Weight Anisotropy in Mean-Field Theory: Learning on Isotropic Data

Niclas Göring NICLAS.GORING@JESUS.OX.AC.UK
Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, United Kingdom

Yoonsoo Nam
Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, United Kingdom

Chris Mingard
Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, United Kingdom

Jake Reid
Experimental Psychology, University of Oxford, Oxford, United Kingdom

Ard A. Louis
Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, United Kingdom

Abstract

Neural networks efficiently learn isotropic data distributions with low-dimensional target structure where fixed kernel limits fail. We trace this advantage to input feature selection (IFS): networks develop strong weight anisotropy along task-relevant coordinates. While standard Mean-Field (MF) theory captures the onset of feature learning, it tracks only first moments and thus misses IFS and underestimates post-transition generalisation. We introduce MF-ARD, augmenting MF with a single additional set of order parameters for coordinate-wise precisions. MF-ARD successfully captures the sharp generalisation transitions of finite-width networks.

1. Introduction

While finite-width networks can discover low-dimensional targets in isotropic data via feature learning (FL) [7, 11, 12], fixed kernels [14, 17] suffer from the curse of dimensionality. Statistical physics and Mean-Field (MF) theory provide a rigorous framework for modeling this FL regime [3, 5, 9, 10, 13, 16, 19–22]. However, as outlined above, standard MF structurally misses **input feature selection (IFS)** because it lacks the second-moment tracking required to model coordinate-dependent regularisation. We resolve this by proposing **MF-ARD**, which integrates automatic relevance determination [18, 27] into the MF framework. By elevating coordinate-wise precisions to self-consistent order parameters, MF-ARD explicitly models the strong weight anisotropy observed in SGLD-trained networks. This minimal extension is sufficient to quantitatively predict the sharp generalisation transitions that standard MF underestimates. Extended related work is deferred to Section A.

2. Background: Mean field theory

2.1. SGLD setup

For analytic tractability, we use SGLD, the limit of full-batch GD with weight decay and injected Gaussian noise.([26, 29], see Section C.1 for details). Given an NN described by $f_{\boldsymbol{\theta}}(\boldsymbol{x}_\mu)$, a dataset

© N. Göring, Y. Nam, C. Mingard, J. Reid & A.A. Louis.

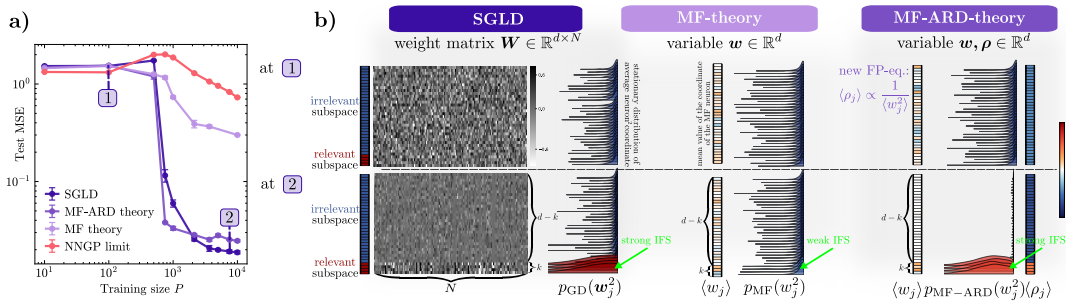

Figure 1: **a)** Test MSE vs. training-set size P. MF-ARD successfully captures the sharp generalization transition of SGLD, which standard MF theory underestimates and the NNGP limit misses entirely. Points $\boxed{1}$ and $\boxed{2}$ mark pre- and post-transition states. **b)** Stationary distributions of the squared weights, $p(w_j^2)$. At $\boxed{2}$, **SGLD** exhibits strong Input Feature Selection (IFS): the k relevant coordinates (red) develop large variances, while the $d-k$ irrelevant ones (blue) remain suppressed. Standard **MF theory** tracks only coordinate means, yielding weak IFS. By introducing a fixed-point equation for coordinate variances, **MF-ARD** accurately reproduces the strong variance separation in $p(w_j^2)$ observed in SGLD, explaining the post-transition performance gains.

