How feature learning breaks the curse of dimensionality in isotropic data: A mean field approach

Niclas Göring [1] **Yoonsoo Nam** [1] **Chris Mingard** [1] **Jake Reid** [2] **Ard A. Louis** [1]

Abstract

A central challenge in deep learning theory is understanding how neural networks leverage feature learning (FL) to achieve polynomial sample complexity on isotropic data distributions with low-dimensional target structure—such as k-sparse parity—while their corresponding neural tangent kernels exhibit exponential sample complexity in the ambient dimension. We explain this gap by showing that neural networks outperform kernels by developing strong anisotropy between relevant and irrelevant weight coordinates during training. This ability, which kernels fundamentally lack, enables superior generalisation. We call this mechanism input feature selection (IFS). While mean-field (MF) theory provides an interpretable framework for describing FL in neural networks, it fails to capture IFS on isotropic data distributions. To overcome this limitation, we derive a tractable extension to MF theory, using automatic relevance determination (MF-ARD), that successfully captures generalisation performance by accounting for weight anisotropy while preserving the interpretability of standard MF theory. We validate MF-ARD on several datasets, confirming that it captures essential FL dynamics that are missed by kernels and standard MF theory.

1. Introduction

Deep neural networks (NNs) achieve polynomial sample complexity on problems where the data distribution is isotropic but the target depends on only a few directions—canonical examples include k-sparse parity (Damian et al., 2022; Daniely & Malach, 2020) and multi-index models (Bietti et al., 2022). In these settings, fixed kernel methods (including the neural tangent kernel (NTK) (Jacot et al., 2018) or the infinite-width Gaussian process limit (Lee et al., 2018)) require exponentially many samples in the ambient dimension, yet finite-width networks trained with noisy gradient methods often exhibit a sharp transition to near-perfect generalisation once sufficient data are available by feature learning (FL). Understanding the theoretical mechanisms underlying this dramatic finite-width advantage remains a central challenge in deep learning theory.

A statistical physics approach to mean field (MF) theory (Mei et al., 2018; Bordelon & Pehlevan, 2023; Pacelli et al., 2023; Baglioni et al., 2024; Montanari & Urbani, 2025; Ringel et al., 2025; Lauditi et al., 2025; Celentano et al., 2025; Aiudi et al., 2025; Howard et al., 2025) provides a rigorous framework for analyzing FL beyond the kernel regime, yet existing MF approximations fall short on isotropic tasks (Figure 1). Standard MF captures the onset of feature learning as a symmetry-breaking phase transition but substantially underestimates post-transition generalisation. Recent work (van Meegen & Sompolinsky, 2025) describes how neurons can specialize differentially (coding schemes)—some aligning strongly with task structure while others remain weakly tuned—but does not address how within a *single* neuron, weights specialize to task-relevant input coordinates. This input coordinate-wise selectivity is essential for finding the task relevant features (Abbe et al., 2023), yet standard MF enforces a homogeneous weight distribution across coordinates, preventing the model from realizing the anisotropy observed in SGLD trained networks.

We address this gap by identifying a key mechanism, **input feature selection (IFS)** and introducing a minimal extension to MF theory to capture it. Through analysis of stochastic gradient Langevin dynamics (SGLD) at the level of per-coordinate weight energies, we show that gradient-weight interactions induce coordinate-dependent effective regularisation: relevant directions experience weakened decay and develop large variance, while irrelevant input coordinates remain suppressed. To incorporate IFS into a tractable MF framework, we propose **MF-ARD**, an automatic relevance determination (MacKay, 1996; Tipping, 2001) extension that endows the weight prior with learnable, coordinate-

[1] Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, United Kingdom [2]Experimental Psychology, University of Oxford, Oxford, United Kingdom. Correspondence to: Niclas Göring <niclas.goring@jesus.ox.ac.uk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

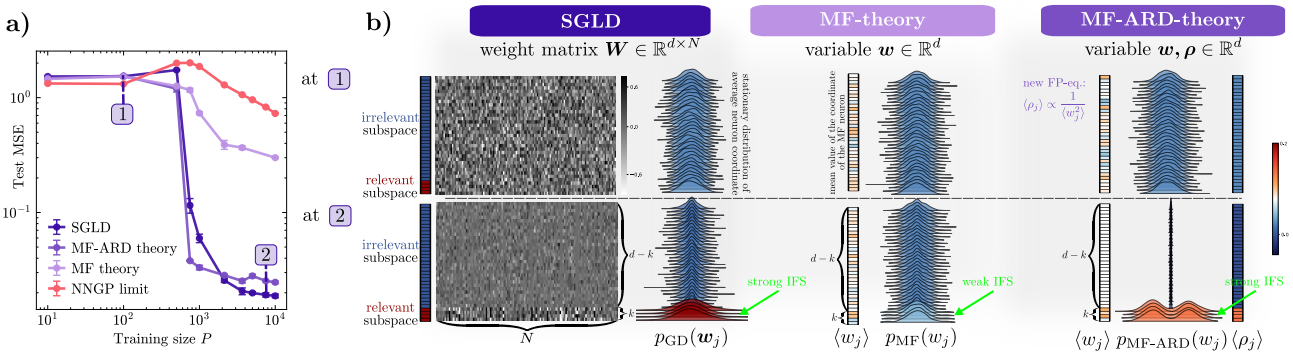

Figure 1. **a)** Test MSE as a function of training-set size P for a finite-width N two-layer network trained with SGLD, its NNGP (kernel) limit, standard MF theory, and our MF–ARD theory. The circled points $\boxed{1}$ and $\boxed{2}$ indicate two representative dataset sizes whose stationary statistics are visualized in b. **b)** Coordinate-wise stationary statistics of the averaged input-coordinate weight (SGLD: weight matrix $\boldsymbol{W} \in \mathbb{R}^{d\times N}$, MF: effective variable $w \in \mathbb{R}^d$), MF-ARD: effective variables $\boldsymbol{w}, \boldsymbol{\rho} \in \mathbb{R}^d$). For each input coordinate $j \in [d]$, we plot the stationary distribution of the corresponding averaged coordinate and its relevant summary statistics, separating the k task-relevant coordinates from the $d-k$ irrelevant ones. **SGLD** exhibits strong IFS after the transition $\boxed{2}$: relevant coordinates develop large variance while irrelevant ones remain suppressed. **MF theory** only has a fixed point (FP) equation in the coordinate means $\{\langle w_j \rangle\}_{j=1}^d$. This "mean-only" description yields weak IFS. **MF–ARD** introduces additional FP equation in the coordinate variances $\{\langle w_j^2 \rangle\}_{j=1}^d$. This new FP relation enables variance tracking that amplifies separation between relevant and irrelevant coordinates and thereby reproduces strong IFS and the post-transition generalisation gains seen in SGLD trained NNs. While the exact form of the stationary distribution $p(w_j)$ differs between SGLD and MF-ARD, they match in their second moment as we will discuss below.

wise variances. This added degree of freedom suffices to reproduce the sharp generalisation transitions observed in practice while preserving MF interpretability. Specifically, we make the following **contributions**:

- We **identify input feature selection (IFS)** as the key mechanism missing from standard MF on isotropic data, and provide a dynamical derivation showing how SGLD induces coordinate-dependent regularisation leading to strong weight anisotropy.
- We **propose MF-ARD**, a principled automatic relevance determination extension of MF theory with coordinate-wise variance parameters, and prove it enables strong IFS while standard MF cannot (Theorem 4.1).
- We establish that **MF-ARD breaks the curse of dimensionality**: unlike plain MF, whose phase boundary scales with ambient dimension d, MF-ARD's transition threshold depends only on the intrinsic task dimension k (Theorem 4.4).
- We demonstrate **quantitative agreement** between MF-ARD predictions and SGLD-trained networks across varying dataset sizes and noise levels where both standard MF and kernel theories fail (Figures 1 and 6).

2. Background

2.1. Data dependent feature learning regimes

A central question in understanding NNs and kernel methods is how data structure shapes their relative performance. Prior work on spiked-covariance inputs—where covariance partially reveals task directions—shows a reduced rich–lazy gap (Ghorbani et al., 2021; Mousavi-Hosseini et al., 2023), but the most interesting case may be the isotropic regime. Here, inputs are spherical ($XX^\top \approx I$ after normalization, where X denotes the data matrix), providing no informative anisotropy to pre-select relevant directions. When targets have low-dimensional structure—e.g., rules supported on a subset $S \subset [d]$ with $|S| = k$ (such as k-sparse parity, Section A.3)—the subsets S and S^c are input-indistinguishable, forcing learners to identify S purely from supervision. In this regime, recent work establishes a sharp NN–kernel separation: rich NNs discover relevant subspaces by developing an anisotropic weight matrix (amplifying S, suppressing S^c) and achieve polynomial sample complexity, whereas kernel methods with fixed representations require exponentially more samples (Bietti et al., 2022; Daniely & Malach, 2020; Kou et al., 2024; Suzuki et al., 2023; Telgarsky, 2022). We therefore focus on isotropic datasets with low-dimensional targets.

2.2. SGLD setup

For analytic tractability, we use SGLD, the limit of full-batch GD with weight decay and injected Gaussian noise.

This links optimisation and Bayesian inference by viewing parameter trajectories as samples from a posterior distribution ((Welling & Teh, 2011; Teh et al., 2015), see Section A.1 for details). Given a NN described by $f_{\boldsymbol{\theta}}(\boldsymbol{x}_\mu)$, a dataset $\mathcal{D} = \{(\boldsymbol{x}_\mu, \boldsymbol{y}_\mu)\}_{\mu=1}^P$, drawn from an input distribution $q(\boldsymbol{x}, \boldsymbol{y})$, the SGLD update equation is:

$$\Delta\theta_{i,t} = -\eta\left[\lambda_i\,\theta_{i,t} + \nabla_{\theta_i}\left(\frac{1}{P}\sum_{\mu=1}^{P}\big(f_{\boldsymbol{\theta}}(\mathbf{x}_\mu) - \boldsymbol{y}_\mu\big)^2\right)\right] + \sqrt{2T\,\eta}\,\xi_{i,t}, \quad \xi_{i,t} \sim \mathcal{N}(0,1), \tag{1}$$

with full-batch gradients on parameters θ_i, injected Gaussian noise ξ, learning rate η, weight decay λ_i, and noise strength set by $T = 2\kappa^2$. Additionally, we focus on two-layer fully-connected networks with input dimension d, hidden width N, input weights $\boldsymbol{w}_i \in \mathbb{R}^d$, output weights a_i, output dimension 1, and nonlinearity ϕ:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{N^\gamma}\sum_{i=1}^{N} a_i\,\phi(\boldsymbol{w}_i^\top \boldsymbol{x}), \tag{2}$$

$$w_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{d}\right), \quad a_i \sim \mathcal{N}(0, \sigma_a^2). \tag{3}$$

The initial parameters are set with Gaussians. The scale factor $N^{-\gamma}$ on the output layer enables interpolation between $\gamma = 1/2$ (NTK scaling) and $\gamma = 1$ (mean field scaling, not to be confused with MF theory, see e.g. (Mei et al., 2019) for details). Using the explicit form of $f_{\boldsymbol{\theta}}$ to rewrite the stationary distribution yields (see Section A.2):

SGLD-posterior

$$\begin{aligned}-\ln p_{\text{GD}} = &\frac{1}{2\sigma_a^2}\sum_{i=1}^{N} a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^{N}\|\boldsymbol{w}_i\|^2 \\ &+ \frac{1}{2\kappa^2 N^{2\gamma}}\sum_{i=1}^{N} a_i^2\,\Sigma(\mathbf{w}_i) \\ &+ \frac{1}{2\kappa^2 N^{2\gamma}}\sum_{i\neq i'} a_i a_{i'} G(\boldsymbol{w}_i, \boldsymbol{w}_{i'}) \\ &- \frac{1}{\kappa^2 N^\gamma}\sum_{i=1}^{N} a_i J_{\mathcal{Y}}(\mathbf{w}_i) + \text{const.},\end{aligned} \tag{4}$$

where $p_{\text{GD}} \to p_{\text{GD}}(\{\boldsymbol{w}_i\}_{i=1}^N, \boldsymbol{a}|\mathcal{D})$ with $\boldsymbol{w}_i \in \mathbb{R}^d$, and output weights $\boldsymbol{a} \in \mathbb{R}^N$.

The posterior is shaped by three competing forces: the self-energy $\Sigma(\boldsymbol{w}) = \frac{1}{P}\sum_\mu \phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)^2$ regularises neuron magnitude; the data coupling $J_{\mathcal{Y}}(\boldsymbol{w}) = \frac{1}{P}\sum_\mu[\phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu) y(\boldsymbol{x}_\mu)]$ rewards target alignment and breaks symmetry; and the interaction kernel $G(\boldsymbol{w}_i, \boldsymbol{w}_{i'}) = \frac{1}{P}\sum_\mu \phi(\boldsymbol{w}_i^\top \boldsymbol{x}_\mu)\,\phi(\boldsymbol{w}_{i'}^\top \boldsymbol{x}_\mu)$ mediates neuron cooperation. As P increases (or κ decreases), the likelihood tilts the posterior away from the isotropic prior, creating anisotropy along task-relevant directions and cooperative alignment via G, yielding a transition to a non-Gaussian posterior on low-rank structures. Each neuron's optimal weight depends on all others, preventing factorisation into independent terms, so that the posterior is generally intractable.

2.3. MF Theory: Removing off-diagonal couplings

To make the posterior in Equation (4) tractable, we use self-consistent MF theory (related to the MF theory in (Rubin et al., 2024; Ringel et al., 2025)), which replaces the correlated posterior $p(\boldsymbol{W}|\mathcal{D})$ with a factorized approximation $p(\boldsymbol{W}|\mathcal{D}) \approx \prod_{i=1}^N p_{\text{MF}}(\boldsymbol{w}_i)$ where each neuron is independent. Each neuron interacts not with all others but with their average behavior, the mean field $\langle f\rangle$:

$$\sum_{i'\neq i} a_{i'}\,G(\boldsymbol{w}_i, \boldsymbol{w}_{i'}) \overset{\text{MF}}{\approx} \frac{1}{P}\sum_\mu \phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)\,N^\gamma\langle f(\boldsymbol{x})\rangle. \tag{5}$$

We require $\langle f(\boldsymbol{x})\rangle = N^{1-\gamma}\cdot\mathbb{E}_{(\boldsymbol{w},a)\sim p(\boldsymbol{w},a)}\big[a\,\phi\big(\boldsymbol{w}^\top \boldsymbol{x}\big)\big]$ for self-consistency. Expanding $\langle f(\mathbf{x})\rangle = \sum_A m_A \chi_A(\mathbf{x})$ in an orthonormal basis $\{\chi_A\}$ with feature coefficients $m_A = \mathbb{E}_{\boldsymbol{x}}\big[\langle f(\boldsymbol{x})\rangle\,\chi_A(\boldsymbol{x})\big]$ yields:

MF theory (fixed point equations)

$$\begin{aligned}-\ln p_{\text{MF}} = &\frac{a^2}{2\sigma_a^2} + \frac{d}{2\sigma_w^2}\sum_{j=1}^{d} w_j^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}}\Sigma(\boldsymbol{w}) \\ &- \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w})\right) \\ &m_A = N^{1-\gamma}\langle a J_A(\boldsymbol{w})\rangle_{p_{\text{MF}}} \quad \forall\, A,\end{aligned} \tag{6}$$

where $p_{\text{MF}} \to p(\boldsymbol{w}, a|\{m_A\}, \mathcal{D})$, $\boldsymbol{w} \in \mathbb{R}^d$, $a \in \mathbb{R}$ and $J_A(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}[\phi(\boldsymbol{w}^\top \boldsymbol{x})\chi_A(\boldsymbol{x})]$.

Interpretation

> m_A: Feature coefficients measure how strongly the average neuron aligns with a certain basis function.

> $G \to \sum_A m_A J_A$: Coupling of the single neuron to the effective field from all the other neurons.

MF replaces the intractable $N \times d$ weight matrix problem with a self-consistent problem for a single d-dimensional vector. These equations are solved by iteration, computing $\Sigma(\boldsymbol{w}), J_A(\boldsymbol{w}), J_{\mathcal{Y}}(\boldsymbol{w})$ directly from the finite training dataset of size P (Section E). The NNGP limit also emerges naturally, see Section A.5.

3. Theory I: Input Feature Selection as Implicit regularisation

For simplicity, we focus on k-sparse parity in this and the following section: $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ with $S = \{1, \ldots, k\}$ and $S^c = \{k+1, \ldots, d\}$, a canonical example of the isotropic datasets in Section 2.1. We identify IFS as the dynamical mechanism driving what can be viewed as an implicit regu-

larisation in gradient-based training on sparse tasks. Prior work (Abbe et al., 2023; 2022; Ba et al., 2022; Shi et al., 2022) established that weight matrix anisotropy is central for generalisation. Here, we provide a simplified analytical derivation focusing on *how* this arises during training to influence the stationary SGLD distribution.

3.1. Input Feature Selection and coordinate-wise effective decay

To understand the origin of IFS, we analyze the dynamics of the weight matrix rows. Let $\boldsymbol{w}_j(t) \in \mathbb{R}^N$ denote the j-th row of the weight matrix $\boldsymbol{W}(t) \in \mathbb{R}^{d\times N}$, corresponding to the j-th input feature. By abuse of notation, we write $\boldsymbol{w}_j^2(t) := \|\boldsymbol{w}_j(t)\|_2^2$. Under SGLD dynamics (see (1)) with weight decay λ, applying Itô's lemma to $\boldsymbol{w}_j^2(t)$ yields the *exact* evolution equation

$$\frac{d}{dt}\langle \boldsymbol{w}_j^2(t)\rangle = -2\eta\left(\lambda\,\langle \boldsymbol{w}_j^2(t)\rangle + \langle \hat{\rho}_j(t)\,\boldsymbol{w}_j^2(t)\rangle\right) + 2T\eta N, \tag{7}$$

where we define the normalized gradient–weight alignment $\hat{\rho}_j(t) = \frac{\langle \boldsymbol{w}_j(t), \nabla_{\boldsymbol{w}_j}\ell(t)\rangle}{\boldsymbol{w}_j^2(t)}$. Note that $\hat{\rho}_j$ can be negative: $\hat{\rho}_j < 0$ means the descent direction $-\nabla_{\boldsymbol{w}_j}\ell$ is *aligned* with $\boldsymbol{w}_j$, which tends to increase $\boldsymbol{w}_j^2$ and can counteract weight decay. On sparse tasks, we observe a clear separation in this alignment signal over training (Fig. 2d): for relevant coordinates $j \in S$, $\hat{\rho}_j(t)$ becomes strongly negative prior to the drop in test error, while for irrelevant coordinates $j \in S^c$ it remains near zero.

- **Irrelevant coordinates ($j \in S^c$):** The gradient signal carries no consistent component along $\boldsymbol{w}_j$, so $\hat{\rho}_j(t) \approx 0$. These rows therefore effectively experience *full* weight decay λ, and their $\boldsymbol{w}_j^2$ are driven toward small values.
- **Relevant coordinates ($j \in S$):** The loss induces systematic alignment, with $\hat{\rho}_j(t) < 0$ for extended periods. This provides an *anti-decay* contribution that partially cancels λ, allowing $\boldsymbol{w}_j^2$ to persist or grow.

In this sense, $\hat{\rho}_j(t)$ acts like a *coordinate-wise force* (or effective potential slope) induced by the loss: SGLD evolves $\boldsymbol{w}_j^2$ under the competition between uniform weight decay λ and the data-dependent drift $\hat{\rho}_j(t)$, which selectively weakens decay along task-relevant directions.

