 \gamma_{2,k}^2}{K_t} \right).$$

For $Q_{B,2,k}$, we note that almost surely:

$$|L_{C_2(j_2),k}\, \varphi_2'(H_2(t,C_2(j_2);x,W))\, w_2(t,C_2(j_2))| \le K |L_{C_2(j_2),k}| \le K^2,$$

by Assumption **??**. Then by concentration inequalities:

$$\mathbb{P}\left(\mathbb{E}_Z[Q_{B,2,k}(X)] \ge K_t \gamma_{1,k} \right) \le \frac{1}{\gamma_{1,k}} \exp\left(-\frac{n_2 \gamma_{1,k}^2}{K_t} \right).$$

Step 5: Combining bounds with union over k. Taking a union bound over $k = 1, \dots, r$ channels, we obtain that for any $\gamma_1, \gamma_2 > 0$ and $t \geq 0$, the event:

$$\begin{aligned}\max \Bigg\{ & \max_{j_2 \leq n_2} \mathbb{E}_Z[|q_2(t, X, j_2, C_2(j_2))|], \\ & \max_{j_1 \leq n_1, 1 \leq k \leq r} \mathbb{E}_Z[|q_{B,k}(t, X)|] \Bigg\} \\ & \leq K_t(1 + rK)(\mathscr{D}_t(W, \tilde{W}) + \gamma_1 + \gamma_2),\end{aligned}$$

holds with probability at least:

$$1 - \frac{n_1 r}{\gamma_1} \exp\left(-\frac{n_2 \gamma_1^2}{K_t(1+rK)^2}\right) - \frac{n_2 r}{\gamma_2} \exp\left(-\frac{n_1 \gamma_2^2}{K_t(1+rK)^2}\right).$$

Step 6: Gronwall argument. Combining the bounds and taking a union bound over a discrete time grid $t \in \{0, \xi, 2\xi, \dots, \lfloor T/\xi \rfloor \xi\}$ for some $\xi \in (0, 1)$, we obtain:

$$\begin{aligned}\max \Bigg\{ & \max_{j_2 \leq n_2} \left| \frac{\partial}{\partial t} \tilde{w}_2(t, j_2) - \frac{\partial}{\partial t} w_2(t, C_2(j_2)) \right|, \\ & \max_{j_1 \leq n_1, 1 \leq k \leq r} \left| \frac{\partial}{\partial t} \tilde{w}_1(t, j_1, k) - \frac{\partial}{\partial t} w_1(t, C_1(j_1), k) \right| \Bigg\} \\ & \leq K_T(1 + rK)(\mathscr{D}_t(W, \tilde{W}) + \gamma_1 + \gamma_2 + \xi), \qquad \forall t \in [0, T],\end{aligned}$$

with probability at least:

$$1 - \frac{T+1}{\xi} \left[\frac{n_1 r}{\gamma_1} \exp\left(-\frac{n_2 \gamma_1^2}{K_T(1+rK)^2}\right) + \frac{n_2 r}{\gamma_2} \exp\left(-\frac{n_1 \gamma_2^2}{K_T(1+rK)^2}\right) \right].$$

The above event implies:

$$\mathscr{D}_t(W, \tilde{W}) \leq K_T(1 + rK) \int_0^t (\mathscr{D}_s(W, \tilde{W}) + \gamma_1 + \gamma_2 + \xi) ds,$$

and hence by Gronwall's lemma and the fact $\mathscr{D}_0(W, \tilde{W}) = 0$:

$$\mathscr{D}_T(W, \tilde{W}) \leq (\gamma_1 + \gamma_2 + \xi) e^{K_T(1+rK)}.$$

The result follows from choosing:

$$\begin{aligned}\xi &= \frac{1}{\sqrt{n_{\max}}}, \\ \gamma_1 &= \frac{K_T(1+rK)}{\sqrt{n_2}} \log^{1/2}\left(\frac{3(T+1)n_{\max}^2 r}{\delta} + e\right), \\ \gamma_2 &= \frac{K_T(1+rK)}{\sqrt{n_1}} \log^{1/2}\left(\frac{3(T+1)n_{\max}^2 r}{\delta} + e\right).\end{aligned}$$

□

G.5 Proof of Lemma ??