$\mathcal{D} = \{(\boldsymbol{x}_\mu, \boldsymbol{y}_\mu)\}_{\mu=1}^P$, 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: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{N^\gamma}\sum_{i=1}^N a_i\,\phi(\boldsymbol{w}_i^\top\boldsymbol{x})$, $\quad w_{ij}\sim\mathcal{N}\Big(0, \frac{\sigma_w^2}{d}\Big)$, $\quad a_i \sim \mathcal{N}(0,\sigma_a^2)$. 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). Using the explicit form of $f_{\boldsymbol{\theta}}$ to rewrite the stationary distribution yields (see Section C.2):

SGLD-posterior

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

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.

2.2. MF Theory: Removing off-diagonal couplings

To make the posterior in Equation (2) tractable, we use self-consistent MF theory (related to the MF theory in [22, 23]), which replaces the correlated posterior $p(\boldsymbol{W}|\mathcal{D})$ with a factorised 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 $\langle f \rangle$: $\sum_{i' \neq i} a_{i'}\, G(\boldsymbol{w}_i, \boldsymbol{w}_{i'}) \overset{\text{MF}}{\approx} \frac{1}{P} \sum_{\mu} \phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)\, N^{\gamma} \langle f(\boldsymbol{x}) \rangle$. 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)

$$-\ln p_{\text{MF}} = \frac{a^2}{2\sigma_a^2} + \frac{d}{2\sigma_w^2} \sum_{j=1}^{d} w_j^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}} \Sigma(\boldsymbol{w}) - \frac{a}{\kappa^2 N^{\gamma}} \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w}) \right) \tag{3}$$

$$m_A = N^{1-\gamma} \langle a J_A(\boldsymbol{w}) \rangle_{p_{\text{MF}}} \quad \forall\, A,$$

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

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 G).

3. Theory I: Input Feature Selection as Implicit Regularisation

We focus on k-sparse parity, $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ with $S = \{1, \ldots, k\}$, as the canonical isotropic test case (Section A). Prior work [1, 2, 4, 24] established that weight-matrix anisotropy is central for generalisation; we give a simplified analytical derivation showing how it arises during SGLD training and shapes the stationary distribution. Let $\boldsymbol{w}_j(t) \in \mathbb{R}^N$ denote the j-th row of $\boldsymbol{W}(t) \in \mathbb{R}^{d \times N}$ and write $\boldsymbol{w}_j^2(t) := \|\boldsymbol{w}_j(t)\|_2^2$. Applying Itô's lemma to (1) yields the *exact* evolution

$$\tfrac{d}{dt} \langle \boldsymbol{w}_j^2 \rangle = -2\eta \big(\lambda \langle \boldsymbol{w}_j^2 \rangle + \langle \hat{\rho}_j\, \boldsymbol{w}_j^2 \rangle \big) + 2T\eta N. \tag{4}$$

$\hat{\rho}_j(t) := \frac{\langle \boldsymbol{w}_j, \nabla_{\boldsymbol{w}_j} \ell \rangle}{\boldsymbol{w}_j^2}$ acts as a coordinate-wise effective decay: when $\hat{\rho}_j < 0$, the descent direction aligns with $\boldsymbol{w}_j$, partially cancelling λ. On sparse tasks, $\hat{\rho}_j(t)$ separates sharply (Figure 5d): for $j \in S^c$ the gradient carries no consistent component along $\boldsymbol{w}_j$, so $\hat{\rho}_j \approx 0$ and the row experiences full decay; for $j \in S$ the loss induces systematic alignment ($\hat{\rho}_j < 0$) prior to the test-error drop, providing an anti-decay contribution that lets $\boldsymbol{w}_j^2$ persist or grow. Setting $\frac{d}{dt} \langle \boldsymbol{w}_j^2 \rangle = 0$ in (4) gives the exact balance $\lambda\, \langle \boldsymbol{w}_j^2 \rangle_\infty + \langle \hat{\rho}_j\, \boldsymbol{w}_j^2 \rangle_\infty = TN$, so a separation in $\langle \hat{\rho}_j \rangle_\infty$ between S and S^c produces an anisotropic stationary distribution (Figure 5b, Figure 1a).

Definition 1 (Input feature selection) *A network undergoes IFS if, as epochs $\to \infty$, the weight-anisotropy ratio satisfies:* $R_{\boldsymbol{W}}^{\infty} = \frac{\sqrt{d-k}\, \langle \boldsymbol{w}_{j \in S}^2 \rangle_\infty}{\sqrt{k}\, \langle \boldsymbol{w}_{j \in S^c}^2 \rangle_\infty} \gg 1.$

Section D.1 confirms empirically that low test error tightly correlates with weight concentration in the relevant subspace (Figure 8).