3.2. The stationary distribution

At stationarity, setting $\frac{d}{dt}\langle \boldsymbol{w}_j^2\rangle = 0$ in Equation (7) gives the *exact* coordinate-wise balance condition:

$$\lambda\,\langle \boldsymbol{w}_j^2\rangle_\infty \;+\; \langle \hat{\rho}_j\,\boldsymbol{w}_j^2\rangle_\infty = TN, \tag{8}$$

where $\langle\cdot\rangle_\infty$ denotes the expectation w.r.t. to this stationary distribution if it exists. A separation between the alignment

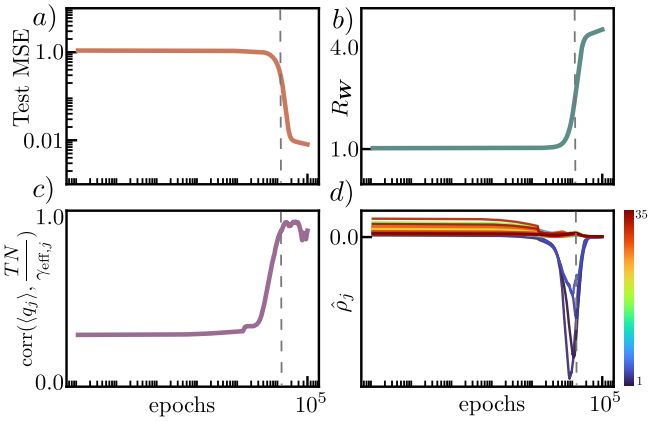

Figure 2. A 1-hidden layer ReLU network ($N = 512$) is trained on $k = 4$ sparse parity in $d = 35$. **a)** Test error versus epochs. **b)** The weight anisotropy ratio R_W increases sharply coincident with the drop in test error. **c)** Validation of the stationary energy balance (8), comparing measured $\boldsymbol{w}_j^2$ with the theoretical stationary prediction by plotting the mean correlation of $\boldsymbol{w}_j^2$ with $\frac{TN}{\lambda+\hat{\rho}_j}$. **d)** Evolution of $\hat{\rho}_j$: support features ($j \in S$) become strongly negative prior to the test error drop, while non-support features ($j \in S^c$) remain near zero (colors indicate coordinate indices).

$\langle \hat{\rho}_j\rangle_\infty$ in equilibrium on S and S^c implies a separation of stationary row-energies $\langle \boldsymbol{w}_j^2\rangle_\infty$, producing an anisotropic stationary distribution Figure 2b) and Figure 1a).

Definition: input feature selection (IFS)
A network undergoes IFS if the stationary distribution is characterized by a strong weight-anisotropy ratio R_W in the epochs $\to \infty$ limit :

$$R_{\boldsymbol{W}}^{\infty} = \frac{\sqrt{d-k}\cdot\langle \boldsymbol{w}_{j\in S}^2\rangle_\infty}{\sqrt{k}\cdot\langle \boldsymbol{w}_{j\in S^c}^2\rangle_\infty} \gg 1. \tag{9}$$

3.3. Empirical evidence

An ensemble of trained NNs In Figure 3, we draw random dataset sizes $P \in [10, 20000]$ and noise levels $\kappa_0 \in [5\cdot 10^{-5}, 5\cdot 10^{-2}]$. We train the full NN for 10^5 epochs and color code the final test error. This confirms empirically that a minimal level of anisotropy is needed to generalise.

Causal validation To show that IFS is the causal driver of generalisation, we introduce a counterfactual regularisation term to suppress IFS. We define the isotropy penalty: $\ell_{\mathrm{iso}}(\boldsymbol{W}) = \lambda_h \sum_{i=1}^{N} \mathrm{Var}_{j\in[d]}\left(|w_{ji}|\right)$. This term penalizes the variance of weight magnitudes across input dimensions j within each neuron i. Minimizing ℓ_{iso} forces the network to maintain isotropic usage of inputs (i.e., $|w_{1i}| \approx |w_{2i}| \approx$

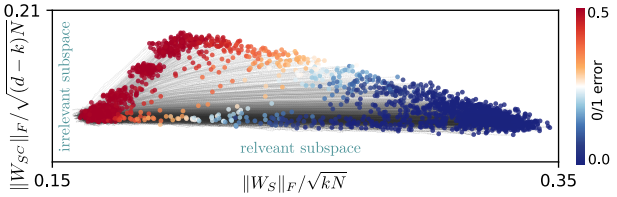

Figure 3. Each point corresponds to a one-hidden-layer $N = 512$ network trained with SGLD on k-sparse parity, with training-set size $P \in [10, 20000]$ and noise level $\kappa_0 \in [5 \cdot 10^{-5}, 5 \cdot 10^{-2}]$. The x-axis shows the normalized Frobenius norm of the first-layer weights restricted to the task-relevant coordinates S, $\|W_S\|_F/\sqrt{kN}$, and the y-axis shows the corresponding norm on the irrelevant coordinates S^c, $\|W_{S^c}\|_F/\sqrt{(d-k)N}$. Color denotes final test error. Low test error is achieved only when weight mass concentrates in the relevant subspace, illustrating that strong IFS $R_{\boldsymbol{W}}^{\infty} \gg 1$ correlates with generalisation in this isotropic setting.

. . .), effectively "switching off" the IFS mechanism while allowing other forms of coordination between neurons (including those discussed in (van Meegen & Sompolinsky, 2025)). Figure 4 clearly shows that growing regularisation causes the NN to generalise progressively worse. This confirms that IFS is not just a correlate of learning, but the necessary mechanism for generalisation in this regime.

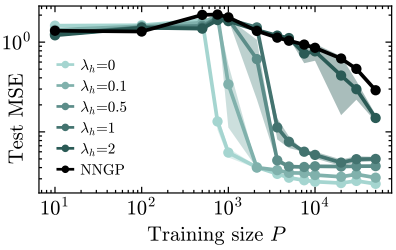

Figure 4. Test loss for k-sparse parity as a function of the regularisation λ_h that forces coordinates within a neuron to be homogeneous (preventing IFS). Increasing λ_h reduces performance till the NN performance degrades back to the Kernel (NNGP) limit. This confirms that coordinate-wise anisotropy is the necessary condition for sample-efficient learning in this setting.

4. Theory II: MF model of IFS

In this section, we focus on deriving an effective model of IFS in a MF context.

4.1. FL in the simple MF model

The simplicity of MF theory allows us to interpret FL as *self-consistent symmetry breaking* in p_{MF}, where learning emerges from competing forces. Below critical values (P_c, κ_c), the only stable fixed point (FP) of Equation (6) is $m_A = 0\ \forall A$, and p_{MF} corresponds to the Gaussian prior. The Gaussian prior and self-energy $\Sigma(\boldsymbol{w})$ regularize toward an isotropic, high-entropy state penalizing large weights. Upon crossing (P_c, κ_c)—when data signal (controlled by P and noise κ) becomes strong enough—the $m_A = 0$ FP destabilizes (see Theorem D.7 for proof and explicit κ_c formula). The neuron-data coupling $J_{\mathcal{Y}}$, acting as an external field, rewards neurons whose activations $\phi(\boldsymbol{w}^\top \boldsymbol{x})$ correlate with target $y(\boldsymbol{x})$, breaking the prior's symmetry. Weights pull toward task-aligned configurations, making p_{MF} non-Gaussian. For target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$, the order parameter is the feature coefficient m_S, and the FP equations in (6) become (derivation in Section D.5):

$$m_S = \frac{N^{1-2\gamma}}{\kappa^2}(1 - m_S) \underbrace{\left\langle \frac{J_S(\boldsymbol{w})^2}{\sigma_a^{-2} + \dfrac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}} \right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)}}_{\omega_0}. \tag{10}$$

The phase transition occurs when $\omega_0 > 1$ (Figure 5): for SGLD-trained networks and MF theory, $m_S = 0$ until critical P_c, then increases abruptly. The NNGP model, lacking non-linear feedback, shows only smooth, continuous rise. In Section A.4, we link this FL phase transition to outlier eigenvalue emergence in the Gram kernel $G(\boldsymbol{w}_i, \boldsymbol{w}_{i'})$.

4.2. Problem: Simple MF underestimates IFS

However, Figure 1**a)** reveals a critical discrepancy: post-transition, the SGLD-trained network generalises much more rapidly than the MF model. This performance gap indicates that standard MF theory misses the structural mechanism driving IFS. The root cause lies in the MF FP (6), where all neurons are modeled as i.i.d. samples from a single distribution $p_{\text{MF}}(a, \boldsymbol{w})$. In MF theory, we replace the full weight matrix $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_d] \in \mathbb{R}^{N\times d}$ by only modeling the mean $w_j = \langle \boldsymbol{w}_j \rangle_{p_{\text{MF}}}$ for each of the d row vectors (each of dimension N). While, as illustrated in Figure 5, tracking the mean is enough to predict the *onset* of the phase transition, this does not, as shown through our dynamical analysis in Section 3, capture the rapid increase in generalisation, which originates instead from a strong symmetry breaking in the *variance* $\boldsymbol{w}_j^2$ of the row-vectors $\boldsymbol{w}_j$. We formally describe the fact that standard MF theory cannot produce a strong anisotropy in the variances (IFS), leading to a wrong prediction in the generalisation error (see Figure 1 a) in the theorem below:

Theorem 4.1. *Let $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ and $\{w_j\}_{j=1}^d$ a solution to the MF FP equations in (6) and $\boldsymbol{w}_S = \{w_j\}_{j\in S}$, $\boldsymbol{w}_S^c = \{w_j\}_{j\in S^c}$. The FP of the plain MF theory (6) cannot develop a strong weight-anisotropic ratio:*

$$R_{\boldsymbol{w}}^{MF} = \frac{\sqrt{d-k}\|\boldsymbol{w}_S\|^2}{\sqrt{k}\|\boldsymbol{w}_{S^c}\|^2} = 1 + O\left(\frac{1}{d}\right). \tag{11}$$

Proof. See Section D.7 for the proof. □

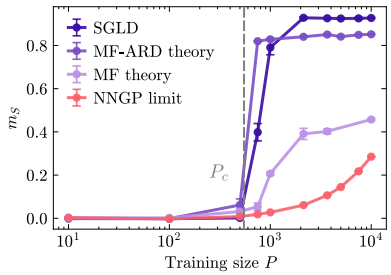

Figure 5. Feature coefficient m_S vs. P for a k-sparse parity target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x}) = \Pi_{j\in S} x_j$ with index $S = \{0, 1, 2, 3\}$ in $d = 35$ with a ReLU network ($N = 512$, see Section A.3; Section F for more details). SGLD, MF and MF-ARD exhibit a phase transition at P_c, NNGP grows smoothly, no phase-transition or FL.

4.3. A minimal MF model of IFS: Tracking Coordinate Variances

How can we capture IFS within the tractable MF framework? As discussed above, MF theory tracks the mean of each row-vector of the weight matrix but not its variance which makes it incapable of modeling IFS and the associated ability of NNs to generalise. Hence, we propose to extend MF theory to track the variances w_j^2 of each row vector of the weight matrix $\boldsymbol{W}(t)$ as explicit variables.

Intuition: Our goal is to introduce the variances such that the second moment of our new MF theory matches the stationary law of SGLD (Equation (8)). Assuming that the gradient-weight alignment in the stationary state acts as a radial fixed force $\langle \hat{\rho}_j \, \boldsymbol{w}_j^2 \rangle_\infty \approx \rho_j^r \langle \boldsymbol{w}_j^2 \rangle_\infty$ and using our standard MF independence assumption $\langle \boldsymbol{w}_j^2 \rangle_\infty \overset{MF}{\approx} N \cdot \langle w_j^2 \rangle_{\text{MF}}$, Equation (8) becomes

$$\lambda \, \langle w_j^2 \rangle_{\text{MF}} \; + \; \rho_j^r \cdot \langle w_j^2 \rangle_{\text{MF}} = T. \tag{12}$$

This is equivalent to assuming the likelihood gradient behaves as if it came from an effective potential $\frac{\rho_j^r}{2} \|\boldsymbol{w}_j\|_2^2$ in coordinate j (linear response, see Figure 2c) for evidence that this assumption holds for SGLD in equilibrium). We now apply a standard technique from statistical physics and look for a minimal enlargement of the state space of the MF theory from $\{(w_j, m_A)\}_{j=1}^d \to \{(w_j, m_A), (w_j^2, \rho_j)\}_{j=1}^d$ to match the stationary law above.

4.4. Automatic Relevance Determination MF theory

The first step is to introduce ρ_j into the MF distribution $p_{\text{MF}}(\boldsymbol{w})$ to create a feed-back loop. Since ρ_j effectively acts as a precision (controlling the variance of the coordinates), the simplest model assumption is that, conditioned on the precision, the weights are Gaussian: $p(w_j \mid \rho_j) = \mathcal{N}(0, \rho_j^{-1})$. Such a conditional prior is called automatic relevance determination (ARD) in Bayesian NNs (MacKay, 1996; Tipping, 2001). As ρ_j is now a variable in our theory, we need to choose a prior. The standard choice is a gamma distribution, which ensures positivity of the variances and provides conjugacy with the Gaussian likelihood, which enables computational tractability.

Theorem 4.2. *Assume the row-wise weights follow a Gaussian $p(\boldsymbol{w}_j \mid \rho_j) = \mathcal{N}(0, \rho_j^{-1}\mathbb{I})$ and the precisions follow a conjugate Gamma prior $p(\rho_j) = \Gamma(\alpha_0, \beta_0)$. Enforcing the consistency condition that $\rho_j \to d/\sigma_w^2$ (isotropic prior) when the data signal vanishes requires setting $\beta_0 = \alpha_0 \frac{\sigma_w^2}{d}$. The resulting self-consistent FP for the variance order parameter is: $\rho_j = \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\langle w_j^2 \rangle_{p_{\text{ARD}}}}$.*

Proof. See Section A.8 for proof. □

The FP equation for MF-ARD has the following form:

MF-ARD theory (fixed point equations)

$$-\ln p_{\text{ARD}} = \frac{a^2}{2\sigma_a^2} + \frac{1}{2}\sum_{j=1}^{d} \rho_j w_j^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}} \Sigma(\boldsymbol{w}) - \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w}) \right) \tag{13}$$

$$m_A = N^{1-\gamma} \langle a J_A(\boldsymbol{w}) \rangle_{p_{\text{ARD}}} \; \forall \, A$$
$$\rho_j = \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\langle w_j^2 \rangle_{p_{\text{ARD}}}} \tag{14}$$

where $p_{\text{ARD}} \to p_{\text{ARD}}(\boldsymbol{w}, a | \{m_A, \rho_i\}, \mathcal{D})$ and $\boldsymbol{w} \in \mathbb{R}^d$, $\boldsymbol{\rho} \in \mathbb{R}^d$, $a \in \mathbb{R}$.

where the parts in blue are the new components compared to the simple MF theory from Equation (6).

Interpretation (means vs. variances): The MF couplings $\{m_A\}$ control *mean alignment* (a first-moment effect),

$$\langle w_j \rangle_{p_{\text{MF}}} \longmapsto m_A \longmapsto p_{\text{MF}}(\boldsymbol{w}) \tag{15}$$

whereas the precisions $\{\rho_j\}$ control *coordinate-wise shrinkage* (a second-moment effect). The map

$$\langle w_j^2 \rangle_{p_{\text{ARD}}} \longmapsto \rho_j \longmapsto p_{\text{ARD}}(\boldsymbol{w}) \tag{16}$$

realises the feedback underlying IFS: relevant coordinates acquire larger $\langle w_j^2 \rangle$, which shrinks ρ_j via (14) and further reduces shrinkage, irrelevant coordinates do the opposite. In this way, MF–ARD remains a MF theory (neurons are still independent), but with an enlarged set of order parameters that is sufficient to capture the coordinate-dependent variance structure induced dynamically by SGLD. See Section C for a comparison to van Meegen & Sompolinsky (2025).

4.5. Connection of MF-ARD to IFS in SGLD

The link between the dynamical IFS mechanism and the stationary Bayesian ARD mechanism becomes explicit at the stationary state of SGLD (Equation (7)). We can show that MF-ARD matches the SGLD stationary distribution in the first (this is true from the initial MF theory) and second moment (up to order $O(\frac{\alpha_0}{N})$) as we initially aimed at (see Section 4.3):

Theorem 4.3. *Under the linear response assumption $\langle \hat{\rho}_j \, \boldsymbol{w}_j^2 \rangle_\infty \approx \rho_j^r \langle \boldsymbol{w}_j^2 \rangle_\infty$, and defining the effective SGLD precision $\rho_{\text{eff},j} := \frac{\lambda+\rho_j^r}{T}$, assuming $\alpha_0 \ll N$, it holds:*

$$\text{SGLD:} \quad \rho_{\text{eff},j} = \frac{N}{\langle \boldsymbol{w}_j^2 \rangle_\infty} \tag{17}$$

$$\text{MF-ARD:} \quad \rho_j = \frac{1}{\langle w_j^2 \rangle_{p_{\text{MF-ARD}}}} \left(1 + O\left(\frac{\alpha_0}{N}\right)\right). \tag{18}$$

Proof. See Section D.8 for the proof. □

We interpret the MF-ARD precision ρ_j as the total effective precision $\rho_{\text{eff},j} = \frac{\lambda+\rho_j^r}{T}$ in SGLD, i.e. it includes the baseline weight decay and the temperature. In practice the assumption $\alpha_0 \ll N$ always holds true (see Section F for the exact hyperparameters used). The stationary laws for the variances are equivalent up to a factor N which vanishes once we use the MF assumption (see Equation (12)) $\langle \boldsymbol{w}_j^2 \rangle_\infty \overset{MF}{\approx} N \cdot \langle \boldsymbol{w}_j^2 \rangle_{\text{MF}}$. The gradient-weight alignment term modulates the effective "precision" (posterior variance) for each coordinate: relevant features experience strong negative alignment that reduces their effective precision and allows large weights, while irrelevant features retain high precision. This produces an anisotropic stationary distribution even with isotropic inputs. MF-ARD captures this implicit dynamical mechanism explicitly within a tractable, static MF FP.

4.6. How MF-ARD beats the curse of dimensionality

Let $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ with $|S| = k$ in d dimensions and $\phi = \text{ReLU}$. Write κ_c^2 for the critical noise at onset of the phase transition. We assume the infinite data limit and focus on the phase transition in the critical noise as it is analytically more tractable than critical data size.

Theorem 4.4. *There exists an outer iteration $t_0 = \mathcal{O}(1)$ of the FP algorithm and a constant $\varepsilon_0 > 0$ (independent of d) such that we get ε_0-level symmetry breaking towards S: $\min_{j\in S} \langle w_j^2 \rangle_{p_{ARD}}^{t_0} - \max_{j\notin S} \langle w_j^2 \rangle_{p_{ARD}}^{t_0} \geq c \cdot \varepsilon_0$ (see Section D.6 for details). Then, the critical noise scales as:*

$$\kappa_c^2 = \begin{cases} \Theta\left(\sqrt{1/(dk)}\right) & \textit{(plain MF)}, \\ \Theta\left(\sqrt{1/k}\right) & \textit{(MF–ARD)}. \end{cases} \tag{19}$$

Proof. See Section D.6 for a proof and exact constants. Here we summarize the scaling. □

ε-symmetry breaking assumes that there is already a very small difference in the variance for weights that are relevant vs. irrelevant for learning the target. Given this asymmetry, MF-ARD can amplify it far more effectively than plain MF. For plain MF, the noise threshold required to transition from high to low generalisation error (equivalently $m_S = 0 \to m_S \neq 0$) scales with the ambient dimension d, requiring noise levels $1/\sqrt{d}$ times smaller than MF-ARD. This dimensional dependence is eliminated in MF-ARD. Consequently, the phase boundary scales with the intrinsic problem dimension k rather than the ambient dimension d. While most theorems are specifically proven for $\phi =$ReLU, we effectively observe the same mechanisms for other activation functions, see Section A.11.

5. Results

5.1. Numerical prediction of generalisation errors across regimes

We evaluate on k-sparse parity, a setting where fixed kernels are known to require super-polynomial samples in d, k, while finite-width networks can have lower sample complexity due to FL (Daniely & Malach, 2020; Damian et al., 2022; Barak et al., 2022). Figure 6 shows test-MSE heatmaps over dataset size P and noise κ for (a) SGLD-trained networks, (b) plain MF, and (c) MF–ARD. All three display a transition from high to low error as P increases or κ decreases. Plain MF detects *when* learning starts but yields a diffuse, shifted boundary and overestimates post-transition error. MF–ARD, with a single additional set of order parameters $\{\rho_j\}$, closely tracks both the *location* and the *sharpness* of the SGLD phase boundary and reproduces the "helpful noise" regime in which moderate κ lowers the critical sample size P_c (the kink around $\kappa = 0.05$).

In Figure 9 (Appendix) we present the same phase diagram analysis for a single index-model task (Bietti et al., 2022), where the target function takes the form $y(\boldsymbol{x}) = g(\mathbf{v}^T \boldsymbol{x})$. Here, $\boldsymbol{v}, \boldsymbol{x} \in \mathbb{R}^d$, the input $\boldsymbol{x}$ follows a standard multivariate normal distribution $\mathcal{N}(0, \mathbb{1})$, $v_{j\in S} = 1/\sqrt{k}$, $v_{j\notin S} = 0$ for an index set with $|S| = k$ and g is a non-linear function (Hermite polynomial in our case).