Proof. The proof follows the same structure as the full-rank case [**?**], with careful adaptations for the low-rank structure. The key difference is that we need to account for the r channels in w_1 and the mixing matrix L in all bounds.

We consider $t \leq T$ for a given terminal time $T \in \epsilon \mathbb{N}_{\geq 0}$. We reuse the notation K_t from the proof of Lemma **??**. Note that $K_t \leq K_T$ for all $t \leq T$. We also note that at initialization, $\mathscr{D}_0(\mathbf{W}, \tilde{W}) = 0$.

For brevity, let us define quantities that relate to the difference in the gradient updates between $\mathbf{W}$ and $\tilde{W}$:

$$\begin{aligned}
q_2(k,z,\tilde{z},j_2) &= d_L(z;\mathbf{W}(k))\,\varphi_2(\mathbf{H}_2(k,j_2;x,\mathbf{W}(k))) \\
&\quad - d_L(\tilde{z};\tilde{W}(k\epsilon))\,\varphi_2(\tilde{H}_2(k\epsilon,j_2;\tilde{x},\tilde{W}(k\epsilon))), \\
r_2(k,z,j_2) &= \xi_2(k\epsilon) d_L(z;\tilde{W}(k\epsilon))\,\varphi_2(\tilde{H}_2(k\epsilon,j_2;x,\tilde{W}(k\epsilon))) \\
&\quad - \xi_2(k\epsilon)\mathbb{E}_Z[d_L(Z;\tilde{W}(k\epsilon))\,\varphi_2(\tilde{H}_2(k\epsilon,j_2;X,\tilde{W}(k\epsilon)))], \\
q_1(k,z,\tilde{z},j_1,k') &= d_L(z;\mathbf{W}(k))\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle)\,\mathbf{B}_{k'}(k;x,\mathbf{W}(k)) \\
&\quad - d_L(\tilde{z};\tilde{W}(k\epsilon))\,\varphi_1(\langle L^0(C_1(j_1)),\tilde{x}\rangle)\,\tilde{B}_{k'}(k\epsilon;\tilde{x},\tilde{W}(k\epsilon)), \\
r_1(k,z,j_1,k') &= \xi_1(k\epsilon) d_L(z;\tilde{W}(k\epsilon))\,\varphi_1(\langle L^0(C_1(j_1)),x\rangle)\,\tilde{B}_{k'}(k\epsilon;x,\tilde{W}(k\epsilon)) \\
&\quad - \xi_1(k\epsilon)\mathbb{E}_Z[d_L(Z;\tilde{W}(k\epsilon))\,\varphi_1(\langle L^0(C_1(j_1)),X\rangle)\,\tilde{B}_{k'}(k\epsilon;X,\tilde{W}(k\epsilon))],
\end{aligned}$$

where $\mathbf{B}_{k'}(k;x,\mathbf{W}) = \frac{1}{n_2}\sum_{j_2=1}^{n_2} L_{C_2(j_2),k'}\,\varphi_2'(\mathbf{H}_2(k,j_2;x,\mathbf{W}))\,\mathbf{w}_2(k,j_2)$.

By time-interpolation estimates (similar to Claim 1 in the full-rank case) and Assumption **??**, we have:

$$\begin{aligned}
|\mathbf{w}_2(\lfloor t/\epsilon\rfloor,j_2) - \tilde{w}_2(t,j_2)| &\le K \max_{j_2\le n_2}\left[Q_{2,1}(\lfloor t/\epsilon\rfloor,j_2) + Q_{2,2}(\lfloor t/\epsilon\rfloor,j_2)\right] + tK_t\epsilon, \\
|\mathbf{w}_1(\lfloor t/\epsilon\rfloor,j_1,k) - \tilde{w}_1(t,j_1,k)| &\le K \max_{j_1\le n_1,1\le k\le r}\left[Q_{1,1}(\lfloor t/\epsilon\rfloor,j_1,k) + Q_{1,2}(\lfloor t/\epsilon\rfloor,j_1,k)\right] + tK_t\epsilon,
\end{aligned}$$