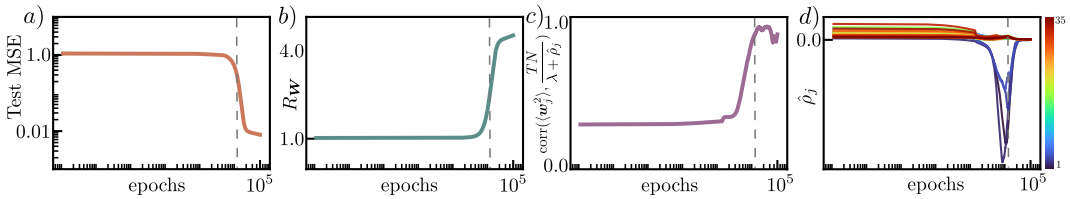

Figure 2: A 1-hidden layer ReLU network ($N = 512$) is trained on $k = 4$ sparse parity in $d = 35$. **a)** Test error versus epochs. **b)** The weight anisotropy ratio R_W increases sharply coincident with the drop in test error. **c)** Validation of the stationary energy balance (3), comparing measured $\boldsymbol{w}_j^2$ 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).

4. Theory II: MF model of IFS

4.1. FL in the simple MF model

MF theory interprets FL as *self-consistent symmetry breaking* in p_{MF}. Below (P_c, κ_c), the only stable FP of Equation (3) is $m_A = 0\ \forall A$ and p_{MF} collapses to the Gaussian prior, regularised toward an isotropic high-entropy state by the prior and self-energy $\Sigma(\boldsymbol{w})$. Above (P_c, κ_c), the neuron-data coupling $J_{\mathcal{Y}}$ acts as an external field that rewards alignment of $\phi(\boldsymbol{w}^\top \boldsymbol{x})$ with $y(\boldsymbol{x})$, destabilising $m_A = 0$ (see Theorem 11 for the explicit κ_c). For $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ the order parameter is m_S, and the FP equations (3) reduce to (Section F.5) $m_S = \frac{N^{1-2\gamma}}{\kappa^2}(1 - m_S)\left\langle \frac{J_S(\boldsymbol{w})^2}{\sigma_a^{-2} + \Sigma(\boldsymbol{w})/(\kappa^2 N^{2\gamma})} \right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)}$. The transition occurs when $\omega_0 > 1$ (Figure 6): SGLD and MF stay at $m_S = 0$ until P_c then jump abruptly, while NNGP rises only smoothly.

4.2. Problem: simple MF underestimates IFS

Figure 1a reveals the gap: post-transition, SGLD generalises far faster than MF. Standard MF parameterises $p_{\text{MF}}(\boldsymbol{w})$ through *first-moment* order parameters $\{m_A\}$ only, so the prior term $\frac{d}{2\sigma_w^2}\sum_j w_j^2$ assigns equal precision to every j regardless of how $\langle w_j^2 \rangle_{p_{\text{MF}}}$ varies between S and S^c. Tracking $\{m_A\}$ predicts the *onset* of the transition (Figure 6) but not the post-transition gain, which Section 3 traces to symmetry breaking in the second moments $\langle w_j^2 \rangle$. We formalise this obstruction:

Theorem 2 *Let $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ and $\{w_j\}_{j=1}^d$ a solution of (3) with $\boldsymbol{w}_S = \{w_j\}_{j\in S}$, $\boldsymbol{w}_{S^c} = \{w_j\}_{j\in S^c}$. The plain-MF FP cannot develop strong weight anisotropy: $R_{\boldsymbol{w}}^{MF} = \sqrt{d-k}\,\|\boldsymbol{w}_S\|^2 / (\sqrt{k}\,\|\boldsymbol{w}_{S^c}\|^2) = 1 + O(1/d)$. See Section F.6 for the proof.*