5.2. Grokking and the scaling with depth

IFS in deep networks: To validate that IFS extends to deep networks, we analyze first-layer weight anisotropy $R_{\boldsymbol{W}^1}$ (Equation (9)) and layer-wise sensitivity $R_{h^{(\ell)}} = \frac{\mathbb{E}_{j\in S}[\|\partial h^{(\ell)}/\partial x_j\|^2]}{\mathbb{E}_{j\notin S}[\|\partial h^{(\ell)}/\partial x_j\|^2]}$, where $h^{(\ell)}(\boldsymbol{x}) = \phi(\boldsymbol{W}^{(\ell)}\boldsymbol{x} + \boldsymbol{b}^{(\boldsymbol{\ell})})$ is the ℓ-th hidden activation. This quantifies the representation's effective signal-to-noise ratio at depth ℓ. As shown

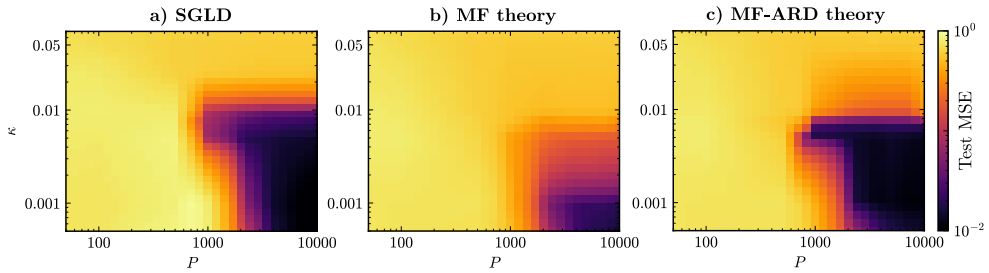

Figure 6. **Generalisation phase diagram (Test MSE) vs. dataset size P and noise κ for k-sparse parity target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ with $S = \{0, 1, 2, 3\}$ in $d = 35$. a) SGLD, b) MF, c) MF-ARD:** All panels show a transition from high to low test error as P increases or κ decreases. However, the plain MF theory strongly underestimates the sharpness of the transition while the MF-ARD theory largely reproduces the shape/location of this boundary. Training details are in Section F.

in Figure 7a), the network solves the task when these ratios diverge. While the first layer initiates selection ($R_{h^1} \approx 2.5$, $R_{\boldsymbol{W}^1} > 1$), subsequent layers progressively amplify filtering, reaching $R_{h^6} \approx 20$ at the final layer. This suggests deep networks overcome the curse of dimensionality by hierarchically refining first-layer IFS dynamics, transforming weak input signals into dominant representations.

IFS in grokking: Figure 7b) illustrates grokking and token-level IFS emergence in a small transformer trained on modular addition with distractor tokens. The top panel shows training and test accuracy versus epochs: the model fits training data early while test accuracy remains at chance, then jumps abruptly to near-perfect generalisation (grokking). The bottom panel tracks two IFS order parameters quantifying whether learning concentrates on the two task-relevant tokens (addends at positions $0, 1$): $M_{\text{attn}}(t)$, mean CLS attention mass on positions $0, 1$ (averaged over heads/layers; see Section A.10), and $M_{\text{grad}}(t)$, the fraction of embedding-gradient norm on positions $0, 1$ relative to all positions. Both remain small during memorization, then increase sharply when test accuracy jumps, indicating generalisation onset coincides with rapid attention and gradient re-routing onto relevant tokens.

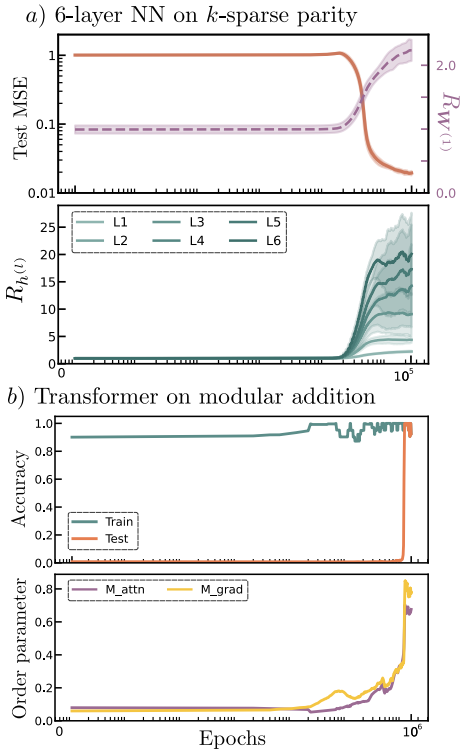

Figure 7. **a)** A 6-layer network on k-sparse parity solves the task only as first-layer weight anisotropy $R_{\mathbf{W}^1}$ and layer-wise sensitivity ratios ($R_{h^{(l)}}$) diverge, with deeper layers strongly amplifying the filtering. **b)** A small transformer on modular addition with distractor tokens exhibits grokking: test accuracy jumps when attention and gradient mass rapidly concentrate on the two relevant tokens, as measured by M_{attn} and M_{grad}.

6. Conclusion

We extend MF theory using ARD to track both mean and variance of the SGLD stationary distribution, producing a tractable mechanistic theory for FL in NNs on sparse isotropic datasets. Our theory explains why finite-width networks outperform fixed-kernel methods: they engage in self-reinforcing dynamics that reshape their feature space to amplify task-relevant signals, dramatically improving sample complexity. In particular, the non-linear IFS mechanism—absent in standard MF but captured by MF-ARD—is central to DNNs' ability to break the curse of dimensionality. Limitations of our approach include the focus on two-layer networks and tasks with sparse structures for which our MF-ARD theory may be best suited. A limitation shared with other works on these sparse datasets is the question of how far the FL mechanisms observed here apply to datasets requiring distributed or smooth representations. This remains an open challenge, see e.g. (Petrini et al., 2022).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Understanding how DNNs work, as we attempt here, may have potential societal consequences, none which we feel must be specifically highlighted here.

LLM Usage

LLMs were used to aid in checking and writing proofs, and also for reviewing the main text.

References

Abbe, E., Cornacchia, E., Hazla, J., and Marquis, C. An Initial Alignment between Neural Network and Target is Needed for Gradient Descent to Learn. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 33–52. PMLR, June 2022. URL https://proceedings.mlr.press/v162/abbe22a.html. ISSN: 2640-3498.

Abbe, E., Boix-Adsera, E., and Misiakiewicz, T. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics, August 2023. URL http://arxiv.org/abs/2302.11055. arXiv:2302.11055 [cs].

Aiudi, R., Pacelli, R., Baglioni, P., Vezzani, A., Burioni, R., and Rotondo, P. Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks. *Nature Communications*, 16(1):568, 2025.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation, May 2022. URL http://arxiv.org/abs/2205.01445. arXiv:2205.01445 [stat].

Baglioni, P., Pacelli, R., Aiudi, R., Di Renzo, F., Vezzani, A., Burioni, R., and Rotondo, P. Predictive power of a bayesian effective action for fully connected one hidden layer neural networks in the proportional limit. *Physical Review Letters*, 133(2):027301, 2024.

Barak, B., Edelman, B., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Bietti, A., Bruna, J., Sanford, C., and Song, M. J. Learning single-index models with shallow neural networks, 2022. URL http://arxiv.org/abs/2210.15651.

Bordelon, B. and Pehlevan, C. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36:9707–9750, 2023.

Celentano, M., Cheng, C., and Montanari, A. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, June 2025. doi: 10.48550/arXiv.2112.07572. URL https://arxiv.org/abs/2112.07572.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html. ISSN: 2640-3498.

Daniely, A. and Malach, E. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When Do Neural Networks Outperform Kernel Methods? *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124009, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3a81. URL http://arxiv.org/abs/2006.13409. arXiv:2006.13409 [stat].

Howard, J. N., Jefferson, R., Maiti, A., and Ringel, Z. Wilsonian renormalization of neural network gaussian processes. *Machine Learning: Science and Technology*, 6(2):025038, 2025.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Kou, Y., Chen, Z., Gu, Q., and Kakade, S. M. Matching the Statistical Query Lower Bound for k-Sparse Parity Problems with Sign Stochastic Gradient Descent, December 2024. URL http://arxiv.org/abs/2404.12376. arXiv:2404.12376 [cs].

Lauditi, C., Bordelon, B., and Pehlevan, C. Adaptive kernel predictors from feature-learning infinite limits of neural networks. *arXiv preprint arXiv:2502.07998*, 2025.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes, 2018. URL https://arxiv.org/abs/1711.00165.

MacKay, D. J. Bayesian non-linear modeling for the prediction competition. In *Maximum Entropy and Bayesian Methods: Santa Barbara, California, USA, 1993*, pp. 221–234. Springer, 1996.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.

Montanari, A. and Urbani, P. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

Mousavi-Hosseini, A., Wu, D., Suzuki, T., and Erdogdu, M. A. Gradient-Based Feature Learning under Structured Data. 2023.

Pacelli, R., Ariosto, S., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12):1497–1507, 2023.

Petrini, L., Cagnetta, F., Vanden-Eijnden, E., and Wyart, M. Learning sparse features can lead to overfitting in neural networks. *Advances in Neural Information Processing Systems*, 35:9403–9416, 2022.

Ringel, Z., Rubin, N., Mor, E., Helias, M., and Seroussi, I. Applications of statistical field theory in deep learning, 2025. URL https://arxiv.org/abs/2502.18553.

Rubin, N., Seroussi, I., and Ringel, Z. Grokking as a first order phase transition in two layer networks. *arXiv preprint arXiv:2310.03789*, May 2024. URL https://arxiv.org/abs/2310.03789. Preprint.

Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

Suzuki, T., Wu, D., Oko, K., and Nitanda, A. Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. *Advances in Neural Information Processing Systems*, 36:34536–34556, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6cc321baf0a8611b1d1bdbd18822667b-Abstract-Conference.html.

Teh, Y. W., Thiéry, A., and Vollmer, S. Consistency and fluctuations for stochastic gradient langevin dynamics, 2015. URL https://arxiv.org/abs/1409.0578.

Telgarsky, M. Feature selection with gradient descent on two-layer networks in low-rotation regimes, August 2022. URL http://arxiv.org/abs/2208.02789. arXiv:2208.02789 [cs].

Tipping, M. E. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1 (Jun):211–244, 2001.

van Meegen, A. and Sompolinsky, H. Coding schemes in neural networks learning classification tasks. 16 (1):3354, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58276-6. URL https://doi.org/10.1038/s41467-025-58276-6.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, ICML'11, pp. 681–688, Bellevue, Washington, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5. doi: 10.5555/3104482.3104568. URL https://dl.acm.org/doi/10.5555/3104482.3104568.

A. Additional Background

A.1. Equivalence of Langevin Dynamics and Bayesian Inference

In this section, we provide a detailed derivation for the equivalence between the stationary distribution of a network trained with Stochastic Gradient Langevin Dynamics (SGLD) and the posterior distribution of a corresponding Bayesian model.

Network and Priors We analyze a two-layer neural network with the functional form:

$$f(\boldsymbol{x}) = \frac{1}{N^\gamma} \sum_{i=1}^{N} a_i \phi(\boldsymbol{w}_i^\top \boldsymbol{x}). \tag{20}$$

The parameters are drawn from independent Gaussian priors, which corresponds to choosing a specific prior in a Bayesian model. The initializations are given by:

- Weights: $w_{ij} \sim \mathcal{N}(0, g_w^2)$ with $g_w^2 = \frac{\sigma_w}{d}$. This implies a prior probability $p(\boldsymbol{w}_i) \propto \exp(-\frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2)$.
- Amplitudes: $a_i \sim \mathcal{N}(0, g_a^2)$ with $g_a^2 = \sigma_a^2$. This implies a prior probability $p(a_i) \propto \exp(-\frac{1}{2g_a^2}\|a_i\|^2)$.

Let $\boldsymbol{\theta}_i = (\boldsymbol{w}_i, a_i)$ denote the parameters for the i-th neuron.

Stochastic Gradient Langevin Dynamics We consider a full-batch Gradient Descent (GD) update for a parameter set θ_i, which includes a weight decay term with coefficient γ and injected isotropic Gaussian noise $\xi_t \sim \mathcal{N}(0, I)$. This algorithm is known as Stochastic Gradient Langevin Dynamics (SGLD):

$$\Delta\theta_{i,t} := \theta_{i,t+1} - \theta_{i,t} \tag{21}$$

$$= -\eta\big(\gamma\, \theta_{i,t} + \nabla_{\theta_i}\ell(f_{\boldsymbol{\theta}})\big) \; + \sqrt{2T\eta}\; \xi_t\,. \tag{22}$$

Here, η is the learning rate, T is a scalar temperature that controls the noise magnitude, and $\ell(f_{\boldsymbol{\theta}}) = \frac{1}{P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2$ is the mean squared error loss over the dataset of size P.

In the continuous-time limit ($\eta \to 0$), this discrete update equation corresponds to a Langevin stochastic differential equation. The stationary distribution of this process, reached as $t \to \infty$, is given by the Gibbs-Boltzmann distribution:

$$p(\boldsymbol{\theta}_i) \propto \exp\Big(-\frac{1}{T}\Big(\frac{\gamma}{2}\|\theta_i\|^2 + \ell(f_{\boldsymbol{\theta}})\Big)\Big) \tag{23}$$

We can separate the weight decay terms γ into two separate parameters for weights and amplitudes, γ_w and γ_a, respectively. The stationary distribution for the parameters of a single neuron $(\boldsymbol{w}_i, a_i)$ is then:

$$p(\boldsymbol{w}_i, a_i) \propto \exp\left(-\frac{1}{T}\left(\frac{\gamma_w}{2}\|\boldsymbol{w}_i\|^2 + \frac{\gamma_a}{2}\|a_i\|^2 + \frac{1}{P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\right)\right). \tag{24}$$

Bayesian Posterior Distribution From a Bayesian perspective, we aim to find the posterior distribution of the parameters given the data $\mathcal{D} = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^{P}$. The posterior is given by Bayes' theorem: $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

- The prior $p(\boldsymbol{\theta})$ is defined by our choice of initialization: $p(\boldsymbol{\theta}) = \prod_i p(\boldsymbol{w}_i)p(a_i) \propto \exp\left(-\sum_i\left(\frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2 + \frac{1}{2g_a^2}\|a_i\|^2\right)\right)$.
- The likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ is chosen to be a Gaussian distribution with variance κ^2, corresponding to the mean squared error loss: $p(\mathcal{D}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\right)$.

Combining these, the log-posterior is proportional to the negative of an energy function $\mathcal{E}(\boldsymbol{\theta}, \mathcal{D})$. The posterior distribution for a single neuron's parameters is:

$$p(\boldsymbol{w}_i, a_i|\mathcal{D}) \propto \exp\left(-\left(\frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2 + \frac{1}{2g_a^2}\|a_i\|^2 + \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\right)\right). \tag{25}$$

Identifying the Distributions To establish the equivalence, we equate the functional forms of the SGLD stationary distribution and the Bayesian posterior. By comparing the exponents term-by-term, we find the following correspondences:

- **Weights:** $\frac{\gamma_w}{2T}\|\boldsymbol{w}_i\|^2 = \frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2 \implies \frac{\gamma_w}{T} = \frac{1}{g_w^2}$
- **Amplitudes:** $\frac{\gamma_a}{2T}\|a_i\|^2 = \frac{1}{2g_a^2}\|a_i\|^2 \implies \frac{\gamma_a}{T} = \frac{1}{g_a^2}$
- **Loss Term:** $\frac{1}{TP}\sum_\mu \ell_\mu = \frac{1}{2\kappa^2 P}\sum_\mu \ell_\mu \implies T = 2\kappa^2$

This implies that the SGLD algorithm with temperature T effectively samples from a Bayesian posterior with data noise variance $\kappa^2 = T/2$, and with prior variances $g_w^2 = T/\gamma_w$ and $g_a^2 = T/\gamma_a$.

Final Update Equations By substituting these relations back into the SGLD update rules, we obtain the dynamics for sampling from the desired posterior:

$$\Delta a_{t,i} = -\eta\left(\frac{T}{g_a^2}a_{t,i} + \nabla_{a_i}\left(\frac{1}{P}\sum_\mu \ell_\mu\right)\right) + \sqrt{2T\eta}\;\xi_{t,a} \tag{26}$$

$$\Delta \boldsymbol{w}_{t,i} = -\eta\left(\frac{T}{g_w^2}\boldsymbol{w}_{t,i} + \nabla_{\boldsymbol{w}_i}\left(\frac{1}{P}\sum_\mu \ell_\mu\right)\right) + \sqrt{2T\eta}\;\xi_{t,w}. \tag{27}$$

For training, this means we must set the temperature to $T = 2\kappa^2$.

A.2. Derivation of (4)

We start from the negative log posterior and the two–layer model in Eq. (3):

$$-\ln p_{\mathrm{GD}}(\boldsymbol{\theta} \mid \mathcal{D}) = \underbrace{\sum_l \frac{1}{2\sigma_l^2}\sum_j \|\boldsymbol{\theta}_{l,j}\|^2}_{-\ln p_{\text{prior}}} + \underbrace{\frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_\mu) - y_\mu\big)^2}_{-\ln p_{\mathrm{L}}}.$$

For our two–layer network, $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{N^\gamma}\sum_{i=1}^N a_i\,\phi(\boldsymbol{w}_i^\top \boldsymbol{x})$. Define the shorthand $\phi_{i\mu} := \phi(\boldsymbol{w}_i^\top \boldsymbol{x}_\mu)$. Then the data term expands as

$$\begin{aligned} -\ln p_{\mathrm{L}} &= \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}\left(\frac{1}{N^\gamma}\sum_{i=1}^N a_i\phi_{i\mu} - y_\mu\right)^2 \\ &= \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}\left[\frac{1}{N^{2\gamma}}\sum_{i=1}^N\sum_{j=1}^N a_i a_j\,\phi_{i\mu}\phi_{j\mu} - \frac{2}{N^\gamma}\sum_{i=1}^N a_i\,\phi_{i\mu}y_\mu + y_\mu^2\right]. \end{aligned} \tag{28}$$

Introduce the dataset–averaged quantities

$$\Sigma(\boldsymbol{w}) := \frac{1}{P}\sum_{\mu=1}^{P}\phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)^2, \qquad G(\boldsymbol{w},\boldsymbol{w}') := \frac{1}{P}\sum_{\mu=1}^{P}\phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)\,\phi(\boldsymbol{w}'^\top \boldsymbol{x}_\mu),$$

$$J_{\mathcal{Y}}(\boldsymbol{w}) := \frac{1}{P}\sum_{\mu=1}^{P}\phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)\,y_\mu. \tag{29}$$

Using $\sum_{i,j} = \sum_{i=j} + \sum_{i\neq j}$ in (28), we obtain

$$-\ln p_{\mathrm{L}} = \frac{1}{2\kappa^2 N^{2\gamma}} \left[\sum_{i=1}^{N} a_i^2 \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu}^2}_{\Sigma(\boldsymbol{w}_i)} + \sum_{i\neq j} a_i a_j \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu}\phi_{j\mu}}_{G(\boldsymbol{w}_i,\boldsymbol{w}_j)} \right] - \frac{1}{\kappa^2 N^{\gamma}} \sum_{i=1}^{N} a_i \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu} y_\mu}_{J_{\mathcal{Y}}(\boldsymbol{w}_i)} + \underbrace{\frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P} y_\mu^2}_{\text{const.}}. \tag{30}$$

The prior part for our parameterization $w_{ij} \sim \mathcal{N}(0, \sigma_w^2/d)$ and $a_i \sim \mathcal{N}(0, \sigma_a^2)$ is

$$-\ln p_{\text{prior}} = \frac{1}{2\sigma_a^2}\sum_{i=1}^{N} a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^{N} \|\boldsymbol{w}_i\|^2. \tag{31}$$

Combining (30) and (31), and discarding the $\boldsymbol{\theta}$-independent constant, yields Eq. (4):

$$-\ln\, p_{\mathrm{GD}}(\boldsymbol{W}, \boldsymbol{a} \mid \mathcal{D}) = \frac{1}{2\sigma_a^2}\sum_{i=1}^{N} a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^{N} \|\boldsymbol{w}_i\|^2 + \frac{1}{2\kappa^2 N^{2\gamma}}\sum_{i=1}^{N} a_i^2\, \Sigma(\boldsymbol{w}_i) + \frac{1}{2\kappa^2 N^{2\gamma}}\sum_{i\neq j} a_i a_j\, G(\boldsymbol{w}_i, \boldsymbol{w}_j) - \frac{1}{\kappa^2 N^{\gamma}}\sum_{i=1}^{N} a_i\, J_{\mathcal{Y}}(\boldsymbol{w}_i) + \text{const.}$$

This derivation holds for any nonlinearity ϕ.