where:

$$\begin{aligned}
Q_{2,1}(\lfloor t/\epsilon\rfloor,j_2) &= \epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1}|q_2(\ell,z(\ell),z(\ell),j_2)|, \\
Q_{2,2}(\lfloor t/\epsilon\rfloor,j_2) &= \left|\epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1} r_2(\ell,z(\ell),j_2)\right|, \\
Q_{1,1}(\lfloor t/\epsilon\rfloor,j_1,k) &= \epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1}|q_1(\ell,z(\ell),z(\ell),j_1,k)|, \\
Q_{1,2}(\lfloor t/\epsilon\rfloor,j_1,k) &= \left|\epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1} r_1(\ell,z(\ell),j_1,k)\right|.
\end{aligned}$$

Bounding the terms: For $Q_{2,1}$, using Assumption **??** and the low-rank structure:

$$\begin{aligned}
|q_2(\ell,z(\ell),z(\ell),j_2)| &\le K|d_L(z(\ell);\mathbf{W}(\ell)) - d_L(z(\ell);\tilde{W}(\ell\epsilon))| \\
&\quad + K|d_L(z(\ell);\tilde{W}(\ell\epsilon))|\,|\mathbf{H}_2(\ell,j_2;x(\ell),\mathbf{W}(\ell)) - \tilde{H}_2(\ell\epsilon,j_2;x(\ell),\tilde{W}(\ell\epsilon))| \\
&\le K_t\mathscr{D}_{\ell\epsilon}(\tilde{W},\mathbf{W}) + K_t\sum_{k=1}^{r}|L_{C_2(j_2),k}|\max_{j_1\le n_1}|\mathbf{w}_1(\ell,j_1,k) - \tilde{w}_1(\ell\epsilon,j_1,k)| \\
&\le K_t(1+rK)\mathscr{D}_{\ell\epsilon}(\tilde{W},\mathbf{W}),
\end{aligned}$$

which yields:

$$\max_{j_2\le n_2} Q_{2,1}(\lfloor t/\epsilon\rfloor,j_2) \le K_t(1+rK)\epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1}\mathscr{D}_{\ell\epsilon}(\tilde{W},\mathbf{W}).$$

For $Q_{1,1}$, similarly:

$$\begin{aligned}
|q_1(\ell,z(\ell),z(\ell),j_1,k)| &\le K|d_L(z(\ell);\mathbf{W}(\ell)) - d_L(z(\ell);\tilde{W}(\ell\epsilon))|\,|\mathbf{B}_k(\ell;x(\ell),\mathbf{W}(\ell))| \\
&\quad + K|d_L(z(\ell);\tilde{W}(\ell\epsilon))|\,|\mathbf{B}_k(\ell;x(\ell),\mathbf{W}(\ell)) - \tilde{B}_k(\ell\epsilon;x(\ell),\tilde{W}(\ell\epsilon))| \\
&\le K_t(1+rK)\mathscr{D}_{\ell\epsilon}(\tilde{W},\mathbf{W}),
\end{aligned}$$

which yields:

$$\max_{j_1\le n_1,1\le k\le r} Q_{1,1}(\lfloor t/\epsilon\rfloor,j_1,k) \le K_t(1+rK)\epsilon\sum_{\ell=0}^{\lfloor t/\epsilon\rfloor-1}\mathscr{D}_{\ell\epsilon}(\tilde{W},\mathbf{W}).$$