4.3. A minimal MF model of IFS: tracking coordinate variances

We extend MF to track the row-wise variances w_j^2 as explicit order parameters. The goal is to match the SGLD stationary law (3): assuming the gradient-weight alignment behaves 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 the MF independence assumption $\langle \boldsymbol{w}_j^2 \rangle_\infty \overset{\text{MF}}{\approx} N \langle w_j^2 \rangle_{\text{MF}}$, (3) reduces to

$$\lambda \, \langle w_j^2 \rangle_{\text{MF}} + \rho_j^r \, \langle w_j^2 \rangle_{\text{MF}} = T, \tag{5}$$

i.e. the likelihood gradient acts as an effective potential $\frac{\rho_j^r}{2} \|\boldsymbol{w}_j\|_2^2$ in coordinate j (linear response; see Figure 5c for empirical evidence). We therefore enlarge the state space minimally: $\{(w_j, m_A)\}_{j=1}^d \to \{(w_j, m_A), (w_j^2, \rho_j)\}_{j=1}^d$.

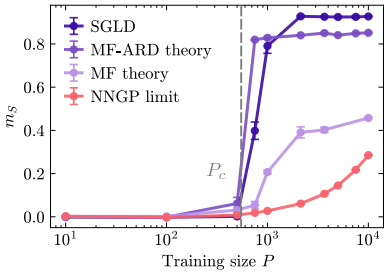

Figure 3: Feature coefficient m_S vs. P for a k-sparse parity target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x}) = \Pi_{j \in S} x_j$ with index $S = \{0, 1, 2, 3\}$ in $d = 35$ with a ReLU network ($N = 512$, see Section C.3; Section H for more details). SGLD, MF and MF-ARD exhibit a phase transition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

4.4. Automatic Relevance Determination MF theory

We endow $p_{\text{MF}}(\boldsymbol{w})$ with a coordinate-wise precision ρ_j via the conditional Gaussian $p(w_j \mid \rho_j) = \mathcal{N}(0, \rho_j^{-1})$, automatic relevance determination (ARD) in Bayesian NNs [18, 27], with a conjugate Gamma prior on ρ_j.

Theorem 3 *Assume $p(\boldsymbol{w}_j \mid \rho_j) = \mathcal{N}(0, \rho_j^{-1}\mathbb{I})$ and $p(\rho_j) = \Gamma(\alpha_0, \beta_0)$. Enforcing $\rho_j \to d/\sigma_w^2$ when the data signal vanishes fixes $\beta_0 = \alpha_0 \sigma_w^2 / d$, giving the self-consistent FP $\rho_j = \left(\alpha_0 + \frac{N}{2}\right) / \left(\frac{\alpha_0}{d} + \frac{N}{2} \langle w_j^2 \rangle_{p_{\text{ARD}}}\right)$. See Section C.7 for the proof.*

MF-ARD theory (fixed point equations)

$$-\ln p_{\text{ARD}} = \frac{a^2}{2\sigma_a^2} + \frac{1}{2} \sum_{j=1}^{d} \rho_j w_j^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}} \Sigma(\boldsymbol{w}) - \frac{a}{\kappa^2 N^\gamma} \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w}) \right) \tag{6}$$

$$m_A = N^{1-\gamma} \langle a J_A(\boldsymbol{w}) \rangle_{p_{\text{ARD}}} \; \forall \, A \qquad \rho_j = \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2} \langle w_j^2 \rangle_{p_{\text{ARD}}}} \tag{7}$$

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

The map $\langle w_j^2 \rangle \mapsto \rho_j \mapsto p_{\text{ARD}}$ realises the IFS feedback: relevant coordinates acquire larger $\langle w_j^2 \rangle$, which lowers ρ_j via (7) and further reduces shrinkage; irrelevant coordinates do the opposite. Neurons remain independent, but the enlarged order-parameter set now captures SGLD's coordinate-dependent variance structure.

4.5. Connection of MF-ARD to IFS in SGLD

At stationarity, MF-ARD matches the SGLD distribution in both first moments (inherited from MF) and second moments up to $O(\alpha_0/N)$:

Theorem 4 *Under the linear-response assumption* $\langle \hat{\rho}_j \, \boldsymbol{w}_j^2 \rangle_\infty \approx \rho_j^r \langle \boldsymbol{w}_j^2 \rangle_\infty$*, defining* $\rho_{\text{eff},j} := (\lambda + \rho_j^r)/T$*, and assuming* $\alpha_0 \ll N$*:* SGLD: $\rho_{\text{eff},j} = \frac{N}{\langle \boldsymbol{w}_j^2 \rangle_\infty}$, MF-ARD: $\rho_j = \frac{1}{\langle w_j^2 \rangle_{p_{\text{MF-ARD}}}} \big(1 + O(\alpha_0/N)\big)$*. See Section F.7 for the proof.*

We thus identify ρ_j with the total effective SGLD precision (the factor N is absorbed by (5), and $\alpha_0 \ll N$ throughout our experiments). Relevant features acquire strong negative alignment and large weights; irrelevant ones retain high precision, yielding an anisotropic stationary distribution from isotropic inputs, captured explicitly by MF-ARD's static FP.