A.3. k-sparse parity target function

A k-sparse parity target function teacher is a single Walsh basis function on the Boolean hypercube. Let $S \subseteq [d]$ with $|S| = k$. The Walsh function indexed by S is

$$\chi_S(\boldsymbol{x}) \;=\; \prod_{j\in S} x_j, \qquad \boldsymbol{x} \in \{\pm 1\}^d,$$

(and, if $\boldsymbol{x} \in [-1,1]^d$, one may use $\chi_S(\boldsymbol{x}) = \prod_{j\in S} \mathrm{sign}(x_j)$). The family $\{\chi_S\}_{S\subseteq[d]}$ forms an orthonormal basis under the uniform product measure, i.e. $\mathbb{E}[\chi_S(\boldsymbol{x})\chi_T(\boldsymbol{x})] = \mathbb{1}\{S = T\}$. In our experiments the teacher is $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$, when $|S| = k$ this is exactly the k-parity problem.

A.4. Relation to kernel eigenvalue outliers

The transition from $m_S = 0$ to $m_S > 0$ manifests as an outlier eigenvalue in the learned kernel. Recall that $m_S = N^{1-\gamma}\langle a J_S(\boldsymbol{w})\rangle$ measures the mean alignment between neurons and the target mode, where $J_S(\boldsymbol{w}) = \mathbb{E}[\phi(\boldsymbol{w}^\top \boldsymbol{x})\chi_S(\boldsymbol{x})]$ quantifies how well a single neuron with weights $\boldsymbol{w}$ correlates with χ_S. When m_S becomes non-zero, it indicates that neurons have collectively aligned their weights to capture the target structure. Their $J_S(\boldsymbol{w}_i)$ values have grown large. This alignment directly impacts the empirical kernel through

$$\widehat{R}_S^{(a)} = \frac{\chi_S^\top K \chi_S}{\mathrm{tr}(K)} = \frac{\sum_{i=1}^{N} a_i^2\, J_S(\boldsymbol{w}_i)^2}{\sum_{i=1}^{N} a_i^2\, \Sigma(\boldsymbol{w}_i)}, \tag{32}$$

where $K_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{w}_i^\top \boldsymbol{x}_\mu)\, \phi(\boldsymbol{w}_i^\top \boldsymbol{x}_\nu)$ is the learned kernel. The numerator $\sum_i a_i^2 J_S(\boldsymbol{w}_i)^2$ grows quadratically with the alignment strengths $J_S(\boldsymbol{w}_i)$, while the denominator (trace) remains roughly constant. In the kernel regime ($m_S = 0$), neurons remain randomly oriented so $J_S(\boldsymbol{w}_i) \sim O(N^{-1/2})$ and the ratio stays $O(1/P)$. However, when FL occurs ($m_S > 0$), the enhanced $J_S(\boldsymbol{w}_i)$ values cause $\chi_S^\top K \chi_S$ to grow substantially, making χ_S an outlying eigendirection of K. This anisotropic deformation, from the isotropic NNGP to a low-rank, plus isotropic structure, provides a direct spectral signature of the FL phase transition.

A.5. The NNGP limit

A.6. NNGP limit: No interactions

The MF model can be simplified even further in the infinite-width limit with $\gamma = 1/2$ (NTK-scaling). For this scaling the limit is well-behaved, the self-energy term in Equation (6) vanishes as well as any m_A-dependent tilt, resulting in $p(\boldsymbol{w})$ collapsing to its prior (see Section D.1 for a *proof*).

NNGP limit (fixed point equations)

$$\begin{aligned} -\ln p_\infty &= \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \\ m_A^\infty &= \frac{\sigma_a^2}{\kappa^2}\big(\langle J_{\mathcal{Y}}(\boldsymbol{w})J_A(\boldsymbol{w})\rangle_{p_\infty} \\ &\quad -\sum_B m_B^\infty \langle J_B(\boldsymbol{w})J_A(\boldsymbol{w})\rangle_{p_\infty}\big) \end{aligned} \tag{33}$$

where $p_\infty \to p_\infty(\boldsymbol{w})$, $\boldsymbol{w} \in \mathbb{R}^d$.

Kernel picture We can define the kernel: $K_{AB} := \mathbb{E}_{\boldsymbol{w}\sim p_\infty}\left[J_A(\boldsymbol{w})\, J_B(\boldsymbol{w})\right]$ reducing the equations above to kernel ridge regression:

$$\begin{aligned} \boldsymbol{m}^\infty &= K\,(K+\tau\mathbb{1})^{-1}\boldsymbol{y}, \text{ where} \\ \tau &= \kappa^2/\sigma_a^2, \quad y_A = \mathbb{E}_{\boldsymbol{x}}[y(\boldsymbol{x})\chi_A(\boldsymbol{x})]. \end{aligned} \tag{34}$$

This NNGP limit represents the most restrictive limit in our approximation hierarchy. While MF theory eliminates inter-neuron interactions ($w_{ij} \leftrightarrow w_{i'j}$), the NNGP limit additionally removes intra-neuron coordinate coupling ($w_{ij} \leftrightarrow w_{ij'}$), see **??** for an illustration. Every weight is sampled from the same distribution (the prior).

No FL mechanism In the infinite-width limit, there is *no* FL in the sense above. With no data-dependent term ($J_{\mathcal{Y}}$) to break symmetry, all neurons remain frozen at their prior. The feature coefficients m_A^∞ are now determined purely by the fixed kernel $K_{AB} = \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_B(\boldsymbol{w})]$ (see Section A.7 how this connects to the usual kernel formulation of the NNGP limit).

A.7. Connection to the standard NNGP formulation

The kernel representation in the function basis $\{\chi_A\}$ can be transformed to recover the standard NNGP formulation in input space. We start with the function-basis kernel

$$K_{AB} = \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_B(\boldsymbol{w})], \tag{35}$$

where $J_A(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}[\phi(\boldsymbol{w}^\top\boldsymbol{x})\chi_A(\boldsymbol{x})]$ projects the neuron's output onto basis function χ_A. Expanding this definition:

$$K_{AB} = \mathbb{E}_{\boldsymbol{w}}\left[\mathbb{E}_{\boldsymbol{x}}[\phi(\boldsymbol{w}^\top\boldsymbol{x})\chi_A(\boldsymbol{x})]\cdot\mathbb{E}_{\boldsymbol{x}'}[\phi(\boldsymbol{w}^\top\boldsymbol{x}')\chi_B(\boldsymbol{x}')]\right] \tag{36}$$

$$= \mathbb{E}_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\phi(\boldsymbol{w}^\top\boldsymbol{x})\phi(\boldsymbol{w}^\top\boldsymbol{x}')\chi_A(\boldsymbol{x})\chi_B(\boldsymbol{x}')\right] \tag{37}$$

$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\chi_A(\boldsymbol{x})\chi_B(\boldsymbol{x}')\cdot\underbrace{\sigma_a^2\mathbb{E}_{\boldsymbol{w}}[\phi(\boldsymbol{w}^\top\boldsymbol{x})\phi(\boldsymbol{w}^\top\boldsymbol{x}')]}_{=:K(\boldsymbol{x},\boldsymbol{x}')}\right], \tag{38}$$

where $K(\boldsymbol{x},\boldsymbol{x}')$ is the standard NNGP kernel in input space. The fixed-point equation $m_A = \frac{\sigma_a^2}{\kappa^2}(\Xi_A - \sum_B K_{AB}m_B)$ with $\Xi_A = \sum_S y_S K_{AS}$ becomes

$$m = K(K+\tau I)^{-1}y, \qquad \tau = \kappa^2/\sigma_a^2. \tag{39}$$

To see this explicitly, consider data points $\{\boldsymbol{x}_\mu\}_{\mu=1}^P$ with labels y_μ. The predictor in the function basis is $f(\boldsymbol{x}) = \sum_A m_A\chi_A(\boldsymbol{x})$, while in the input basis it becomes $f(\boldsymbol{x}_\mu) = \sum_{\nu=1}^P \alpha_\nu K(\boldsymbol{x}_\mu,\boldsymbol{x}_\nu)$ where $\alpha = (K+\tau I)^{-1}y$. The equivalence follows from the change of basis: if $\chi_A(\boldsymbol{x}) = \delta_{\boldsymbol{x},\boldsymbol{x}_A}$ (point evaluation basis), then $K_{AB} = K(\boldsymbol{x}_A,\boldsymbol{x}_B)$ directly recovers the Gram matrix. For general orthogonal bases, the kernel ridge regression solution remains invariant under this transformation.

A.8. Derivation of the ARD precision fixed point

Recall the ARD prior on per-coordinate precisions $\boldsymbol{\rho} = (\rho_1, \dots, \rho_d)$ and weights:

$$p(\boldsymbol{w} \mid \boldsymbol{\rho}) = \prod_{j=1}^{d} \mathcal{N}\big(w_j \mid 0, \rho_j^{-1}\big), \qquad p(\boldsymbol{\rho}) = \prod_{j=1}^{d} \Gamma(\rho_j \mid \alpha_0, \beta_0), \tag{40}$$

and the single-neuron MF–ARD action (Equation (13) in the main text)

$$-\ln p_{\mathrm{ARD}}(\boldsymbol{w}, a \mid \{m_A\}, \boldsymbol{\rho}, \mathcal{D}) = \frac{a^2}{2\sigma_a^2} + \frac{1}{2}\sum_{j=1}^{d} \rho_j\, w_j^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}}\Sigma(\boldsymbol{w}) - \frac{a}{\kappa^2 N^{\gamma}}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w})\right). \tag{41}$$

For a width-N two-layer network, the joint action is the sum over neurons $i = 1, \dots, N$ of Equation (41), and the (negative) log-evidence (free energy) for fixed $\{m_A\}$ is

$$\mathcal{F}(\boldsymbol{\rho}; \{m_A\}) = -\ln Z(\boldsymbol{\rho}; \{m_A\}) - \ln p(\boldsymbol{\rho}) \quad \text{with} \quad Z(\boldsymbol{\rho}; \{m_A\}) = \int \prod_{i=1}^{N} d\boldsymbol{w}_i\, da_i\; e^{-\sum_{i=1}^{N} S_{\mathrm{ARD}}(\boldsymbol{w}_i, a_i)}. \tag{42}$$

Stationarity of the negative log-evidence w.r.t. ρ_j gives the ARD FP:

$$0 = \frac{\partial \mathcal{F}}{\partial \rho_j} = \underbrace{\left\langle \frac{\partial}{\partial \rho_j} \sum_{i=1}^{N} S_{\mathrm{ARD}}(\boldsymbol{w}_i, a_i) \right\rangle_{p_{\mathrm{ARD}}}}_{\text{energy term}} \underbrace{- \frac{N}{2}\frac{1}{\rho_j}}_{\text{Gaussian normalizer}} \underbrace{- \frac{\partial \ln p(\boldsymbol{\rho})}{\partial \rho_j}}_{\text{prior term}} = \frac{1}{2}\sum_{i=1}^{N} \left\langle w_{ij}^2 \right\rangle_{p_{\mathrm{ARD}}} - \frac{N}{2}\frac{1}{\rho_j} - \frac{\alpha_0 - 1}{\rho_j} + \beta_0, \tag{43}$$

where we used $\partial S_{\mathrm{ARD}}/\partial \rho_j = \frac{1}{2}\sum_i w_{ij}^2$ and $\partial \ln \Gamma(\rho_j \mid \alpha_0, \beta_0)/\partial \rho_j = (\alpha_0 - 1)/\rho_j - \beta_0$. By MF symmetry, all neurons are i.i.d. under p_{ARD}, so $\sum_{i=1}^{N} \langle w_{ij}^2 \rangle_{p_{\mathrm{ARD}}} = N \langle w_j^2 \rangle_{p_{\mathrm{ARD}}}$, and Equation (43) yields the closed-form FP update

$$\rho_j^\star \;=\; \frac{\alpha_0 - 1 + \frac{N}{2}}{\beta_0 + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}} \;\overset{\text{MAP corr.}}{\approx}\; \frac{\alpha_0 + \frac{N}{2}}{\beta_0 + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}}, \tag{44}$$

where the MAP conjugacy correction (absorbing the -1 in α_0) gives the form used in the main text (Equation (14)). With the scale-matching choice $\beta_0 = \alpha_0/d$, Equation (44) becomes

$$\rho_j^\star \;=\; \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}}. \tag{45}$$

A.9. NNGP limit of the MF-ARD model

NNGP-limit of MF-ARD

$$\mathcal{S}_\infty^{\mathrm{FL}}(\mathbf{w}|\rho) = \frac{1}{2}\sum_{j=1}^{d} \rho_j w_j^2, \qquad p_{\infty,\rho}(\mathbf{w}) = \frac{1}{\mathcal{Z}}\exp\big(-\mathcal{S}_\infty^{\mathrm{FL}}(\mathbf{w}|\rho)\big), \tag{46}$$

$$m_A^\infty(\rho) = \frac{\sigma_a^2}{\kappa^2}\left\langle \big(J_{\mathcal{Y}}(\mathbf{w}) - \sum_B m_B^\infty(\rho) J_B(\mathbf{w})\big) J_A(\mathbf{w}) \right\rangle_{p_{\infty,\rho}}, \quad \forall A. \tag{47}$$

$$0 = \frac{1}{2}\mathrm{tr}\big[(A - (Ay)(Ay)^\top)\partial_{\rho_j} K_\rho\big] + \beta_0 - \frac{\alpha_0 - 1}{\rho_j}, \qquad A = (K_\rho + \tau I)^{-1} \tag{48}$$

Kernel picture We can define the kernel : $K_{\rho,AB} = \mathbb{E}_{\mathbf{w}\sim p_{\infty,\rho}}[J_A(\mathbf{w}) J_B(\mathbf{w})]$ reducing the solution above to KRR

$$m^\infty(\rho) = K_\rho (K_\rho + \tau I)^{-1} y, \quad \tau = \kappa^2/\sigma_a^2 \tag{49}$$

Usual (kernel) form Let $C_\rho = \mathrm{diag}(\rho)^{-1}$ and $K_\rho(\mathbf{x}, \mathbf{x}') = \sigma_a^2 \, \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,C_\rho)}[\phi(\mathbf{w}^\top \mathbf{x})\phi(\mathbf{w}^\top \mathbf{x}')]$. Then

$$m^\infty(\rho) = K_\rho(K_\rho + \tau I)^{-1} y, \quad K_\rho \chi_A = \lambda_A(\rho)\chi_A \;\Rightarrow\; m_A^\infty(\rho) = \frac{\lambda_A(\rho)}{\lambda_A(\rho) + \tau}\, y_A. \tag{50}$$

This shows that ρ rescales coordinates before the nonlinearity, so $K_\rho = \mathbb{E}_{\mathbf{z}}[J_A(C_\rho^{1/2}\mathbf{z})\, J_B(C_\rho^{1/2}\mathbf{z})]$ is a nonlinear deformation of the isotropic kernel. Any nontrivial change in ρ therefore produces an $O(1)$ change in the Gaussian measure over $\mathbf{w}$, hence an $O(1)$ change in K (it does not vanish with width). ARD improves spectral alignment by increasing $\lambda_A(\rho)$ along task-aligned directions. When $\lambda_A(\rho)$ crosses the scale τ, m_A^∞ exhibits a jump, yielding the leading-order FL effect at infinite width.

A.10. Details on transformers

Let the input sequence be $(x_0, x_1, \ldots, x_{T-1})$, where (x_0, x_1) are the two addends, $(x_2, \ldots, x_{T-2})$ are distractors, and x_{T-1} is a CLS token used for classification. For each layer ℓ and head h, let $\alpha^{(\ell,h)}_{\mathrm{CLS}\to j}(t)$ be the attention probability from the CLS query (position $T-1$) to key position j at training epoch t. Define

$$M_{\mathrm{attn}}(t) = \mathbb{E}_{\ell,h}\left[\alpha^{(\ell,h)}_{\mathrm{CLS}\to 0}(t) + \alpha^{(\ell,h)}_{\mathrm{CLS}\to 1}(t)\right] \in [0,1].$$

Let $e_j(t) \in \mathbb{R}^{d_{\mathrm{model}}}$ denote the embedding vector at position j, and define $g_j(t) = \mathbb{E}\left[|\nabla_{e_j} L(t)|\right]$ (expectation over a fixed probe batch or the data distribution). Then

$$M_{\mathrm{grad}}(t) = \frac{g_0(t) + g_1(t)}{\sum_{j=0}^{T-1} g_j(t)} \in [0,1].$$

In the experiments, expectations are approximated by averaging over a probe batch and over heads/layers (for M_{attn}).

A.11. Universality of IFS

Our IFS mechanism is largely agnostic to the specific choice of nonlinearity, provided the nonlinearity allows for feature learning. As shown in Figure 8 networks trained with Sigmoid activations exhibit the same coordinate-wise heavy-tailed distributions on relevant input dimensions as ReLU networks. While the population-level organization (sparse vs. redundant neurons) may differ as predicted by (van Meegen & Sompolinsky, 2025), the fundamental mechanism of discovering the relevant input subspace via IFS remains the same.

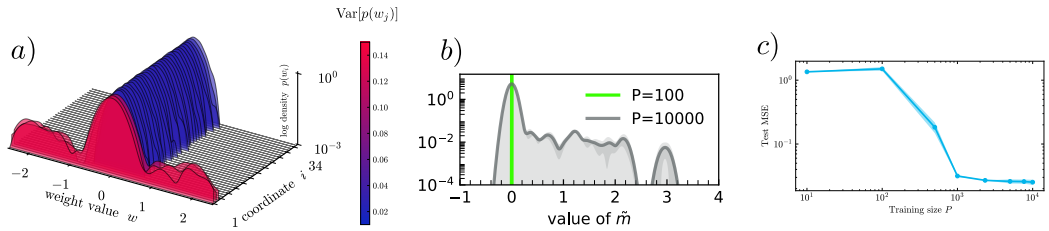

Figure 8. Similar to the ReLU results in the main text, SGLD training with Sigmoid activations leads to strong coordinate-wise symmetry breaking (high variance on relevant features $j \in S$). This demonstrates that IFS is a fundamental mechanism for subspace discovery, distinct from the neuron-level coding schemes discussed in (van Meegen & Sompolinsky, 2025).

B. Additional figures

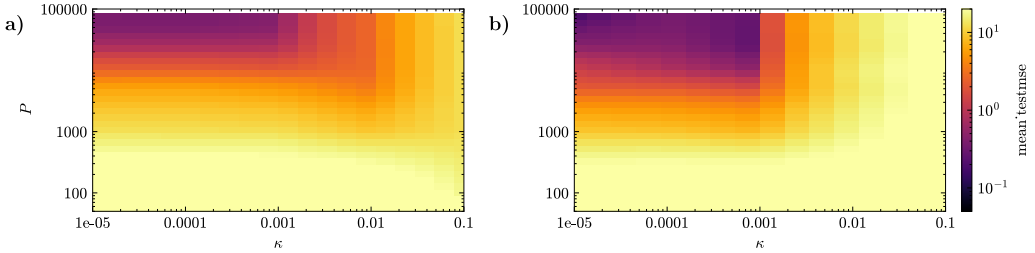

Figure 9. Same plot as Figure 6 **a)** SGLD, **b)** MF-ARD, but for a single-index model with Gaussian inputs $\boldsymbol{x}! \sim! \mathcal{N}(0, \mathbb{1})$ and teacher $y = \mathrm{He}_p(\boldsymbol{w}^\top \boldsymbol{x})$, where $\boldsymbol{w}_j = 1/\sqrt{k}$ on a k-sized support and 0 otherwise, here $d = 18$, $k = 2$, $p = 4$.

C. Comparison with recent mean field works

In this section, we clarify the relationship and fundamental differences between our MF-ARD framework and the recent work by van Meegen & Sompolinsky (2025). While both works utilize mean field approximations to study the posterior of neural networks, they address fundamentally different phenomena occurring at different levels of the network hierarchy.