For $Q_{2,2}$ and $Q_{1,2}$, we use martingale concentration. The martingale differences are bounded: $|r_2(\ell, z(\ell), j_2)| \le K_t$ and $|r_1(\ell, z(\ell), j_1, k)| \le K_t(1 + rK)$ almost surely by Assumption **??** and the low-rank structure. Then by martingale concentration inequalities:

$$\mathbb{P}\left(\max_{j_2 \le n_2} \max_{\ell \in \{0,1,\ldots,T/\epsilon\}} Q_{2,2}(\ell, j_2) \ge \xi\right) \le 2n_2 \exp\left(-\frac{\xi^2}{K_T(T+1)\epsilon}\right);$$

$$\mathbb{P}\left(\max_{j_1 \le n_1, 1 \le k \le r} \max_{\ell \in \{0,1,\ldots,T/\epsilon\}} Q_{1,2}(\ell, j_1, k) \ge \xi\right) \le 2n_1 r \exp\left(-\frac{\xi^2}{K_T(1+rK)^2(T+1)\epsilon}\right).$$

Putting everything together: All the above results give us:

$$\mathscr{D}_{\lfloor t/\epsilon \rfloor \epsilon}(\tilde{W}, \mathbf{W}) \le K_T(1 + rK)\epsilon \sum_{\ell=0}^{\lfloor t/\epsilon \rfloor - 1} \mathscr{D}_{\ell\epsilon}(\tilde{W}, \mathbf{W}) + \xi + TK_T\epsilon, \qquad \forall t \le T,$$

which holds with probability at least:

$$1 - 2n_1 r \exp\left(-\frac{\xi^2}{K_T(1+rK)^2(T+1)\epsilon}\right) - 2n_2 \exp\left(-\frac{\xi^2}{K_T(T+1)\epsilon}\right).$$

The above event implies, by Gronwall's lemma:

$$\mathscr{D}_T(\tilde{W}, \mathbf{W}) \le (\xi + \epsilon) e^{K_T(1+rK)}.$$

Choosing $\xi = K_T(1 + rK)\sqrt{(T+1)\epsilon \log(2n_1 n_2 r/\delta)}$ completes the proof. □

G.6 Discussion: Width Requirement and Exponential Factor

The bound in Theorem **??** shows that the approximation error scales as:

$$\mathscr{D}_T(W, \mathbf{W}) \le e^{K_T(1+rK)} \left(\frac{1}{\sqrt{n_{\min}}} + \sqrt{\epsilon}\right) \log^{1/2}\left(\frac{3(T+1)n_{\max}^2}{\delta} + e\right).$$

For typical values $T \le 10$, $K \approx 1$, $r \le 100$, and $\epsilon \approx 0.001$, we have $K_T(1+rK) \le 1000$ (roughly), so $e^{K_T(1+rK)} \le e^{1000}$, which is an extremely large exponential factor. This reflects the worst-case scenario where all r channels must be learned simultaneously. However, in practice, channel specialization enables a more favorable learning regime: channels progressively capture different frequency components, avoiding the worst-case exponential scaling. The actual convergence rate is determined by the favorable loss landscape structure (see Section **??**) rather than this worst-case bound.

The key qualitative insight is that the error *decreases* as $n_{\min}$ increases, with rate $O(1/\sqrt{n_{\min}})$ independent of data dimension d. The exponential factor $e^{K_T(1+rK)}$ reflects the multi-channel structure but does not dominate in practice due to channel specialization.

Empirical validation in [**?**] confirms that networks with width ≈ 1000 trained on high-dimensional real-world data (e.g., $d \approx 1000$) exhibit mean-field behaviors, supporting our theoretical prediction that the required width is independent of d.