5. Results

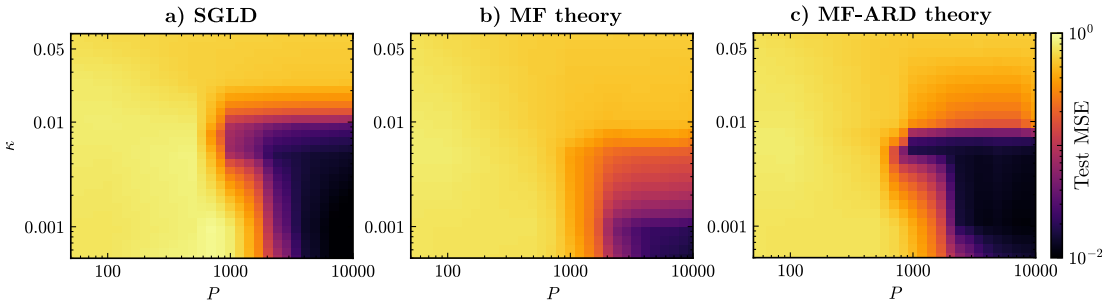

Figure 4: **Generalisation phase diagram (Test MSE) vs. dataset size P and noise κ for k-sparse parity target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ with $S = \{0, 1, 2, 3\}$ in $d = 35$. a) SGLD, b) MF, c) MF-ARD:** All panels show a transition from high to low test error as P increases or κ decreases. 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 H.

Figure 4 shows test-MSE heatmaps over (P, κ) for SGLD, plain MF, and MF–ARD on k-sparse parity. Plain MF detects *when* learning starts but yields a diffuse, shifted boundary; MF–ARD, with the single added set $\{\rho_j\}$, tracks both the location and sharpness of the SGLD boundary and reproduces the "helpful noise" kink near $\kappa = 0.05$. The same picture holds for a single-index Hermite target [7] (Figure 10, Appendix).

Limitations. MF–ARD's diagonal precision restricts feature selection to axis-aligned subspaces; rotated targets would require a low-rank covariance extension. The analysis uses the SGLD stationary distribution and is currently limited to two-layer networks; minibatch SGD and deeper architectures are left to future work.

References

[1] Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, and Christopher Marquis. 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*, pages 33–52. PMLR, June 2022. URL https://proceedings.mlr.press/v162/abbe22a.html. ISSN: 2640-3498.

[2] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. 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].

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

[4] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. 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].

[5] P Baglioni, R Pacelli, R Aiudi, F Di Renzo, A Vezzani, R Burioni, and P Rotondo. 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.

[6] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

[7] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks, 2022. URL http://arxiv.org/abs/2210.15651.

[8] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35: 32240–32256, 2022.

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

[10] Michael Celentano, Chen Cheng, and Andrea Montanari. 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.

[11] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html. ISSN: 2640-3498.

[12] Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

[13] Jessica N Howard, Ro Jefferson, Anindita Maiti, and Zohar Ringel. Wilsonian renormalization of neural network gaussian processes. *Machine Learning: Science and Technology*, 6(2): 025038, 2025.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[15] Yiwen Kou, Zixiang Chen, Quanquan Gu, and Sham M. Kakade. 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].

[16] Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks. *arXiv preprint arXiv:2502.07998*, 2025.

[17] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes, 2018. URL https://arxiv.org/abs/1711.00165.

[18] David JC MacKay. Bayesian non-linear modeling for the prediction competition. In *Maximum Entropy and Bayesian Methods: Santa Barbara, California, USA, 1993*, pages 221–234. Springer, 1996.