C.1. Coordinate-wise vs. Neuron-wise Sparsification

The most critical distinction lies in the domain where symmetry breaking occurs:

- **(Neuron-wise Sparsification):** (van Meegen & Sompolinsky, 2025) investigates how the population of N neurons organizes to represent the target. They find that the posterior distribution over the weight vectors $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N\}$ breaks permutation symmetry, resulting in a solution where only a subset of neurons ($O(1)$ out of N) strongly correlate with the target, while others remain agnostic or redundant. This is a textitinter-neuron phenomenon.
- **Our Work (Coordinate-wise Sparsification / IFS):** Our theory models the posterior of a *single representative neuron* $\boldsymbol{w} = [w_1, \ldots, w_d] \in \mathbb{R}^d$. We investigate how the symmetry between input coordinates $j = 1, \ldots, d$ is broken. We find that for tasks with isotropic data but low-dimensional targets, the posterior marginals $p(w_j)$ become heavy-tailed and high-variance only for task-relevant coordinates ($j \in S$), while remaining Gaussian and narrow for irrelevant coordinates ($j \notin S$). This is an *intra-neuron* phenomenon.

Standard mean field theories (including the base model in (van Meegen & Sompolinsky, 2025)) typically assume that the weight distribution is isotropic over input coordinates unless the prior enforces otherwise. Consequently, while (van Meegen & Sompolinsky, 2025) successfully predicts that some neurons become "active" and others "dormant," it does not explicitly model the self-reinforcing mechanism that allows an *active* neuron to selectively amplify specific input coordinates to overcome the curse of dimensionality.

C.2. The Necessity of Coordinate Anisotropy: A Control Experiment

To rigorously prove that coordinate-wise symmetry breaking (the core of our MF-ARD theory) is the causal mechanism for beating the kernel regime, we performed a control experiment enforcing coordinate homogeneity.

We trained networks on the k-sparse parity task using a modified loss function that penalizes anisotropy within neurons, without penalizing sparsity across neurons:

$$\ell_{\text{total}} = \ell_{\text{MSE}} + \lambda_h \sum_{j=1}^{N} \left[\frac{1}{d} \sum_{i=1}^{d} \left(|w_{ij}| - \left(\frac{1}{d} \sum_{k=1}^{d} |w_{kj}| \right) \right)^2 \right]. \tag{51}$$

Result: As λ_h increases, the learning curve of the neural network reverts to the NNGP learning curve (see Figure 4). Even

if the network is allowed to sparsify at the neuron level (some neurons large, some small), suppressing the variance *between coordinates* w_{ij} prevents the network from learning the target efficiently.

This confirms that the performance gap between Finite-Width NNs and Kernels on isotropic data is driven by the mechanism modeled in this paper (IFS/Coordinate Anisotropy), a mechanism that is distinct from and complementary to the population-level phenomena described in (van Meegen & Sompolinsky, 2025).

D. Proofs

D.1. Kernel limit

Theorem D.1. *With $\gamma = 1/2$ the infinite width limit has the following FP equations:*

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2}\left(\mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_{\mathcal{Y}}(\boldsymbol{w})] - \sum_B \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_B(\boldsymbol{w})]m_B^\infty\right). \tag{52}$$

Proof. The proof proceeds by taking the $N \to \infty$ limit of the self-consistency equations derived from the cavity method's free energy.

1. Self-Consistency from Free Energy. We begin with the MF free energy functional derived from the cavity method:

$$\mathcal{F}(\{m_A\}) = \frac{1}{2\kappa^2}\sum_A (y_A - m_A)^2 - N \ln \mathcal{Z}_1(\{m_A\}) \tag{53}$$

where $\mathcal{Z}_1(\{m_A\}) = \int d\boldsymbol{w}\, e^{-\bar{S}_{\text{eff}}^{(\infty)}(\boldsymbol{w};\{m_A\})}$ is the single-neuron partition function. The effective action for a single weight vector $\boldsymbol{w}$, after integrating out the amplitude a, is:

$$\bar{S}_{\text{eff}}^{(\infty)}(\boldsymbol{w};\{m_A\}) = \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\ln\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right) - \frac{\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w})\right)^2}{2\kappa^4 N^{2\gamma}\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right)} \tag{54}$$

The equilibrium state is found by the stationarity condition $\partial\mathcal{F}/\partial m_A = 0$, which yields the self-consistency equation:

$$m_A = N^{1-\gamma}\left\langle \mu(\boldsymbol{w})J_A(\boldsymbol{w})\right\rangle_{p(\boldsymbol{w}|\{m_B\})} \tag{55}$$

where $p(\boldsymbol{w}|\{m_B\}) = \frac{1}{\mathcal{Z}_1}e^{-\bar{S}_{\text{eff}}^{(\infty)}}$ is the posterior distribution on a single neuron's weights, and $\mu(\boldsymbol{w})$ is the posterior mean of its amplitude a:

$$\mu(\boldsymbol{w}) = \frac{\frac{1}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B J_B(\boldsymbol{w})\right)}{\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}} \tag{56}$$

2. The Infinite-Width Limit with Critical Scaling. To obtain a non-trivial limit as $N \to \infty$, we must set the scaling to the critical value $\gamma = 1/2$. For $\gamma > 1/2$, the prefactor $N^{1-\gamma} \to 0$, decoupling the neurons and leading to $m_A \to 0$. For $\gamma < 1/2$, the interaction term diverges. With $\gamma = 1/2$, the terms of order $1/N$ in the effective action $\bar{S}_{\text{eff}}^{(\infty)}$ vanish. Specifically:

$$\frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} = \frac{\Sigma(\boldsymbol{w})}{N\kappa^2} \xrightarrow{N\to\infty} 0 \tag{57}$$

As a result, the data-dependent and m_A-dependent terms in the exponent of $p(\boldsymbol{w}|\{m_A\})$ vanish. The distribution over weights collapses to its prior:

$$p(\boldsymbol{w}|\{m_A\}) \xrightarrow{N\to\infty} p_\infty(\boldsymbol{w}) \propto \exp\left(-\frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2\right) \tag{58}$$

Simultaneously, the denominator in $\mu(\boldsymbol{w})$ simplifies to $1/\sigma_a^2$. The expression for the posterior mean amplitude becomes:

$$\mu(\boldsymbol{w}) \xrightarrow{N\to\infty} \frac{\sigma_a^2}{\kappa^2 N^{1/2}}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B J_B(\boldsymbol{w})\right) \tag{59}$$

3. Deriving the Linear System. We now substitute these limiting forms back into the self-consistency equation, using m_A^∞ to denote the solution in this limit:

$$m_A^\infty = N^{1-1/2} \left\langle \left[\frac{\sigma_a^2}{\kappa^2 N^{1/2}} \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B^\infty J_B(\boldsymbol{w}) \right) \right] J_A(\boldsymbol{w}) \right\rangle_{p_\infty(\boldsymbol{w})} \tag{60}$$

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2} \left\langle \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B^\infty J_B(\boldsymbol{w}) \right) J_A(\boldsymbol{w}) \right\rangle_{p_\infty(\boldsymbol{w})} \tag{61}$$

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2} \left(\mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_{\mathcal{Y}}(\boldsymbol{w})] - \sum_B \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_B(\boldsymbol{w})] m_B^\infty \right) \tag{62}$$

Let us define the NNGP kernel matrix K with elements $K_{AB} := \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_B(\boldsymbol{w})]$ and the data-kernel coupling vector Ξ with elements $\Xi_A := \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_{\mathcal{Y}}(\boldsymbol{w})] = \sum_S y_S K_{AS}$. The equation becomes a linear system for the vector m^∞:

$$m^\infty = \frac{\sigma_a^2}{\kappa^2}(\Xi - K m^\infty) = \frac{\sigma_a^2}{\kappa^2}(Ky - K m^\infty) \tag{63}$$

4. Solution as Kernel Ridge Regression. Rearranging the linear system, we get:

$$\left(I + \frac{\sigma_a^2}{\kappa^2} K \right) m^\infty = \frac{\sigma_a^2}{\kappa^2} K y \tag{64}$$

$$\left(\frac{\kappa^2}{2\sigma_a^2} I + \frac{1}{2} K \right) m^\infty = \frac{1}{2} K y \tag{65}$$

Let the ridge be $\tau = \kappa^2/(2\sigma_a^2)$. The equation is $(\tau I + K) m^\infty = Ky$. Solving for m^∞ gives the final KRR solution:

$$m^\infty = K(K + \tau I)^{-1} y \tag{66}$$

This completes the proof. The solution depends only on the NNGP kernel K, which is fixed by the network architecture and priors, not the training data labels. This demonstrates the absence of FL in this specific infinite-width limit. □

D.2. Integrating out a

As the action is quadratic in a we can integrate it out.

Theorem D.2. *The distribution is given by*

$$S_{\text{MF}}(\boldsymbol{w}, \{m_A\}) = const. + \frac{1}{2} \ln \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right) - \frac{\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w}) \right)^2}{2N^{2\gamma}\kappa^4 \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right)} \tag{67}$$

$$+ \frac{d}{2\sigma_w^2} \|\boldsymbol{w}\|^2 \tag{68}$$

$$p(a|\boldsymbol{w}) = \mathcal{N}(\mu(\boldsymbol{w}), \sigma^2(\boldsymbol{w})) \tag{69}$$

$$\sigma(\boldsymbol{w})^2 = \frac{1}{2\alpha} = \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right)^{-1} \tag{70}$$

$$\mu(\boldsymbol{w}) = \sigma^2 \beta = \frac{\beta}{2\alpha} = \frac{\frac{1}{N^\gamma \kappa^2} \left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w}) \right)}{\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right)} \tag{71}$$

Proof. We start with the standard Gaussian integral

$$\int da d\boldsymbol{w} e^{-[\alpha \cdot a^2 - \beta \cdot a + c]} = \int d\boldsymbol{w} \sqrt{\frac{\pi}{\alpha}} e^{\frac{\beta^2}{4\alpha} - c} \tag{72}$$

We have

$$\mathcal{S}_{\text{MF}}(\boldsymbol{w}, a, \{m_A\}) = \frac{1}{2\sigma_a^2}\sum_{i=1}^{N} a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^{d}\|\boldsymbol{w}_i\|^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}}\Sigma(\boldsymbol{w}) \tag{73}$$

$$- \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w})\right) \tag{74}$$

For 1 neuron we get

$$\mathcal{S}_{\text{MF}}(\boldsymbol{w}, a, \{m_A\}) = a^2\left(\frac{1}{2\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{2\kappa^2 N^{2\gamma}}\right) - \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w})\right) + \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \tag{75}$$

with

$$\alpha := \frac{1}{2\sigma_a^2} + \frac{\Sigma(w)}{2N^{2\gamma}\kappa^2}, \tag{76}$$

$$\beta := \frac{J_{\mathcal{Y}}(w) - \sum_A m_A J_A(w)}{N^\gamma \kappa^2}. \tag{77}$$

$$c = \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \tag{78}$$

and $\alpha(a - \frac{\beta}{2\alpha})^2 - \frac{\beta^2}{4\alpha} + c$ comparing to $\frac{-(a-\mu)^2}{2\sigma^2}$ we get $\alpha = \frac{1}{2\sigma^2}$ and $\mu = \frac{\beta}{2\alpha}$ for the Gaussian identification. This gives

$$\int d\boldsymbol{w}\sqrt{\frac{\pi}{\alpha}}e^{\frac{\beta^2}{4\alpha} - c} = \int d\boldsymbol{w} e^{-[\frac{1}{2}\ln(\pi) - \frac{1}{2}\ln(\alpha) + \frac{\beta^2}{4\alpha} - c]} \tag{79}$$

where the exponent is identified as the effective action. □

D.3. MF- FP equation

Lemma D.3. *Consider a single target mode $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ and assume other overlaps vanish. Using the self–consistency $m_A = N^{1-\gamma}\langle a\, J_A(\boldsymbol{w})\rangle$ and the conditional mean $\mu(\boldsymbol{w})$ above, we obtain*

$$m_S \;=\; N^{1-\gamma}\left\langle \mu(\boldsymbol{w})\, J_S(\boldsymbol{w})\right\rangle \;=\; \frac{N^{1-2\gamma}}{\kappa^2}(1 - m_S)\left\langle \frac{J_S(\boldsymbol{w})^2}{\sigma_a^{-2} + \dfrac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}}\right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)}. \tag{80}$$

and in the $\kappa \to 0$ limit it is

$$m_S \;\approx\; (1 - m_S)\, N\left\langle \frac{J_S(\boldsymbol{w})^2}{\Sigma(\boldsymbol{w})}\right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)} \tag{81}$$

Proof. Consider a single target mode $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ and assume other overlaps vanish. Using the self–consistency $m_A = N^{1-\gamma}\langle a\, J_A(\boldsymbol{w})\rangle$ and the conditional mean $\mu(\boldsymbol{w})$ above, we obtain

$$m_S \;=\; N^{1-\gamma}\left\langle \mu(\boldsymbol{w})\, J_S(\boldsymbol{w})\right\rangle \;=\; \frac{N^{1-2\gamma}}{\kappa^2}(1 - m_S)\left\langle \frac{J_S(\boldsymbol{w})^2}{\sigma_a^{-2} + \dfrac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}}\right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)}.$$

In the small–noise regime $\kappa \to 0$ (with $\Sigma(\boldsymbol{w}) > 0$), the denominator simplifies as $\sigma_a^{-2} + \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}} \approx \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}$, so that

$$m_S \;\approx\; (1 - m_S)\, N\left\langle \frac{J_S(\boldsymbol{w})^2}{\Sigma(\boldsymbol{w})}\right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)} \quad\Longrightarrow\quad m_S \;\approx\; \frac{N\Big\langle J_S(\boldsymbol{w})^2/\Sigma(\boldsymbol{w})\Big\rangle}{1 + N\Big\langle J_S(\boldsymbol{w})^2/\Sigma(\boldsymbol{w})\Big\rangle}$$

By Cauchy–Schwarz, $J_S(\boldsymbol{w})^2 \le \Sigma(\boldsymbol{w})$, so $0 \le m_S < 1$ unless all mass concentrates on perfectly aligned $\boldsymbol{w}$. □

D.4. Solution to the fixed point equation

Lemma D.4. *Consider* $\phi = \text{ReLU}$. *Let the inputs* $x_j \in \{\pm 1\}$ *be i.i.d. and* $y(\boldsymbol{x}) = \chi_S(\mathbf{x}) = \prod_{j\in S} x_j$ *with* $S = \{0, 1, ..., k-1\}$. *We define* $R_k(\boldsymbol{w}) = \frac{J_S(\boldsymbol{w})^2}{\Sigma(\boldsymbol{w})}$. *It holds that*

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} R_k(\boldsymbol{w}) \tag{82}$$

is given by $\boldsymbol{w}^* = (\overbrace{\alpha, \ldots, \alpha}^{k}, 0, ..., 0)$.

Proof. Decompose $w_S = \alpha \frac{\mathbb{1}_S}{\sqrt{k}} + u, \quad u \perp \mathbb{1}_S$ and keep arbitrary $\boldsymbol{w}^C$. Because the data distribution is invariant to permutations of the k coordinates in S the only S-dependent statistic that survives in χ_S-weighted expectations is the sum $s(\boldsymbol{x}) = \sum_{j\in S} x_j$. Any component orthogonal to $\mathbb{1}_S$ averages out inJ_S but increases $\Sigma(\boldsymbol{w})$ (by convexity of $x \mapsto \phi(x)^2$). Likewise, weights in S^C contribute variance to Σ but contribute nothing to J_S (they are independent of χ_S. and average to zero under the sign symmetry). Thus the maximizer of $R_k(\boldsymbol{w})$ lives in the span of $\mathbb{1}_S$ and $\boldsymbol{w}^C$. □

Lemma D.5. *Consider* $\phi = \text{ReLU}$. *Let the inputs* $x_j \in \{\pm 1\}$ *be i.i.d. and* $y(\boldsymbol{x}) = \chi_S(\mathbf{x}) = \prod_{j\in S} x_j$ *with* $|S| = k$. *We assume* $\boldsymbol{w}^*$ *from Theorem D.4. Then, the ratio of the squared neuron-target coupling* $J_S(\mathbf{w})^2$ *to the neuron's self-energy* $\Sigma(\mathbf{w})$ *is a constant* R_k *independent of the scale* α.

Proof. Let $r \sim \text{Binomial}(k, 1/2)$ be the number of components $x_j = -1$ for $j \in S$. The inner product is $s(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \alpha \sum_{j\in S} x_j = \alpha(k - 2r)$, and the target function is $\chi_S(\mathbf{x}) = (-1)^r$.

The neuron-target coupling $J_S(\mathbf{w})$ is the expectation $\mathbb{E}[\phi(\mathbf{w}^\top \mathbf{x})\chi_S(\mathbf{x})]$.

$$J_S(\mathbf{w}) = \mathbb{E}\left[\alpha \cdot [k - 2r]_+ \cdot (-1)^r\right] \tag{83}$$

$$= \alpha \sum_{r=0}^{k} P(r) \cdot [k - 2r]_+ \cdot (-1)^r \tag{84}$$

$$= \alpha \cdot 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)(-1)^r \tag{85}$$

where $[z]_+ = \max(0, z)$, and the sum is restricted to the terms where $k - 2r > 0$.

The neuron's self-energy $\Sigma(\mathbf{w})$ is the expectation $\mathbb{E}[\phi(\mathbf{w}^\top \mathbf{x})^2]$.

$$\Sigma(\mathbf{w}) = \mathbb{E}\left[(\alpha \cdot [k - 2r]_+)^2\right] \tag{86}$$

$$= \alpha^2 \sum_{r=0}^{k} P(r) \cdot [k - 2r]_+^2 \tag{87}$$

$$= \alpha^2 \cdot 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)^2 \tag{88}$$

We define the scale-independent constants D_k and C_k:

$$D_k := 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)(-1)^r \tag{89}$$

$$C_k := 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)^2 \tag{90}$$

such that $J_S(\mathbf{w}) = \alpha D_k$ and $\Sigma(\mathbf{w}) = \alpha^2 C_k$. The ratio is then

$$R_k := \frac{J_S(\mathbf{w})^2}{\Sigma(\mathbf{w})} = \frac{(\alpha D_k)^2}{\alpha^2 C_k} = \frac{D_k^2}{C_k}$$

which is independent of α, thus proving the proposition.

□

D.5. Exact solution of the FP equation

Theorem D.6. *Using the setup from Theorem D.5, espeically the proposed $\boldsymbol{w}^*$, the FP is given by*

$$m_S = (1-m_S)NR_k\left(1-\frac{\sigma_a^{-2}}{A^\star(m_S)}\right), \quad A^\star(m_S) = \frac{-C_k+\sqrt{C_k^2+4\left(\frac{dk}{\sigma_w^2}\right)N^{2\gamma}(1-m_S)^2D_k^2\,\sigma_a^{-2}}}{2\left(\frac{dk}{\sigma_w^2}\right)\kappa^2N^{2\gamma}} \tag{91}$$

Proof. From Theorem D.3, we know:

$$m_S = \frac{N^{1-2\gamma}}{\kappa^2}(1-m_S)\left\langle\frac{J_S(\mathbf{w})^2}{\sigma_a^{-2}+\Sigma(\mathbf{w})/(\kappa^2N^{2\gamma})}\right\rangle_{\mathbf{w}|m_S}$$

Assuming the posterior distribution of weights is sharply peaked around the optimal direction given by the ansatz in Theorem D.5, the expectation $\langle\cdot\rangle_{\mathbf{w}|m_S}$ reduces to an evaluation at the optimal weight scale $\alpha^\star$. The value of $\alpha^\star$ is determined by minimizing the single-neuron effective action $\mathcal{S}_{\text{eff}}(\alpha)$:

$$\mathcal{S}_{\text{eff}}(\alpha) = \underbrace{\frac{dk}{2\sigma_w^2}\alpha^2}_{\text{Weight Prior}} + \underbrace{\frac{1}{2}\ln A(\alpha)}_{\text{Normalizer}} - \underbrace{\frac{1}{2}\frac{\left[(1-m_S)J_S(\alpha)\right]^2/(\kappa^2N^{2\gamma})}{A(\alpha)}}_{\text{Data Gain}}$$

where $A(\alpha) = \sigma_a^{-2}+\Sigma(\alpha)/(\kappa^2N^{2\gamma}) = \sigma_a^{-2}+\alpha^2C_k/(\kappa^2N^{2\gamma})$.