H Channel-wise partial functions

This section rewrites the r channel summaries $m_k(t; \cdot)$ as integrals against a measure μ_0 on the *untrained* first-layer features, together with a pushforward that incorporates the *trained* first-layer weights w_1. Let $w_0 : \Omega_1 \to^d$ denote the untrained first-layer map (in the main text, $w_0 = L^0$) and let ρ^1 be the law of C_1. Define the pushforward $\mu_0 \equiv (w_0)_\# \rho^1$ on ^{d}. For each $t \ge 0$ and $k \in \{1, \ldots, r\}$, the signed measure $\mu_{1,k}^t \equiv (w_0)_\#(w_1(t, \cdot, k)\, \rho^1)$ satisfies $\mu_{1,k}^t(d\theta) = \bar{w}_{1,k}(t, \theta)\, \mu_0(d\theta)$ with $\bar{w}_{1,k}(t, \theta)$ the conditional average of $w_1(t, \cdot, k)$ given $w_0(c_1) = \theta$. Then $m_k(t; X, W) = \int_d \varphi_1(\theta X)\, \mu_{1,k}^t(d\theta) = \int_d \bar{w}_{1,k}(t, \theta)\, \varphi_1(\theta X)\, \mu_0(d\theta)$.

H.1 Time variation of the r partial functions and the induced PDE

Define $f_k(t,x) \equiv m_k(t;x,W) = \int_d \bar{w}_{1,k}(t,\theta)\,\varphi_1(\theta x)\,\mu_0(d\theta)$. From the MF ODE (**??**), with $B_k(t;X) \equiv_{C_2} [L_{C_2,k}\,\varphi_2'(H_2(t,C_2;X,W(t)))\,w_2(t,C_2)]$, the coefficient field evolves as

$$\partial_t \bar{w}_{1,k}(t,\theta) = -\xi_1(t)\,_{Z=(X,Y)}\big[d_L\big(Z;W(t)\big)\,\varphi_1(\theta X)\,B_k(t;X)\big]. \tag{9}$$

Differentiating under the integral and defining $K_{\mu_0}(x,x') \equiv \int_d \varphi_1(\theta x)\,\varphi_1(\theta x')\,\mu_0(d\theta)$ yields

$$\partial_t f_k(t,x) = -\xi_1(t)\,_{Z=(X,Y)}\big[d_L\big(Z;W(t)\big)\,B_k(t;X)\,K_{\mu_0}(x,X)\big], \qquad k=1,\dots,r. \tag{10}$$

For ReLU $\varphi_2(u)=u_+$, $B_k(t;X) =_{C_2} [L_{C_2,k}\,w_2(t,C_2)\,\{H_2(t,C_2;X,W(t))>0\}]$. In a one-spike reduction at $x_\star$ with $f_k(t)=f_k(t,x_\star)$, $d(t)=d_L((x_\star,y_\star);W(t))$, the dynamics are

$$\partial_t f_k(t) = -\xi_1(t)\,d(t)\,B_k(t), \tag{11}$$

$$\partial_t w_2(t,c_2) = -\xi_2(t)\,d(t)\,\big(L_{c_2} f(t)\big)_+, \tag{12}$$

with

$$B_k(t) =_{C_2} \big[L_{C_2,k}\,w_2(t,C_2)\,\big\{L_{C_2} f(t)>0\big\}\big], \qquad k=1,\dots,r. \tag{13}$$

Under Assumption **??** (symmetric independent coordinates of L_{C_2}), nonnegative $w_2(0,\cdot)$ and $d(t)\le 0$, a one-sparse initial $f(0)=a_0 e_j$ stays one-sparse: $f(t)=a(t)e_j$ with $a(t)\ge a_0$, $B_k(t)=0$ for $k\neq j$, and $B_j(t)>0$ (Lemma **??**). Half-space symmetry gives $[U\,\{aU>0\}] = \mathrm{sign}(a)\,\frac{1}{2}[|U|]$ (Lemma **??**).

Assumption H.1 (Random mixing vector, symmetric and independent)**.** The random vector $L_{C_2}=(L_{C_2,1},\dots,L_{C_2,r})$ has independent coordinates, each symmetric about 0: $L_{C_2,k} \overset{d}{=} -L_{C_2,k}$ and $[L_{C_2,k}^2]<\infty$ for all k.