[19] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

[20] Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

[21] Rosalba Pacelli, Sebastiano Ariosto, Mauro Pastore, Francesco Ginelli, Marco Gherardi, and Pietro Rotondo. A statistical mechanics framework for bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12):1497–1507, 2023.

[22] Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of statistical field theory in deep learning, 2025. URL https://arxiv.org/abs/2502.18553.

[23] Noa Rubin, Inbar Seroussi, and Zohar Ringel. 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.

[24] Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

[25] Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. 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.

[26] Yee Whye Teh, Alexandre Thiéry, and Sebastian Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics, 2015. URL https://arxiv.org/abs/1409.0578.

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

[28] Alexander van Meegen and Haim Sompolinsky. 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.

[29] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, ICML'11, pages 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.

LLM Usage

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

Appendix A. Related work: Data-dependent feature learning and mean-field theories

We focus on k-sparse parity and sparse multi-index models because these isotropic tasks provide the sharpest stress test for feature learning (FL): the input distribution carries no anisotropy that would help a fixed kernel, so NTK/NNGP methods pay a sample-complexity exponent in d while finite-width networks do not [6, 11, 12, 15]. A rich line of gradient-based analyses proves that networks *do* recover the relevant subspace on such tasks [2, 4, 6, 11], but these results establish *that* learning happens rather than *predicting* generalisation error across the full (P, κ) plane. Our goal is a tractable theory whose fixed points predict the complete phase diagram (Figure 4), including the location and sharpness of the P_c transition. Several mean-field-style frameworks go beyond the kernel regime (Table A): classical MF and its Langevin extensions [19, 25]; DMFT-style theories tracking $O(P^2T^2)$ kernel order parameters [8, 9, 16]; DMFT in the proportional limit for single-index models ($k = 1$, with $k \geq 2$ deferred) [20]; and Bayesian/replica approaches [21, 28]. Despite their power, none has demonstrated a sample-complexity transition P_c on high-dimensional isotropic tasks (Table A). Two barriers explain this. First, DMFT tracks $O(P^2T^2)$ order parameters requiring Monte Carlo solvers [8, 16], which becomes prohibitive at the dataset sizes ($P \sim 10^3$–10^4) needed to resolve the transition on sparse parity. Second, each existing closure is organised around the observable natural to its limit—sample–time kernels, latent-subspace projections, or replica Q-matrices—not around the input-coordinate row energies $\{\|W_{j,:}\|_2^2\}_{j=1}^d$ that govern feature selection on isotropic inputs. MF–ARD addresses both gaps: it augments standard MF with only d coordinate-wise precisions $\{\rho_j\}$, the minimal closure that exposes IFS while remaining valid for any P, N. We use ARD [18, 27] as the mechanism that turns coordinate-wise second moments into self-consistent order parameters, not as a pruning heuristic.

Theory	**# OP**	**Regime**	**Data tested**	P_c **on isotropic**
Rubin et al. [23]	$O(1)$	Bayesian, $N \to \infty$	cubic, mod. add.	No
Bordelon and Pehlevan [8]	$O(P^2T^2)$	$T, P = O_N(1)$	small CIFAR, Gaussian	No
Lauditi et al. [16]	$O(P^2T^2)$	$N \to \infty$, μP	CIFAR subsets, Gaussian	No
Montanari and Urbani [20]	$O(T^2)$	$n/d \to \alpha$, $m \to \infty$	$k = 1$ single-index Gaussian	$k \geq 2$ deferred
Pacelli et al. [21]	$O(1)$	Bayesian, $P/N \to \alpha$	MNIST, CIFAR (proportional)	No
van Meegen and Sompolinsky [28]	$O(P)$	Bayesian, $N \to \infty$	MNIST, CIFAR	No
MF-ARD (ours)	$\boldsymbol{O(d)}$	**any** P, N	**high-**d **isotropic (**k**-parity,** k**-index)**	**Yes (Fig. 6)**

Table 1: Mean-field-style theories of feature learning. "# OP" = order-parameter count; "P_c on isotropic" = whether a sample-complexity transition on a high-d isotropic task has been demonstrated.