Setting $\frac{\partial\mathcal{S}_{\text{eff}}}{\partial\alpha}=0$ yields a quadratic equation for the optimal value $A^\star = A(\alpha^\star)$:

$$\left(\frac{dk}{\sigma_w^2}\right)\kappa^2N^{2\gamma}(A^\star)^2 + C_k\,A^\star - \frac{(1-m_S)^2D_k^2\,\sigma_a^{-2}}{\kappa^2} = 0$$

The positive root of this equation gives the solution for $A^\star(m_S)$:

$$A^\star(m_S) = \frac{-C_k+\sqrt{C_k^2+4\left(\frac{dk}{\sigma_w^2}\right)N^{2\gamma}(1-m_S)^2D_k^2\,\sigma_a^{-2}}}{2\left(\frac{dk}{\sigma_w^2}\right)\kappa^2N^{2\gamma}}$$

We now substitute this back into the self-consistency equation. Using the definitions of J_S, Σ, and $A^\star$, we have $(\alpha^\star)^2 = \frac{\kappa^2N^{2\gamma}}{C_k}(A^\star-\sigma_a^{-2})$.

$$\begin{aligned}
m_S &= (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{J_S(\alpha^\star)^2}{A^\star} && (92)\\
&= (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{(\alpha^\star)^2D_k^2}{A^\star} && (93)\\
&= (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{1}{A^\star}\left[\frac{\kappa^2N^{2\gamma}}{C_k}(A^\star-\sigma_a^{-2})\right]D_k^2 && (94)\\
&= (1-m_S)N\frac{D_k^2}{C_k}\frac{A^\star-\sigma_a^{-2}}{A^\star} && (95)\\
&= (1-m_S)NR_k\left(1-\frac{\sigma_a^{-2}}{A^\star(m_S)}\right) && (96)
\end{aligned}$$

This gives the final fixed-point equation for the order parameter m_S. □

Theorem D.7. *Consider the fixed–point equation in the infinite P-limit*

$$m_S \;=\; (1-m_S)\, N R_k\!\left(1-\frac{\sigma_a^{-2}}{A^\star(m_S)}\right), \tag{97}$$

with $A^\star(m_S)$ given implicitly as the positive root of

$$\left(\tfrac{dk}{\sigma_w^2}\right)\kappa^2 N^{2\gamma}\left(A^\star\right)^2 \;+\; C_k\, A^\star \;-\; (1-m_S)^2 D_k^2\,\sigma_a^{-2} \;=\; 0, \tag{98}$$

and define $C := \frac{dk}{\sigma_w^2}$. Then the critical noise *level*

$$\kappa_c^2 \;=\; \frac{\sqrt{C_k^2 + 4\,C\,N^{2\gamma} D_k^2 \sigma_a^{-2}} \;-\; C_k}{2\,C\,\sigma_a^{-2}\,N^{2\gamma}} \tag{99}$$

marks a phase transition:

(i) If $\kappa^2 \geq \kappa_c^2$, then $A^\star(0) \leq \sigma_a^{-2}$ and $m_S = 0$ is a fixed point; in particular for $\kappa^2 = \kappa_c^2$ we have $A^\star(0) = \sigma_a^{-2}$ and the only solution is $m_S = 0$.

(ii) If $\kappa^2 < \kappa_c^2$, then $A^\star(0) > \sigma_a^{-2}$ and there exists a unique nontrivial solution $m_S \in (0,1)$, thus the system exhibits symmetry breaking $m_S : 0 \to m_S > 0$ as κ crosses κ_c from above.

Proof. Set $m_S = 0$ in (98) to get

$$C\,\kappa^2 N^{2\gamma}\,(A^\star)^2 + C_k A^\star - D_k^2 \sigma_a^{-2} = 0. \tag{100}$$

At the onset of FL the trivial fixed point $m_S = 0$ changes stability precisely when the right–hand side of the FP map ceases to vanish at $m_S = 0$, i.e. when

$$N R_k\!\left(1-\frac{\sigma_a^{-2}}{A^\star(0)}\right) = 0 \quad\Longleftrightarrow\quad A^\star(0) = \sigma_a^{-2}. \tag{101}$$

Plugging $A^\star = \sigma_a^{-2}$ into the quadratic and solving for κ^2 gives

$$C\,\kappa_c^2 N^{2\gamma}\sigma_a^{-4} + C_k \sigma_a^{-2} - D_k^2 \sigma_a^{-2} = 0, \tag{102}$$

which, after rearrangement, yields

$$\kappa_c^2 = \frac{\sqrt{C_k^2 + 4\,C\,N^{2\gamma} D_k^2 \sigma_a^{-2}} \;-\; C_k}{2\,C\,\sigma_a^{-2}\,N^{2\gamma}}, \tag{103}$$

the stated expression.

For $\kappa^2 > \kappa_c^2$ the quadratic gives $A^\star(0) \leq \sigma_a^{-2}$, hence $1 - \sigma_a^{-2}/A^\star(0) \leq 0$ and the FP map evaluates to 0 at $m_S = 0$, so $m_S = 0$ is a (and in fact the only) solution. For $\kappa^2 < \kappa_c^2$ we have $A^\star(0) > \sigma_a^{-2}$ so the FP map at $m_S = 0$ is strictly positive, continuity and the fact that the map is strictly decreasing in m_S (because $(1 - m_S)$ and $A^\star(m_S)$ both decrease with m_S) imply a unique intersection with the diagonal in $(0,1)$, i.e. a unique $m_S \in (0,1)$ solves the FP equation. □

D.6. How MF-ARD beats the curse of dimensionality

We use the following notation. Let $S \subset [d]$ with $|S| = k$ denote the (unknown) support. For a weight coordinate w_j at a given outer iterate, write

$$v_j := \langle w_j^2 \rangle_p, \qquad v_j'(\rho) := \mathbb{E}_{p_\rho}[w_j^2] \tag{104}$$

for the current and next second moments, respectively. For a vector of ARD precisions $\rho \in \mathbb{R}_+^d$ define the explicit ARD map

$$\rho_j(v_j) \;=\; \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\, v_j} \;=:\; \frac{A}{B(v_j)}, \qquad A := \alpha_0 + \frac{N}{2},\;\; B(v) := \frac{\alpha_0}{d} + \frac{N}{2} v. \tag{105}$$

We write $w_{-j} := (w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d)$ for all coordinates except j and write p_ρ for the posterior p_{ARD} to make the ρ dependence explicit.

Assumption Here, we will state the assumption that is needed to prove the theorem: ε **symmetry breaking towards** S. There exists an outer iterate $t_0 = \mathcal{O}(1)$ and a constant $\varepsilon_0 > 0$ (independent of d) such that

$$\min_{j \in S} v_j^{t_0} \;-\; \max_{j \notin S} v_j^{t_0} \;\geq\; c\,\varepsilon_0, \qquad v_j^t := \langle w_j^2 \rangle_p \text{ at outer time } t. \tag{106}$$

We need to establish the following global bound.

Lemma D.8. *Fix $j \in [d]$ and w_{-j}. Assume $\|x\|_\infty \leq 1$ (e.g. $x \in \{\pm 1\}^d$) and $\phi(z) = \max(0, z)$, as well as $m_S \leq 1$ Let*

$$g(\boldsymbol{w}) = \frac{1}{2} \ln\Big(\sigma_a^{-2} + \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}\Big) - \frac{\big(J_{\mathcal{Y}}(\boldsymbol{w}) - m_S J_S(\boldsymbol{w})\big)^2}{2\kappa^4 N^{2\gamma}\big(\sigma_a^{-2} + \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}\big)}. \tag{107}$$

Then there exists $L_\star > 0$, independent of d and w_{-j}, such that the map $t \mapsto g(w_{-j}, t)$ satisfies

$$\big|\partial_t^2 g(w_{-j}, t)\big| \;\leq\; L_\star \qquad \textit{for a.e. } t \in \mathbb{R}. \tag{108}$$

Consequently, for all $t \in \mathbb{R}$,

$$g(w_{-j}, t) \;\geq\; g(w_{-j}, 0) - \frac{L_\star}{2}\, t^2. \tag{109}$$

Proof. Fix $j \in [d]$ and w_{-j}. Write $t := w_j$ and, for each input x, set $z(t, x) := \boldsymbol{w}^\top x = t\, x_j + c_x$ with $c_x := w_{-j}^\top x_{-j}$. Throughout, $\phi(z) = \max(0, z)$ and $\|x\|_\infty \leq 1$.

1. Derivative of Σ and J_A

For $\Sigma(\boldsymbol{w}) = \mathbb{E}_x[\phi(z)^2] = \mathbb{E}_x[z(t,x)^2\, \mathbf{1}\{z(t,x) > 0\}]$ we have, for a.e. t,

$$\partial_t \Sigma \;=\; 2\,\mathbb{E}_x\big[z \mathbf{1}_{\{z>0\}}\, x_j\big], \qquad \partial_t^2 \Sigma \;=\; 2\,\mathbb{E}_x\big[x_j^2\, \mathbf{1}_{\{z>0\}}\big] \;\leq\; 2, \tag{110}$$

because $|x_j| \leq 1$. By Cauchy–Schwarz,

$$|\partial_t \Sigma| \;\leq\; 2\big(\mathbb{E}_x[z^2 \mathbf{1}_{\{z>0\}}]\big)^{1/2}\big(\mathbb{E}_x[x_j^2 \mathbf{1}_{\{z>0\}}]\big)^{1/2} \;\leq\; 2\sqrt{\Sigma}. \tag{111}$$

For $J_A(\boldsymbol{w}) = \mathbb{E}_x[\phi(z)\, \chi_A(x)]$ (with $|\chi_A| \leq 1$), we get for a.e. t:

$$\partial_t J_A \;=\; \mathbb{E}_x[\mathbf{1}_{\{z>0\}}\, x_j\, \chi_A(x)]\,, \quad \partial_t^2 J_A \;=\; 0 \;\text{(a.e.)}, \qquad |\partial_t J_A| \;\leq\; \mathbb{E}|x_j| \;\leq\; 1, \tag{112}$$

and by Cauchy–Schwarz again

$$|J_A| \;\leq\; \big(\mathbb{E}_x[\phi(z)^2]\big)^{1/2}\big(\mathbb{E}_x[\chi_A(x)^2]\big)^{1/2} \;=\; \sqrt{\Sigma}. \tag{113}$$

2. Decompose g and bound each second derivative

Let

$$a := \sigma_a^{-2}, \qquad c := \frac{1}{\kappa^2 N^{2\gamma}}, \qquad J_\Delta := J_{\mathcal{Y}} - m_S J_S, \qquad q := J_\Delta^2, \tag{114}$$

so

$$g(\boldsymbol{w}) \;=\; \underbrace{\tfrac{1}{2}\log\big(a + c\Sigma\big)}_{=:g_1(\Sigma)} \;-\; \underbrace{\frac{q}{2\kappa^4 N^{2\gamma}\,(a + c\Sigma)}}_{=:g_2(\Sigma, J_\Delta)}. \tag{115}$$

Term 1: Set $h_1(s) := \frac{1}{2}\log(a + cs)$, so $h_1'(s) = \frac{c}{2(a+cs)}$ and $h_1''(s) = -\frac{c^2}{2(a+cs)^2}$. By the chain rule,

$$\partial_t^2 g_1 \;=\; h_1''(\Sigma)\,(\partial_t \Sigma)^2 \;+\; h_1'(\Sigma)\,\partial_t^2 \Sigma. \tag{116}$$

Using (111) and $\partial_t^2 \Sigma \le 2$,

$$\Big|h_1''(\Sigma)\,(\partial_t\Sigma)^2\Big| \;\le\; \frac{c^2}{2(a + c\Sigma)^2}\cdot 4\Sigma \;=\; 2c^2\,\frac{\Sigma}{(a + c\Sigma)^2} \;\le\; \frac{c\,\sigma_a^2}{2}, \tag{117}$$

where the last inequality follows by maximizing $u \mapsto \frac{u}{(a+cu)^2}$ at $u = a/c$. Also $\big|h_1'(\Sigma)\,\partial_t^2\Sigma\big| \le \frac{c}{2(a+c\Sigma)}\cdot 2 \le c\,\sigma_a^2$. Hence

$$\big|\partial_t^2 g_1\big| \;\le\; \frac{3}{2}\,c\,\sigma_a^2\,. \tag{118}$$

Term 2: Write $g_2 = -\alpha\,\frac{q}{a+c\Sigma}$ with $\alpha := \frac{1}{2\kappa^4 N^{2\gamma}}$. Differentiating twice and grouping terms gives (for a.e. t):

$$\partial_t^2 g_2 = -\alpha\left[\frac{q''}{a + c\Sigma} - \frac{2\,q'\,c\,\Sigma'}{(a + c\Sigma)^2} - \frac{q\,c\,\Sigma''}{(a + c\Sigma)^2} + \frac{2\,q\,(c\Sigma')^2}{(a + c\Sigma)^3}\right],$$

where primes denote ∂_t. We now bound q, q', q'' with step 1 and using (113) and $|\partial_t J_A| \le 1$.

- q: Using $|J_\Delta| \;\le\; |J_{\mathcal{Y}}| + |m_S|\,|J_S| \;\le\; (1 + |m_S|)\sqrt{\Sigma} =: C_\Delta\sqrt{\Sigma}$ we get $q = J_\Delta^2 \le C_\Delta^2\,\Sigma$.
- $|q'|$: We get

$$\begin{aligned}|q'| &= 2|J_\Delta||\partial_t J_\Delta| \le 2C_\Delta\sqrt{\Sigma}\cdot\big(|\partial_t J_{\mathcal{Y}}| + |m_S||\partial_t J_S|\big) \le 2C_\Delta\sqrt{\Sigma}(1 + |m_S|)\\ &\le 2C_\Delta^2\sqrt{\Sigma}\end{aligned}$$

- $|q''|$: We get $q'' = 2(\partial_t J_\Delta)^2 \le 2C_\Delta^2$

Together with $\Sigma' \le 2\sqrt{\Sigma}$ and $\Sigma'' \le 2$, we get

$$\Big|\tfrac{q''}{a+c\Sigma}\Big| \le \frac{2C_\Delta^2}{a}, \tag{119}$$

$$\Big|\tfrac{2\,q'\,c\,\Sigma'}{(a+c\Sigma)^2}\Big| \le \frac{2\cdot 2C_\Delta^2\sqrt{\Sigma}\cdot c\cdot 2\sqrt{\Sigma}}{(a + c\Sigma)^2} = 8C_\Delta^2 c\,\frac{\Sigma}{(a + c\Sigma)^2} \;\le\; \frac{2C_\Delta^2}{a}, \tag{120}$$

$$\Big|\tfrac{q\,c\,\Sigma''}{(a+c\Sigma)^2}\Big| \le \frac{C_\Delta^2\,\Sigma\cdot c\cdot 2}{(a + c\Sigma)^2} \;\le\; \frac{C_\Delta^2}{a}, \tag{121}$$

$$\Big|\tfrac{2\,q\,(c\Sigma')^2}{(a+c\Sigma)^3}\Big| \le \frac{2\,C_\Delta^2\,\Sigma\cdot c^2\cdot 4\Sigma}{(a + c\Sigma)^3} = 8C_\Delta^2 c^2\,\frac{\Sigma^2}{(a + c\Sigma)^3} \;\le\; \frac{32C_\Delta^2}{27\,a}, \tag{122}$$

where in the last two lines we used that $u \mapsto \frac{u}{(a+cu)^2}$ and $u \mapsto \frac{u^2}{(a+cu)^3}$ are maximized at $u = \frac{a}{c}$ and $u = \frac{2a}{c}$, respectively, with finite maxima depending only on a, c. Therefore $\left|\partial_t^2 g_2\right| \leq \alpha\, K_2$ for a constant K_2 depending only on (a, c, m_S), and independent of d, w_{-j} and t.

3: Uniform bound on $\partial_t^2 g$

Combining (118) and the bound for $\partial_t^2 g_2$, there exists

$$L_\star \;:=\; \frac{3}{2}\, c\, \sigma_a^2 \;+\; \alpha\, K_2 \tag{123}$$

such that $|\partial_t^2 g(w_{-j}, t)| \leq L_\star$ for a.e. $t \in \mathbb{R}$. This proves the first claim.

4: Standard taylor expansion

For any twice–differentiable h with $|h''| \leq L_\star$ a.e., the 1D Taylor inequality gives

$$h(t) \;\geq\; h(0) \;+\; h'(0)\, t \;-\; \tfrac{L_\star}{2}\, t^2 \qquad (\forall t \in \mathbb{R}).$$

Applying this to $h(t) = g(w_{-j}, t)$ yields

$$g(w_{-j}, t) \;\geq\; g(w_{-j}, 0) \;+\; \partial_t g(w_{-j}, 0)\, t \;-\; \frac{L_\star}{2}\, t^2.$$

In the parity setting and for $j \notin S$ (off–support), symmetry implies $\partial_t g(w_{-j}, 0) = 0$, giving the stated global quadratic lower bound $g(w_{-j}, t) \geq g(w_{-j}, 0) - \frac{L_\star}{2} t^2$. □

This lemma provides a precise bound on the conditional second moment of off-support w_j. The key insight is that when the precision parameter ρ_j is large enough to dominate the coupling term, the second moment behaves essentially like that of a Gaussian with precision ρ_j, up to small corrections.

Lemma D.9. *Let $j \notin S$ be an off-support coordinate and condition on w_{-j}. Given Lemma D.8, there exists $L_\star > 0$ (independent of d and w_{-j}) such that the coupling function $t \mapsto g(w_{-j}, t)$ satisfies:*

$$|\partial_t^2 g(w_{-j}, t)| \leq L_\star \quad \textit{for a.e. } t \in \mathbb{R}.$$

*Assume furthermore the **off-support symmetry condition**:*

$$\partial_j g(w_{-j}, 0) = 0.$$

Consider the conditional distribution for w_j:

$$p(w_j \mid w_{-j}, \rho) \propto \exp\big(-U_j(w_j)\big),$$

where the potential is given by:

$$U_j(w_j) = \frac{1}{2}\rho_j w_j^2 + g(w_j, w_{-j}).$$

If $\rho_j > L_\star$, then for every fixed $\theta \in (0, 1]$, there exist constants $C, c > 0$ (depending only on $L_\star$ and θ, but independent of d and w_{-j}) such that:

$$\mathbb{E}\big[w_j^2 \mid w_{-j}\big] \leq \frac{e^{L_\star \theta^2}}{\rho_j - L_\star} + C\, e^{-c\,(\rho_j - L_\star)\,\theta^2}.$$

In particular, for any $\varepsilon > 0$, one can choose $\theta \in (0, 1]$ small enough so that $e^{L_\star \theta^2} \leq 1 + \varepsilon$, yielding:

$$\mathbb{E}[w_j^2 \mid w_{-j}] \leq \frac{1 + \varepsilon}{\rho_j - L_\star} + C\, e^{-c\,(\rho_j - L_\star)}.$$

Hence, if $\rho_j = \Omega(d)$, then:

$$\mathbb{E}[w_j^2 \mid w_{-j}] \leq \frac{1 + o_d(1)}{\rho_j}$$

with the $o_d(1)$ term uniform in w_{-j}.

Proof. The proof uses a careful splitting argument to handle the near-Gaussian behavior in the core region while controlling exponential tails. The key is to leverage the smoothness bound to create uniform upper and lower envelopes for the potential function.

Fix $j \notin S$ and condition on w_{-j}. For notational simplicity, write $w := w_j$, $\rho := \rho_j$, and define:

$$U(w) = \frac{1}{2}\rho w^2 + g(w),$$

where $g(w) := g(w, w_{-j})$ is the coupling term.

Step 1: Establishing potential envelopes

By Lemma D.8, we have $|g''(t)| \leq L_\star$ for almost every t, and the off-support symmetry gives $g'(0) = 0$.

Using Taylor expansion with integral remainder, for all $w \in \mathbb{R}$:

$$-\frac{L_\star}{2}w^2 \leq g(w) - g(0) \leq \frac{L_\star}{2}w^2. \tag{124}$$

Define the effective precision $m := \rho - L_\star > 0$. From (124), we obtain two crucial envelopes:

Global lower envelope:

$$U(w) \geq U(0) + \frac{m}{2}w^2 \quad \text{for all } w \in \mathbb{R}. \tag{125}$$

Local upper envelope: For any $\theta > 0$ and $|w| \leq \theta$:

$$U(w) \leq U(0) + \frac{\rho}{2}w^2 + \frac{L_\star}{2}w^2 \leq U(0) + \frac{m}{2}w^2 + L_\star\theta^2. \tag{126}$$

The global lower envelope ensures integrability, while the local upper envelope allows us to approximate the distribution by a Gaussian in the core region.