Lemma H.1 (Half-space symmetry identities)**.** *Let U be a real random variable with $U \overset{d}{=} -U$ and $[|U|]<\infty$. Then for any $a\in\backslash\{0\}$, $[U\,\{aU>0\}] = \mathrm{sign}(a)\,\frac{1}{2}[|U|]$. If $[U^2]<\infty$, then $[U^2\,\{aU>0\}] = \frac{1}{2}[U^2]$.*

Lemma H.2 (One-sparse invariant manifold and emergent sign-coherence)**.** *Assume the one-spike system above, Assumption **??**, $d(t)\le 0$ on $[0,T]$, $w_2(0,c_2)\equiv w_2^0\ge 0$, and $f(0)=a_0e_j$ with $a_0>0$. Then for all $t\in[0,T]$: $f(t)=a(t)e_j$ with $a(t)\ge a_0$; $B_k(t)=0$ for $k\neq j$; $B_j(t)>0$ with $B_j(t)=\frac{w_2^0}{2}[|L_{C_2,j}|]+\frac{[L_{C_2,j}^2]}{2}\int_0^t \xi_2(s)\,(-d(s))\,a(s)\,ds$; and $\partial_t a(t)\ge 0$.*

H.2 Finite-support reduction and scalar ODEs

The kernel K_{μ_0} and the evolution (**??**) are derived in the preceding section. For the two-point support $\{x_0,x_1\}$, $K_{\mu_0}(x_0,x_1)$ is *positive* and *fastly decaying* in δ: $0<K_{\mu_0}(x_0,x_1)\le\psi(\delta)$ with $\psi(\delta)\to 0$ rapidly as $\delta\to\infty$, which holds for such NNGP kernels in 1D or under suitable geometry. The positivity reinforces the leading local term; the fast decay keeps the non-local remainder small.

Finite-support data. Assume the input marginal is supported on $x^{(1)},\dots,x^{(m)}$ with $(X=x^{(p)})=\pi_p$ and targets $y^{(p)}$. Set $d_p(t)=d_L((x^{(p)},y^{(p)});W(t))$ and $B_{k,p}(t)=B_k(t;x^{(p)})$. Then (**??**) becomes

$$\partial_t \bar{w}_{1,k}(t,\theta) = -\xi_1(t)\sum_{p=1}^m \pi_p\,d_p(t)\,B_{k,p}(t)\,\varphi_1(\theta x^{(p)}), \qquad k=1,\dots,r. \tag{14}$$

Integrating in time yields $\bar{w}_{1,k}(t,\theta)=\bar{w}_{1,k}(0,\theta)-\sum_{p=1}^m \varphi_1(\theta x^{(p)})\,\Gamma_{k,p}(t)$ with

$$\Gamma_{k,p}(t) \equiv \int_0^t \xi_1(s)\,\pi_p\,d_p(s)\,B_{k,p}(s)\,ds. \tag{15}$$

Plugging into f_k gives the explicit superposition

$$f_k(t,x) = f_k(0,x) - \sum_{p=1}^m \Gamma_{k,p}(t)\,K_{\mu_0}(x,x^{(p)}), \qquad k=1,\dots,r. \tag{16}$$

The spike shape is determined by K_{μ_0}; all learning dynamics reduce to the scalar coefficients $\Gamma_{k,p}(t)$.

One-spike reduction. If $m = 1$ with $x^{(1)} = x_\star$, then $f_k(t,x) = f_k(0,x) - \Gamma_{k,1}(t)\, K_{\mu_0}(x, x_\star)$. For $f_k(0,\cdot) \equiv 0$, $f_k(t,\cdot)$ is exactly a kernel bump at $x_\star$ and spike learning reduces to solving $\Gamma_{k,1}(t)$.

Two-sided step ($m = 2$). Take $x^{(1)} = x_0 = -\delta$, $x^{(2)} = x_1 = +\delta$ ($\delta > 0$), $y^{(1)} = +A$, $y^{(2)} = -A$. Then

$$f_k(t,x) = f_k(0,x) - \Gamma_{k,1}(t)\, K_{\mu_0}(x, x_0) - \Gamma_{k,2}(t)\, K_{\mu_0}(x, x_1).$$

The kernel K_{μ_0} (NNGP, [**?**]) satisfies $K_{\mu_0}(x_p, x_p) = K_0 > 0$ for $p \in \{0,1\}$, and the off-diagonal $K_{\mu_0}(x_0, x_1) = K_{\mu_0}(-\delta, +\delta)$ is positive and fastly decaying in δ: $0 < K_{\mu_0}(x_0, x_1) \le \psi(\delta)$ with $\psi(\delta) \to 0$ rapidly as $\delta \to \infty$. The full evolution is

$$\partial_t f_k(t, x_p) = -\xi_1(t)\, K_{\mu_0}(x_p, x_p)\, d_p(t)\, B_{k,p+1}(t) + E_p(t),$$

with $|E_p(t)| \le C'\, \psi(\delta)$ for $\psi(\delta)$ fastly decaying in δ. The positivity of $K_{\mu_0}(x_0, x_1)$ gives even better positivity in the log-ratio; the fast decay keeps the remainder small. This is the setting of Theorem **??**.

H.3 Proof of Theorem ?? (two-sided step, x_0)

We prove the theorem at x_0; the argument at x_1 is symmetrical.

At x_0. By the log-ratio derivative identity at x_0,

$$\partial_t R_{12}(t, x_0) = \frac{\operatorname{sign}(f_1)\, \partial_t f_1}{|f_1|} - \frac{\operatorname{sign}(f_2)\, \partial_t f_2}{|f_2|}.$$

Insert the full evolution (**??**): $\partial_t f_k(t, x_0) = -\xi_1\, K_{\mu_0}(x_0, x_0)\, d_0\, B_{k,1} + E_0(t)$ with $|E_0(t)| \le C'\, \psi(\delta)$ for $\psi(\delta)$ fastly decaying in δ. Then

$$\partial_t R_{12}(t, x_0) = \xi_1(t)\, K_{\mu_0}(x_0, x_0)\, (-d_0(t)) \Big(\frac{\operatorname{sign}(f_1) B_{1,1}}{|f_1|} - \frac{\operatorname{sign}(f_2) B_{2,1}}{|f_2|} \Big) + \varepsilon_0(t), \qquad |\varepsilon_0(t)| \le C''\, \psi(\delta),$$

where $\varepsilon_0(t) = E_0(t) \left(\frac{\operatorname{sign}(f_1)}{|f_1|} - \frac{\operatorname{sign}(f_2)}{|f_2|} \right)$ inherits the bound from E_0. Use $-d_0 \ge 0$. The sign condition gives $\operatorname{sign}(f_1) B_{1,1} = |B_{1,1}|$; $-\operatorname{sign}(f_2) B_{2,1} \ge -|B_{2,1}|$. The dominance inequality at x_0 is $|B_{2,1}| \le \rho_0 \frac{|f_2|}{|f_1|} |B_{1,1}|$, hence

$$\frac{|B_{1,1}|}{|f_1|} - \frac{\operatorname{sign}(f_2) B_{2,1}}{|f_2|} \ge \frac{|B_{1,1}|}{|f_1|} - \frac{|B_{2,1}|}{|f_2|} \ge (1 - \rho_0)\, \frac{|B_{1,1}|}{|f_1|},$$

so $\partial_t R_{12}(t, x_0) = (1 - \rho_0)\, \xi_1\, K_{\mu_0}(x_0, x_0)\, (-d_0)\, \frac{|B_{1,1}|}{|f_1|} + \varepsilon_0(t)$. The leading term is ≥ 0; since $K_{\mu_0}(x_0, x_1) > 0$ and fastly decaying, the off-diagonal coupling reinforces the leading term (even better positivity) while $|\varepsilon_0| \le C''\, \psi(\delta)$ is negligible for δ large. Thus $\partial_t R_{12}(t, x_0) \ge 0$ whenever the leading term dominates $|\varepsilon_0|$.

At x_1. The same argument applies by symmetry (indices $1 \leftrightarrow 2$, $x_0 \leftrightarrow x_1$): channel 2 dominates at x_1 and $\partial_t R_{21}(t, x_1) \ge 0$.