Step 2: Setting up the moment computation

Let μ be the conditional measure with density proportional to $e^{-U(w)}$. Define the normalization and second moment integrals:

$$Z := \int_{\mathbb{R}} e^{-U(w)}\, dw, \qquad N := \int_{\mathbb{R}} w^2 e^{-U(w)}\, dw,$$

so that $\mathbb{E}_\mu[w^2] = N/Z$.

We partition $\mathbb{R}$ into the **inside region** $I := \{|w| \leq \theta\}$ and the **outside region** $O := \{|w| > \theta\}$, writing $Z = Z_I + Z_O$ and $N = N_I + N_O$.

Step 3: Bounding the inside region contribution

Using the local upper envelope (126):

$$Z_I = \int_{|w|\leq\theta} e^{-U(w)}\, dw \geq e^{-U(0)} e^{-L_\star\theta^2} \int_{|w|\leq\theta} e^{-\frac{m}{2}w^2}\, dw.$$

Using the global lower envelope (125):

$$N_I = \int_{|w|\leq\theta} w^2 e^{-U(w)}\, dw \leq e^{-U(0)} \int_{|w|\leq\theta} w^2 e^{-\frac{m}{2}w^2}\, dw.$$

Combining these bounds:

$$\frac{N_I}{Z} \leq \frac{N_I}{Z_I} \leq e^{L_\star\theta^2} \cdot \frac{\int_{|w|\leq\theta} w^2 e^{-\frac{m}{2}w^2}\, dw}{\int_{|w|\leq\theta} e^{-\frac{m}{2}w^2}\, dw} \leq \frac{e^{L_\star\theta^2}}{m}. \tag{127}$$

The final inequality follows because the ratio represents the truncated second moment of a zero-mean Gaussian with precision m, which is bounded by the untruncated value $1/m$.

Step 4: Controlling the outside region contribution

For the tail contribution, we use the global lower envelope (125):

$$N_O = \int_{|w|>\theta} w^2 e^{-U(w)}\, dw \le e^{-U(0)} \int_{|w|>\theta} w^2 e^{-\frac{m}{2}w^2}\, dw.$$

A standard Gaussian tail bound (obtained by integration by parts) gives, for $a > 0$ and $\theta > 0$:

$$\int_\theta^\infty t^2 e^{-at^2}\, dt \le \left(\frac{\theta}{2a} + \frac{1}{4a^2\theta}\right) e^{-a\theta^2}.$$

Applying this with $a = m/2$ and doubling for both tails:

$$\int_{|w|>\theta} w^2 e^{-\frac{m}{2}w^2}\, dw \le \frac{2}{m}\left(\theta + \frac{1}{m\theta}\right) e^{-\frac{m}{2}\theta^2}. \tag{128}$$

Using the lower bound on Z_I from Step 3 and (128):

$$\frac{N_O}{Z} \le \frac{N_O}{Z_I} \le \frac{e^{L_\star\theta^2}}{c_0} \cdot \frac{2}{m}\left(\theta + \frac{1}{m\theta}\right) e^{-\frac{m}{2}\theta^2}\sqrt{m} \le C_1 e^{-c_1 m\theta^2},$$

where $c_0, C_1, c_1 > 0$ are absolute constants depending only on $L_\star$ and θ. The polynomial factors in m are absorbed into the constant since the exponential term dominates for $m > 0$.

Step 5: Combining the contributions

From (127) and the outside region bound:

$$\mathbb{E}_\mu[w^2] = \frac{N}{Z} = \frac{N_I}{Z} + \frac{N_O}{Z} \le \frac{e^{L_\star\theta^2}}{m} + Ce^{-cm\theta^2},$$

which establishes the main inequality with $m = \rho_j - L_\star$.

For the refined bound, choosing $\theta > 0$ small enough so that $e^{L_\star\theta^2} \le 1 + \varepsilon$ gives the result.

Finally, if $\rho_j = \Omega(d)$, then $m = \rho_j - L_\star = \Omega(d)$, so the exponential tail term is $e^{-\Omega(d)}$ uniformly in w_{-j}, and:

$$\frac{1+\varepsilon}{\rho_j - L_\star} = \frac{1+\varepsilon}{\rho_j\left(1 - \frac{L_\star}{\rho_j}\right)} = \frac{1 + o_d(1)}{\rho_j}.$$

□

The next lemma establishes that off-support w_j contract towards equilibrium values of order $O(1/d)$ when initialized in a suitable bootstrap region. The key insight is that by controlling the initialization within $O(1/d)$, we can ensure the dynamics remain stable and converge.

Lemma D.10. *Let $r > 0$ be a radius parameter and let $L_\star$ be the global smoothness constant from Lemma D.8. We define the following key quantities:*

$$A := \alpha_0 + \frac{N}{2}, \qquad \textit{(total prior mass)} \tag{129}$$

$$B(v) := \frac{\alpha_0}{d} + \frac{N}{2} \cdot v, \qquad \textit{(effective local prior)} \tag{130}$$

For any threshold $K_\star > 0$ and dimension $d \in \mathbb{N}$, we introduce the bootstrap parameters:

$$U_d(K_\star) := \frac{\alpha_0 + \frac{N}{2}K_\star}{d}, \qquad \textit{(maximum local prior)} \tag{131}$$

$$a_d(r, K_\star) := \frac{e^{L_\star r^2} \cdot \alpha_0}{A - L_\star \cdot U_d(K_\star)}, \qquad \textit{(affine drift term)} \tag{132}$$

$$b_d(r, K_\star) := \frac{e^{L_\star r^2} \cdot (N/2)}{A - L_\star \cdot U_d(K_\star)}. \qquad \textit{(contraction coefficient)} \tag{133}$$

Note that as $d \to \infty$, we have the asymptotic behavior:

$$b_d(r, K_\star) \to b_\infty(r) := \frac{e^{L_\star r^2} \cdot (N/2)}{A}.$$

Choose $r > 0$ sufficiently small so that $b_\infty(r) < 1$. This ensures asymptotic contraction. Then there exist $d_0 \in \mathbb{N}$, a finite threshold $K_\star > 0$ (both independent of d), and positive constants $C, c, c' > 0$ such that for all $d \geq d_0$:

1. ***One-step contraction bound:*** *If $v^t_{\text{off}} \leq K_\star/d$, then with $\rho^t_{\text{off}} = A/B(v^t_{\text{off}})$,*

$$v^{t+1}_{\text{off}} \leq \frac{e^{L_\star r^2} \cdot B(v^t_{\text{off}})}{A - L_\star B(v^t_{\text{off}})} + C\, e^{-c\,(\rho^t_{\text{off}} - L_\star)\, r^2} \tag{134}$$

$$\leq \frac{a_d(r, K_\star)}{d} + b_d(r, K_\star) \cdot v^t_{\text{off}} + \eta_d, \tag{135}$$

   where the exponential tail satisfies $\eta_d \leq Ce^{-c'd}$.

2. ***Bootstrap invariance:*** *There exists a finite $K_\star$ such that*

$$v^t_{\text{off}} \leq \frac{K_\star}{d} \quad \Longrightarrow \quad v^{t+1}_{\text{off}} \leq \frac{K_\star}{d} \quad \textit{for all } t \geq 0.$$

   In particular, any initialization with $v^0_{\text{off}} \leq K_\star/d$ remains within the bootstrap region for all time.

3. ***Equilibrium bound:*** *Along any trajectory that remains in the bootstrap region,*

$$v^\star_{\text{off}} \leq \frac{a_d(r, K_\star)}{(1 - b_d(r, K_\star)) \cdot d} + O(e^{-c'd}) = \Theta\left(\frac{1}{d}\right).$$

Proof. The proof proceeds in four main steps: establishing a one-step bound, deriving an affine upper bound, proving bootstrap invariance, and analyzing convergence.

Fix $r > 0$ and let $L_\star$ be the smoothness constant from Lemma D.8. Consider an off-support coordinate at outer iteration t. We write $v := v^t_{\text{off}}$, $B(v) = \frac{\alpha_0}{d} + \frac{N}{2}v$, and $\rho := \rho^t_{\text{off}} = \frac{A}{B(v)}$ with $A = \alpha_0 + \frac{N}{2}$ (see (105)).

Throughout, we assume $v \leq K_\star/d$ for some threshold $K_\star > 0$ to be determined.

Step 1: Establishing the one-step bound

We begin by applying the second moment bound from Lemma D.9. For any $r > 0$ and whenever $\rho > L_\star$, this yields:

$$v^{t+1}_{\text{off}} = \mathbb{E}[w_j^2] \leq \frac{e^{L_\star r^2}}{\rho - L_\star} + C\, e^{-c\,(\rho - L_\star)\, r^2}. \tag{136}$$

Substituting $\rho = \frac{A}{B(v)}$, the main term becomes:

$$\frac{e^{L_\star r^2}}{\rho - L_\star} = e^{L_\star r^2} \cdot \frac{B(v)}{A - L_\star B(v)}.$$

Since we assume $v \leq K_\star/d$, we have the upper bound:

$$B(v) \leq U_d(K_\star) := \frac{\alpha_0 + \frac{N}{2}K_\star}{d} =: \frac{C_B}{d},$$

which implies $\rho = \frac{A}{B(v)} \geq \frac{A \cdot d}{C_B}$. This lower bound on ρ allows us to control the exponential tail:

$$C\, e^{-c\,(\rho - L_\star)\, r^2} \leq C\, e^{-c' d} =: \eta_d \tag{137}$$

for some $c' = c'(A, C_B, L_\star, r) > 0$ that depends only on the fixed parameters.

Step 2: Deriving the affine upper bound

To obtain a tractable recursion, we upper bound the main term using convexity. Define the function:

$$f(u) := \frac{u}{A - L_\star u},$$

which is convex on the interval $[0, A/L_\star)$.

For all $u \in [0, U_d(K_\star)]$, monotonicity of the denominator gives:

$$f(u) = \frac{u}{A - L_\star u} \leq \frac{u}{A - L_\star U_d(K_\star)}.$$

Applying this with $u = B(v)$, we obtain the key affine upper bound:

$$e^{L_\star r^2} \cdot \frac{B(v)}{A - L_\star B(v)} \leq \frac{e^{L_\star r^2}}{A - L_\star U_d(K_\star)} \left(\frac{\alpha_0}{d} + \frac{N}{2} v \right) =: \frac{a_d(r, K_\star)}{d} + b_d(r, K_\star) \cdot v, \tag{138}$$

where we have defined:

$$a_d(r, K_\star) := \frac{e^{L_\star r^2} \cdot \alpha_0}{A - \frac{L_\star C_B}{d}}, \tag{139}$$

$$b_d(r, K_\star) := \frac{e^{L_\star r^2} \cdot (N/2)}{A - \frac{L_\star C_B}{d}}. \tag{140}$$

Note that $a_d(r, K_\star) = \Theta(1)$ and $b_d(r, K_\star) \to b_\infty(r) = \frac{e^{L_\star r^2} \cdot (N/2)}{A}$ as $d \to \infty$.

Step 3: Establishing bootstrap invariance

The key is to choose parameters ensuring contraction. First, pick $r > 0$ small enough so that:

$$b_\infty(r) = \frac{e^{L_\star r^2} \cdot (N/2)}{A} < 1.$$

This is possible because $\frac{N/2}{A} < 1$ (since $\alpha_0 > 0$) and $e^{L_\star r^2} \to 1$ as $r \to 0$.

With r fixed, there exists d_0 such that for all $d \geq d_0$:

$$b_d(r, K_\star) \leq \frac{b_\infty(r) + 1}{2} < 1.$$

The bound is uniform in $K_\star$ because $U_d(K_\star) = O(1/d)$.

Combining equations (136), (138), and (137), for $d \geq d_0$ and $v \leq K_\star/d$:

$$v_{\text{off}}^{t+1} \leq \frac{a_d(r, K_\star)}{d} + b_d(r, K_\star) \cdot v + \eta_d.$$

Now choose $K_\star$ large enough so that:

$$K_\star \geq \sup_{d \geq d_0} \frac{a_d(r, K_\star)}{1 - b_d(r, K_\star)} < \infty.$$

The supremum is finite because $a_d(r, K_\star) = \Theta(1)$ and $1 - b_d(r, K_\star)$ is bounded away from zero for $d \geq d_0$.

Finally, enlarge d_0 if necessary so that $\eta_d \leq \frac{1}{2}(1 - b_d(r, K_\star))\frac{K_\star}{d}$ for all $d \geq d_0$. This ensures bootstrap invariance:

$$v_{\text{off}}^t \leq \frac{K_\star}{d} \quad \Longrightarrow \quad v_{\text{off}}^{t+1} \leq \frac{K_\star}{d}.$$

Step 4: Convergence analysis

Within the invariant bootstrap region, the affine recursion:

$$v_{\text{off}}^{t+1} \leq \frac{a_d(r, K_\star)}{d} + b_d(r, K_\star) \cdot v_{\text{off}}^t + \eta_d$$

contracts towards its fixed point. Since $b_d(r, K_\star) < 1$, the equilibrium value satisfies:

$$v_{\text{off}}^\star \leq \frac{a_d(r, K_\star)}{(1 - b_d(r, K_\star)) \cdot d} + O(e^{-c'd}) = \Theta\left(\frac{1}{d}\right).$$

□

Lemma D.11. *Assume Theorem D.10 3. For all sufficiently large d,*

$$\min_{j \in S} v_j^{t_0} \;\geq\; c\,\varepsilon_0 + \max_{j \notin S} v_j^{t_0} \;\geq\; c\,\varepsilon_0 - O(d^{-1}) \;\geq\; \tfrac{c}{2}\varepsilon_0 \;=\; \Theta(1).$$

Consequently, at (and after) time t_0, the ARD precisions satisfy $\rho_{on} = \Theta(1)$ while $\rho_{off} = \Theta(d)$.

Proof. The first display follows directly from the assumption D.6 and $v_{\text{off}}^{t_0} = O(d^{-1})$ (bootstrap entry). For $j \in S$, $B(v_j) = \frac{\alpha_0}{d} + \frac{N}{2}v_j \geq \frac{N}{2} \cdot \frac{c}{2}\varepsilon_0$ for large d, so $\rho_j = A/B(v_j) = \Theta(1)$ uniformly in d. For $j \notin S$, by Theorem D.10 3), $v_j = \Theta(d^{-1})$, hence $B(v_j) = \Theta(d^{-1})$ and $\rho_j = \Theta(d)$. □

Lemma D.12. *For a neuron with equal weights on S:*

$$C_{\text{MF}} = \frac{dk}{\sigma_w^2} \qquad \textit{and} \qquad C_{\text{ARD}}(t) = \sum_{j \in S} \rho_j^t = \Theta(k) \;\; \textit{for all } t \geq t_0 \textit{ under Theorem D.11.}$$

Proof. MF: With prior penalty $\frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2$, along $\boldsymbol{w} = \alpha\,\mathbf{1}_S$ we get $\frac{dk}{2\sigma_w^2}\alpha^2$, hence $C_{\text{MF}} = \frac{dk}{\sigma_w^2}$. *ARD:* By Theorem D.11, $\rho_j^t = \Theta(1)$ for $j \in S$ and $t \geq t_0$, so $C_{\text{ARD}}(t) = \Theta(k)$. □

Lemma D.13. *At the symmetric initialization point $m_S = 0$, the FL threshold satisfies:*

$$\kappa_c^2 = \frac{\sqrt{C_k^2 + 4CN^{2\gamma}D_k^2\sigma_a^{-2}} - C_k}{2C\sigma_a^{-2}N^{2\gamma}}, \tag{141}$$

where:

- *C is the quadratic curvature along the equal-weights S-direction,*
- *C_k, D_k are geometric constants depending only on the support size k,*
- *N is the number of training samples, γ is a scaling exponent,*
- *σ_a controls the output noise level.*

For large curvature C (corresponding to strong regularisation), the threshold simplifies to:

$$\kappa_c^2 \sim \frac{|D_k|\sigma_a}{N^\gamma\sqrt{C}} =: \frac{\Lambda_k}{\sqrt{C}}, \tag{142}$$

where $\Lambda_k := |D_k|\sigma_a/N^\gamma$.

Proof. By Lemma D.4, the optimal weight configuration is $\mathbf{w}^* = \alpha\mathbf{1}_S$. By Lemma D.5, this gives scale-independent constants $C_k = \mathbb{E}[Z_+^2]$ and $D_k = \mathbb{E}[Z_+\chi_S(x)]$ such that:

$$\Sigma(\alpha) = C_k\alpha^2, \qquad J_S(\alpha) = D_k\alpha, \qquad R_k = \frac{D_k^2}{C_k}.$$

From Theorem D.7, the critical threshold occurs when $A^*(0) = \sigma_a^{-2}$, where A^* solves the quadratic equation at $m_S = 0$.

Step 2: Taylor expansion analysis

We analyze the one-dimensional effective potential along $\mathbf{w} = \alpha\mathbf{1}_S$:

$$U(\alpha) = \frac{C}{2}\alpha^2 + g(\alpha),$$

where C is the prior curvature (Lemma D.12) and $g(\alpha)$ is the coupling term from Lemma D.8.

At onset ($m_S = 0$), we have $J_\Delta = J_S = D_k\alpha$. The coupling term becomes:

$$g(\alpha) = \frac{1}{2}\log\left(a + c\Sigma(\alpha)\right) - \frac{J_S(\alpha)^2}{2\kappa^4 N^{2\gamma}(a + c\Sigma(\alpha))},$$

with $a = \sigma_a^{-2}$, $c = \frac{1}{\kappa^2 N^{2\gamma}}$, $\Sigma(\alpha) = C_k\alpha^2$, and $J_S(\alpha) = D_k\alpha$.

Step 3: Computing the Taylor expansion

Expanding around $\alpha = 0$:

For the logarithmic term:

$$\frac{1}{2}\log\left(a + cC_k\alpha^2\right) = \frac{1}{2}\log a + \frac{cC_k}{2a}\alpha^2 + O(\alpha^4).$$

For the rational term:

$$\frac{D_k^2\alpha^2}{2\kappa^4 N^{2\gamma}(a + cC_k\alpha^2)} = \frac{D_k^2\alpha^2}{2\kappa^4 N^{2\gamma}a} + O(\alpha^4) = \frac{\sigma_a^2 D_k^2}{2\kappa^4 N^{2\gamma}}\alpha^2 + O(\alpha^4).$$

Therefore:

$$U(\alpha) = \text{const} + \frac{1}{2}\left[C + \frac{\sigma_a^2 C_k}{\kappa^2 N^{2\gamma}} - \frac{\sigma_a^2 D_k^2}{\kappa^4 N^{2\gamma}}\right]\alpha^2 + O(\alpha^4).$$

Step 4: Onset condition and threshold

FL onset occurs when the quadratic coefficient vanishes:

$$C + \frac{\sigma_a^2 C_k}{\kappa^2 N^{2\gamma}} - \frac{\sigma_a^2 D_k^2}{\kappa^4 N^{2\gamma}} = 0.$$

Multiplying by $\kappa^4 N^{2\gamma}/\sigma_a^2$ and rearranging:

$$CN^{2\gamma}\sigma_a^{-2}\kappa^4 + C_k\kappa^2 - D_k^2 = 0.$$

Solving this quadratic in κ^2 gives:

$$\kappa_c^2 = \frac{\sqrt{C_k^2 + 4CN^{2\gamma}D_k^2\sigma_a^{-2}} - C_k}{2C\sigma_a^{-2}N^{2\gamma}}.$$

Step 5: Asymptotic approximation

For large C, the square root is dominated by the term $4CN^{2\gamma}D_k^2\sigma_a^{-2}$, so:

$$\sqrt{C_k^2 + 4CN^{2\gamma}D_k^2\sigma_a^{-2}} \approx 2|D_k|\sigma_a^{-1}\sqrt{CN^{2\gamma}}.$$

This yields:

$$\kappa_c^2 \sim \frac{2|D_k|\sigma_a^{-1}\sqrt{CN^{2\gamma}}}{2C\sigma_a^{-2}N^{2\gamma}} = \frac{|D_k|\sigma_a}{N^\gamma\sqrt{C}} = \frac{\Lambda_k}{\sqrt{C}}. \tag{143}$$

□

Now the proof of Theorem 4.4 follows by the lemmata above.

Proof. MF: Combine $C_{\text{MF}} = \frac{dk}{\sigma_w^2}$ (Theorem D.12) with (143) to get $\kappa_{c,\text{MF}}^2 \asymp \Lambda_k\sqrt{\sigma_w^2/(dk)} = \Theta(\sqrt{1/(dk)})$. *MF–ARD:* By Theorem D.11, for $t \geq t_0$ we have $C_{\text{ARD}}(t) = \Theta(k)$, hence $\kappa_{c,\text{ARD}}^2 \asymp \Lambda_k/\sqrt{k} = \Theta(\sqrt{1/k})$ from (143). □

D.7. Proof of Theorem 4.1

Proof. In plain MF, the single-neuron posterior has the form

$$p_{\text{MF}}(\boldsymbol{w} \mid m_S) \;\propto\; \exp\Big(-\frac{\rho}{2}\|\boldsymbol{w}\|_2^2 - g(\boldsymbol{w})\Big), \qquad \rho := \frac{d}{\sigma_w^2},$$

with g as in Lemma D.8. For the parity target $y(x) = \chi_S(x)$, the law is invariant under permutations within S and within S^c, hence the second moments are constant on each group:

$$q_{\text{on}} := \mathbb{E}[w_j^2]\ (j \in S), \qquad q_{\text{off}} := \mathbb{E}[w_j^2]\ (j \notin S).$$

Off-support. Fix $j \notin S$ and condition on w_{-j}. In plain MF we have $\rho_j = \rho = \Omega(d)$, and by off-support symmetry $\partial_j g(w_{-j}, 0) = 0$. Lemma D.9 therefore yields

$$\mathbb{E}[w_j^2 \mid w_{-j}] \leq \frac{1 + o_d(1)}{\rho} \quad \text{uniformly in } w_{-j}.$$

Averaging over w_{-j} gives $q_{\text{off}} \leq \frac{1+o_d(1)}{\rho}$.

On-support. For $j \in S$, the same curvature bound $|\partial_t^2 g(w_{-j}, t)| \leq L_\star$ implies the 1D conditional potential $U_j(t) = \frac{\rho}{2}t^2 + g(w_{-j}, t)$ is $(\rho - L_\star)$-strongly convex, hence $(w_j \mid w_{-j}) \leq (\rho - L_\star)^{-1}$. Moreover $|\partial_t g(w_{-j}, 0)|$ is $O(1)$ uniformly in w_{-j} (by the same bounds used in Lemma D.8), so the conditional mean satisfies $|\mathbb{E}[w_j \mid w_{-j}]| = O(\rho^{-1})$, and therefore

$$\mathbb{E}[w_j^2 \mid w_{-j}] = (w_j \mid w_{-j}) + \mathbb{E}[w_j \mid w_{-j}]^2 \leq \frac{1}{\rho - L_\star} + O(\rho^{-2}) = \frac{1 + O(1/\rho)}{\rho}.$$

Averaging gives $q_{\text{on}} \leq \frac{1+O(1/\rho)}{\rho}$.

Finally, the same argument applied to $-g$ (or equivalently using the opposite quadratic envelope from $|\partial_t^2 g| \leq L_\star$) gives matching lower bounds $q_{\text{on}}, q_{\text{off}} \geq \frac{1-O(1/\rho)}{\rho}$, so

$$\frac{q_{\text{on}}}{q_{\text{off}}} = 1 + O\Big(\frac{1}{\rho}\Big) = 1 + O\Big(\frac{1}{d}\Big).$$

Since $\|\boldsymbol{w}_S\|_2^2 = \sum_{j\in S} w_j^2$ and $\|\boldsymbol{w}_{S^c}\|_2^2 = \sum_{j\notin S} w_j^2$, we get

$$R_{\boldsymbol{w}}^{\mathrm{MF}} = \frac{\sqrt{d-k}\,\|\boldsymbol{w}_S\|_2}{\sqrt{k}\,\|\boldsymbol{w}_{S^c}\|_2} = \sqrt{\frac{q_{\mathrm{on}}}{q_{\mathrm{off}}}} = 1 + O\Big(\frac{1}{d}\Big).$$

□

D.8. Proof of the moment closure

Proof. At stationarity, the exact row-energy balance reads

$$\gamma\langle q_j\rangle_\infty + \langle \hat{\rho}_j q_j\rangle_\infty = TN.$$

Applying the linear response assumption

$$(\gamma + \rho_j^r)\langle q_j\rangle_\infty \;\approx\; TN.$$

Dividing by T and using the definition of $\rho_{\mathrm{eff},j}$ yields

$$\langle \boldsymbol{w}_j^2\rangle_\infty \;\approx\; \frac{N}{\rho_{\mathrm{eff},j}}.$$

Finally, by the MF assumption we have $\langle \boldsymbol{w}_j^2\rangle_\infty = N w_j^2$, hence $w_j^2 \approx 1/\rho_{\mathrm{eff},j}$.

The MF–ARD fixed point equation for ρ_j is, by construction,

$$\rho_j = \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2} Q_j},$$

For $\alpha_0 \ll N$, expanding the prefactor ratio gives $\rho_j = (1/w_j^2)(1 + O(\alpha_0/N))$, proving (**??**). This completes the proof upon identifying $\rho_j \equiv \rho_{\mathrm{eff},j}$.

Step 3 (first-moment consistency). The MF–ARD equation for m_A is unchanged in form from the plain MF theory; only the single-neuron distribution entering the expectation changes from p_{MF} to p_{ARD} due to the additional quadratic term $\frac{1}{2}\sum_j \rho_j w_j^2$. Hence the first-moment closure is preserved while the second-moment closure is augmented by $\{\rho_j\}$, proving (iii). □

E. Algorithms

E.1. FP algorithm

Model. Given (X, y) with $X \in \{\pm 1\}^{P\times d}$ and $y \in \mathbb{R}^{P\times 1}$, we approximate the predictor by a particle ensemble,

$$f(\boldsymbol{x}) = s_f \sum_{b=1}^{B} a_b\, \phi(\boldsymbol{w}_b^\top \boldsymbol{x}), \qquad s_f = \frac{N^{1-\gamma}}{B}. \tag{144}$$

We *draw a single dataset of size P once at initialization and keep it fixed for the entire run*. On the training set let $f_p = f(\boldsymbol{x}_p)$ and $r_p = y_p - f_p$. The Langevin temperature is fixed by the likelihood noise as $T = 2\,\kappa^2$ (Section A.1). This choice makes SGLD asymptotically sample from the Bayesian posterior. Because all objectives and gradients are computed as empirical averages over this fixed sample (via $1/P$ factors), the dynamics naturally exhibit finite-P fluctuations.

Sufficient statistics We define the following low-rank statistics per particle b, with $z_{pb} = \boldsymbol{x}_p^\top \boldsymbol{w}_b$:

$$C_{1,b} = \sum_{p=1}^{P} \phi(z_{pb})\, r_p, \qquad C_{2,b} = \sum_{p=1}^{P} \phi(z_{pb})^2, \tag{145}$$

$$G_b^{\text{data}} = -\frac{2}{P} \sum_{p=1}^{P} \Big(r_p - s_f a_b\, \phi(z_{pb})\Big)\, \phi'(z_{pb})\, \big(s_f a_b\big)\, x_p \;\in \mathbb{R}^d. \tag{146}$$

These quantities are the only per-pass summaries we need to form gradients for a_b and w_b.

Prior and SGLD potential We impose a diagonal (ARD) Gaussian prior on the weights and an i.i.d. Gaussian prior on amplitudes:

$$E_{\text{prior}} = \tfrac{1}{2} \sum_{b=1}^{B} \rho^\top (\boldsymbol{w}_b \odot \boldsymbol{w}_b) + \tfrac{1}{2\sigma_a^2} \sum_{b=1}^{B} a_b^2, \quad \mathcal{L}_{\text{data}} = \frac{1}{P} \sum_{p=1}^{P} (y_p - f_p)^2, \tag{147}$$

and update parameters by Langevin dynamics on the potential

$$U(W, a) \;=\; T\, E_{\text{prior}} + \mathcal{L}_{\text{data}}. \tag{148}$$

Gradients used by SGLD From the streamed statistics we get closed-form gradients:

$$\nabla_{a_b} U = \frac{T}{\sigma_a^2}\, a_b \;-\; \frac{2\, s_f}{P}\, C_{1,b} \;+\; \frac{2\, s_f^2}{P}\, C_{2,b}\, a_b, \qquad \nabla_{w_b} U = G_b^{\text{data}} \;+\; T\,(\rho \odot \boldsymbol{w}_b). \tag{149}$$

These are the only quantities used inside the inner SGLD loop.

SGLD updates We apply Euler–Maruyama steps with isotropic Gaussian noise:

$$w_b \leftarrow w_b - \eta\, \nabla_{w_b} U + \sqrt{2T\eta}\, \xi_{w_b}, \qquad a_b \leftarrow a_b - \eta\, \nabla_{a_b} U + \sqrt{2T\eta}\, \xi_{a_b}, \tag{150}$$

where $\xi_{w_b} \sim \mathcal{N}(0, I_d)$ and $\xi_{a_b} \sim \mathcal{N}(0, 1)$. We use polynomial decay on η always matching the SGLD-trained full NNs.

ARD update The ARD update is:

$$\alpha_{\text{post}} = \alpha_0 + \tfrac{B}{2}, \qquad \beta_{\text{post}} = \beta_0 + \tfrac{1}{2} \sum_{b=1}^{B} \|w_b\|_2^2, \qquad \rho \leftarrow (1-\lambda)\, \rho + \lambda\, \frac{\alpha_{\text{post}}}{\beta_{\text{post}}}. \tag{151}$$

Fixed-point view and the K inner steps The outer loop implements a fixed-point iteration on the network f. Writing $r = y - f$, define the map $\mathcal{G}_K$ as: (i) run K inner SGLD steps on $(\boldsymbol{w}_b, a)$ using the current residual r; (ii) recompute $f^{\text{new}}(x_p) = s_f \sum_b a_b\, \phi(w_b^\top x_p)$. As $K \to \infty$ and $\eta \to 0$ the inner Markov chain approaches its stationary law, and the iteration solves the MF FP equations described in the theory sections.

Algorithm 1 Simple streaming SGLD for (a, w) with optional ARD

Require: (X, y); $B, N, \gamma, \sigma_a, \sigma_w, \kappa$; activation ϕ; steps T_{out}, inner steps K, step size η; optional ARD $(\alpha_0, \beta_0, \lambda)$
Ensure: Final particles $\{(w_b, a_b)\}_{b=1}^{B}$ and predictor $f(x) = s_f \sum_b a_b \phi(w_b^\top x)$
0: **Init:** $w_b \sim \mathcal{N}(0, \sigma_w^2 I_d/d)$, $a_b \sim \mathcal{N}(0, \sigma_a^2)$; set $\rho \leftarrow d/\sigma_w^2$; set $T \leftarrow 2\kappa^2$
0: **for** $t = 1..T_{\text{out}}$ **do**
0: compute $f_p = s_f \sum_b a_b \phi(x_p^\top w_b)$ and residuals $r_p = y_p - f_p$
0: **for** $k = 1..K$ **do**
0: compute $\{C_{1,b}, C_{2,b}, G_b^{\text{data}}\}$ via formulas above
0: Form $\nabla_{a_b} U, \nabla_{w_b} U$; update $w_b \leftarrow w_b - \eta \nabla_{w_b} U + \sqrt{2T\eta}\, \xi_{w_b}$, $a_b \leftarrow a_b - \eta \nabla_{a_b} U + \sqrt{2T\eta}\, \xi_{a_b}$
0: refresh f_p, r_p after the last inner step
0: **end for**
0: ARD update ρ: $\alpha_{\text{post}} = \alpha_0 + \frac{B}{2}$, $\beta_{\text{post}} = \beta_0 + \frac{1}{2} \sum_b \|w_b\|^2$, $\rho \leftarrow (1-\lambda)\rho + \lambda\, \alpha_{\text{post}}/\beta_{\text{post}}$
0: **end for**=0

Empirical calculation of m_S and generalisation error Let the teacher be a single Walsh mode χ_S, so $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$. On a held-out set $\{x_\mu\}_{\mu=1}^{P_{\text{eval}}}$ define the vector $c \in \mathbb{R}^{P_{\text{eval}}}$ by $c_\mu = \chi_S(\boldsymbol{x}_\mu)$, the empirical Gram scalar

$$g \;=\; \frac{1}{P_{\text{eval}}}\, c^\top c, \tag{152}$$

and the (scalar) empirical overlap

$$m_S \;=\; \frac{1}{P_{\text{eval}}} \sum_{\mu=1}^{P_{\text{eval}}} \chi_S(\boldsymbol{x}_\mu)\, f(x_\mu) \;=\; \frac{1}{P_{\text{eval}}}\, c^\top f. \tag{153}$$

Let $\bar{f}^2 = \frac{1}{P_{\text{eval}}} \sum_\mu f(\boldsymbol{x}_\mu)^2$. Then the empirical test MSE decomposes as

$$\widehat{\mathcal{E}}_{\text{test}} = \frac{1}{2P_{\text{eval}}} \sum_{\mu=1}^{P_{\text{eval}}} \big(f(\boldsymbol{x}_\mu) - \chi_S(\boldsymbol{x}_\mu)\big)^2 = \underbrace{\tfrac{1}{2}(1 - m_S)^2}_{\text{mode term}} \;+\; \underbrace{\tfrac{1}{2}\big(\bar{f}^2 - 2\, m_S^2 + g\, m_S^2\big)}_{\text{noise / orthogonal term}} \tag{154}$$

$$= \frac{1}{2}\Big(\bar{f}^2 - 2\, m_S + g\Big), \tag{155}$$

where the second line is the direct empirical expression. When the Walsh basis is orthonormal on the evaluation set ($g = 1$), this reduces to $\widehat{\mathcal{E}}_{\text{test}} = \frac{1}{2}(\bar{f}^2 - 2m_S + 1)$.

F. Training details (hyperparameters)

Here, we present the hyperparameters for SGLD and MF-ARD for the different figures. The hyperparameters for Figure 6 are specified below.

- Data in Figure 1 is a slice of Figure 6 for $\kappa = 5 \cdot 10^{-3}$.
- Data in Figure 5 is a slice of Figure 6 for $\kappa = 5 \cdot 10^{-3}$.
- **??** are averages over 3 trained models from the data of Figure 1 for $\kappa = 5 \cdot 10^{-3}$.

Hyperparameter	**SGLD**
d (input dimension)	35
P (train set sizes)	$\{10,\ 100,\ 500,\ 750,\ 1000,\ 2133,\ 3666,\ 5000,\ 7500,\ 10000\}$
κ values	$\{5{\times}10^{-4},\ 10^{-3},\ 5{\times}10^{-3},\ 7.5{\times}10^{-3},\ 10^{-2},\ 5{\times}10^{-2},\ 10^{-1}\}$
E (experiments / config)	3
teacher set S	$\{0,1,2,3\}$
data distribution	$X\in\{-1,+1\}^d,\ y=\prod_{j\in S} X_j$
activation	ReLU
N (hidden units)	512
γ (scaling exponent)	0.5
$g_w,\ g_a$ (prior variances)	$1.0,\ 1.0$
initialization	$w\sim\mathcal{N}(0,g_w/d),\ \ a\sim\mathcal{N}(0,g_a)$
temperature T	$2\,\kappa^2$
epochs (max)	7,500,000
batch size	full-batch
loss	mean MSE
learning rate η (final)	$5{\times}10^{-4}$
start LR η_{start}	$5{\times}10^{-3}$
LR scheduler	polynomial (power 2): $\eta_{\text{start}}\to\eta$ over $2{\times}10^6$ steps

Table 1. Algorithm outlined in Section A.1. Hyperparamters for Figure 6 **a)**.

Hyperparameter	**MF-ARD**
d (input dimension)	35
P (train set sizes)	$\{10,100,500,750,1000,2133,3666,5000,7500,10000\}$
κ values	$\{5{\times}10^{-4},\ 10^{-3},\ 5{\times}10^{-3},\ 7.5{\times}10^{-3},\ 10^{-2},\ 5{\times}10^{-2},\ 10^{-1}\}$
E (experiments / config)	3
teacher set S	$\{0,1,2,3\}$
data distribution	$X\in\{-1,+1\}^d,\ y=\prod_{j\in S} X_j$
activation	ReLU
B (particles)	512
N	512
γ	0.5
$\sigma_a,\ \sigma_w$	$1.0,\ 1.0$
initialization	$w\sim\mathcal{N}(0,\sigma_w^2/d),\ a\sim\mathcal{N}(0,\sigma_a^2)$
outer steps	7,500,000
learning rate scheduler	poly-2 decay: $5\times10^{-3}\to5{\times}10^{-4}$ over $2{\times}10^6$ steps
SGLD inner steps K	$K_0{=}12\ \to\ K_{\min}{=}2$ (decay over $6{\times}10^5$ steps)
temperature T	$2\kappa^2$
ARD	on: $\alpha_0{=}4.0$, $\beta_0{=}4/35$, EMA 0.5, $\rho\in[0,10^{18}]$

Table 2. Algorithm outlined in Section E.1. Hyperparamters for Figure 6 **b)** with ARD disabled and **c)** with ARD enabled.

For the single index model we introduced a bias vector in the architecture.

Hyperparameter	SGLD (Hermite single-index)
d (input dimension)	18
P (train set sizes)	$\{75{,}000, 50{,}000, 25{,}000, 10{,}000, 5{,}000, 1{,}000, 100, 50\}$
κ values	$\{10^{-5}, 10^{-4}, 10^{-3}10^{-2}, 10^{-1}\}$
E (experiments / config)	4
teacher type	single-index Hermite
Hermite degree p	4
support size k	2 (random per experiment)
data distribution	$X \sim \mathcal{N}(0, I_d)$
teacher w	$w_i = \frac{1}{\sqrt{k}}$ on support, else 0
labels	$y = \mathrm{He}_p(Xw)$
activation	ReLU
N	1024
γ	0.5
$\sigma_a,\ \sigma_w,\ \sigma_b$	$1.0,\ 0.5,\ 1.0$
initialization	$w \sim \mathcal{N}(0, \sigma_w^2/d),\ a \sim \mathcal{N}(0, \sigma_a^2),\ b \sim \mathcal{N}(0, \sigma_b^2)$
temperature T	$2\,\kappa^2$
epochs (max)	4,000,000
batch size	full-batch
loss	mean MSE
learning rate η (final)	5×10^{-4}
start LR η_{start}	2×10^{-3}
LR scheduler	polynomial (power 2): $\eta_{\text{start}} \to \eta$ over 2×10^6 steps

Table 3. Algorithm outlined in Section A.1. Hyperparamters for Figure 9 **a)**.

Hyperparameter	MF-ARD (Hermite single-index)
d (input dimension)	18
P (train set sizes)	$\{75{,}000, 50{,}000, 25{,}000, 10{,}000, 5{,}000, 1{,}000, 100, 50\}$
κ values	$\{10^{-5}, 10^{-4}, 10^{-3}10^{-2}, 10^{-1}\}$
E (experiments / config)	4
teacher type	single-index Hermite
Hermite degree p	4
support size k	2 (random per experiment)
data distribution	$X \sim \mathcal{N}(0, I_d)$
teacher w	$w_i = \frac{1}{\sqrt{k}}$ on support, else 0
labels	$y = \mathrm{He}_p(Xw)$
activation	ReLU
B (particles)	1024
N	1024
γ	0.5
$\sigma_a,\ \sigma_w,\ \sigma_b$	$1.0,\ 0.5,\ 1.0$
initialization	$w \sim \mathcal{N}(0, \sigma_w^2/d),\ a \sim \mathcal{N}(0, \sigma_a^2),\ b \sim \mathcal{N}(0, \sigma_b^2)$
outer steps	4,000,000
learning rate scheduler	poly-2 decay: $2 \times 10^{-3} \to 5\times10^{-4}$ over 2.5×10^6 steps
SGLD inner steps K	$K_0{=}12\ \to\ K_{\min}{=}2$ (decay over 6×10^5 steps)
temperature T	$2\kappa^2$
ARD	on: $\alpha_0{=}0.1$, $\beta_0{=}\alpha_0/d{=}0.1/18$, EMA 0.5, $\rho \in [0, 10^{18}]$

Table 4. Algorithm outlined in Section E.1. Hyperparamters for Figure 9 **b)** with ARD enabled.