Conclusion. The log-ratio R_{12} at x_0 is non-decreasing on I; strict dominance of channel 1 at x_0 cannot be lost and is amplified whenever $|B_{1,1}|$ is not too small. By symmetry, channel 2 dominates at x_1.

H.4 On the hypothesis: when (i)–(iii) hold in practice

The hypothesis of Theorem **??** is conditional: (i) $-d_0(t) \ge 0$, (ii) $B_{1,1}(t)$ has the same sign as $f_1(t, x_0)$, and (iii) $|B_{2,1}(t)| \le \rho_0 \frac{|f_2(t,x_0)|}{|f_1(t,x_0)|} |B_{1,1}(t)|$ for some $\rho_0 \in [0, 1)$. In practice, these conditions are observed to hold in standard training setups. Under *Xavier initialization* (or similar scale-corrected schemes) for the frozen feature and mixing matrices, and *sub-Gaussian initialization* for the trainable weights w_1, w_2, the initial f_k and backprop signals $B_{k,p}$ are well-balanced across channels. The gradient-flow dynamics then tend to amplify small asymmetries: once one channel leads at x_0, (ii) and (iii) are maintained because the dominant channel receives a larger B_k and thus a larger $\partial_t f_k$, while the weaker channel's backprop stays proportionally smaller. Condition (i) holds when the network under-predicts at x_0; for squared loss $d_L \propto \hat{y} - y$, $-d_0 \ge 0$ corresponds to $\hat{y}(x_0) \le y(x_0) = +A$, which is typical before convergence. The same reasoning applies for *non-convex losses* such as cross-entropy: the loss derivative has a different form, but under-prediction at a class boundary and emergent sign-coherence under gradient flow still yield (i)–(iii) on an interval I in common regimes. Thus, although the theorem does not prove (i)–(iii) from first principles, they are consistent with and typically observed under Xavier and sub-Gaussian initialization and under losses including cross-entropy, which are standard in practice.

I Additional Experimental Details

I.1 Dataset

Target: $f(x) = \cos(f_1 \pi x^2) - 0.8\cos(f_2 \pi x^2)$ on $[-1, 1]$, symmetric in x. We use $(f_1, f_2) \in \{(36, 12), (72, 24), (144, 48)\}$ with training sizes 1000, 2000, 4000 and test sizes 1234, 2468, 4936. Train: uniform grid; test: different distribution. Sample count scales with frequency to match Nyquist density.

I.2 Hyperparameters

Table 4: Hyperparameters. Main runs: batch 100; we also try 50 and 200.

Hyperparameter	Values
Architecture	
Depth L	8 layers
Width n	1024 neurons per layer
Activation	Leaky ReLU
Rank r	$\{10, 15, 20, 25, 50, 100, 1024\}$
fixWb	True (frozen), False (trainable)
Optimization	
Optimizer	Adam
Learning rate	0.001 (initial)
Scheduler	StepLR ($\gamma = 0.9$, step_size=100)
Gradient clipping	max_norm=1.0
Batch size	$\{50, 100, 200\}$
Training epochs	10,000 (all runs)
Data	
Frequency pairs (f_1, f_2)	$(36, 12)$, $(72, 24)$, $(144, 48)$
Training samples	1000, 2000, 4000 (by frequency)
Test samples	1234, 2468, 4936 (by frequency)
Input domain	$x \in [-1, 1]$
Training Details	
Random seed	42
Device	CUDA (GPU)
Checkpointing	Every 500 epochs
Evaluation frequency	Every 50 epochs

Roughly 90 configurations (rank, fixWb, frequency, batch size); 10k epochs, no early stopping.

I.3 Plots and log-ratio setup

We plot loss and error evolution, final predictions vs. target, and layer-wise channel partials (up to 36 per layer) to inspect hierarchical frequency learning (Section **??**). **Log-ratio** (Figure **??**): 3-layer, n=1024, r=15, $\cos(8\pi x)$; SGD lr 0.01, batch 160, 10k epochs; heatmap of $R_{i,j}$ at x=0.