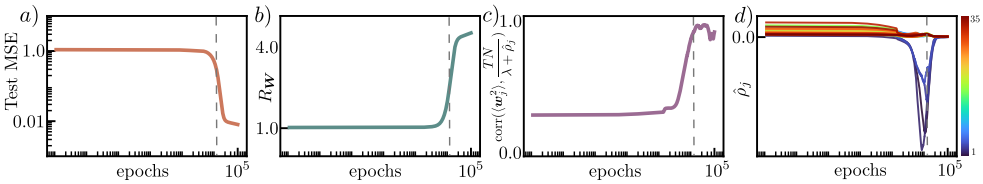

Figure 5: A 1-hidden layer ReLU network ($N = 512$) is trained on $k = 4$ sparse parity in $d = 35$. **a)** Test error versus epochs. **b)** The weight anisotropy ratio R_W increases sharply coincident with the drop in test error. **c)** Validation of the stationary energy balance (3), comparing measured $\boldsymbol{w}_j^2$ 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).

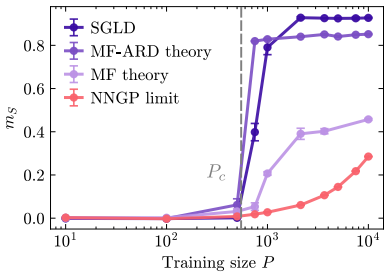

Figure 6: Feature coefficient m_S vs. P for a k-sparse parity target $y(\boldsymbol{x}) = \chi_S(\boldsymbol{x}) = \Pi_{j\in S} x_j$ with index $S = \{0, 1, 2, 3\}$ in $d = 35$ with a ReLU network ($N = 512$, see Section C.3; Section H for more details). SGLD, MF and MF-ARD exhibit a phase transition at P_c, NNGP grows smoothly, no phase-transition or FL.

Appendix B. Main text figures

Appendix C. Additional Background

C.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:

$$f(\boldsymbol{x}) = \frac{1}{N^\gamma} \sum_{i=1}^{N} a_i \phi(\boldsymbol{w}_i^\top \boldsymbol{x}). \tag{8}$$

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):

$$\Delta\theta_{i,t} := \theta_{i,t+1} - \theta_{i,t} \tag{9}$$

$$= -\eta\big(\gamma\,\theta_{i,t} + \nabla_{\theta_i}\ell(f_{\boldsymbol{\theta}})\big) + \sqrt{2T\eta}\;\xi_t\,. \tag{10}$$

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:

$$p(\boldsymbol{\theta}_i) \propto \exp\Big(-\frac{1}{T}\Big(\frac{\gamma}{2}\|\theta_i\|^2 + \ell(f_{\boldsymbol{\theta}})\Big)\Big) \tag{11}$$

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:

$$p(\boldsymbol{w}_i, a_i) \propto \exp\left(-\frac{1}{T}\left(\frac{\gamma_w}{2}\|\boldsymbol{w}_i\|^2 + \frac{\gamma_a}{2}\|a_i\|^2 + \frac{1}{P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\right)\right). \tag{12}$$

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: $p(\boldsymbol{\theta}) = \prod_i p(\boldsymbol{w}_i)p(a_i) \propto \exp\Big(-\sum_i\Big(\frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2 + \frac{1}{2g_a^2}\|a$
- 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: $p(\mathcal{D}|\boldsymbol{\theta}) \propto \exp\Big(-\frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\Big)$.

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:

$$p(\boldsymbol{w}_i, a_i|\mathcal{D}) \propto \exp\left(-\left(\frac{1}{2g_w^2}\|\boldsymbol{w}_i\|^2 + \frac{1}{2g_a^2}\|a_i\|^2 + \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2\right)\right). \tag{13}$$

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:

$$\Delta a_{t,i} = -\eta\left(\frac{T}{g_a^2}a_{t,i} + \nabla_{a_i}\left(\frac{1}{P}\sum_\mu \ell_\mu\right)\right) + \sqrt{2T\eta}\;\xi_{t,a} \tag{14}$$

$$\Delta \boldsymbol{w}_{t,i} = -\eta\left(\frac{T}{g_w^2}\boldsymbol{w}_{t,i} + \nabla_{\boldsymbol{w}_i}\left(\frac{1}{P}\sum_\mu \ell_\mu\right)\right) + \sqrt{2T\eta}\;\xi_{t,w}. \tag{15}$$

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

C.2. Derivation of (2)

We start from the negative log posterior and the two–layer model in Eq. (2.1):

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

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

$$\begin{aligned} -\ln p_{\mathrm{L}} &= \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}\left(\frac{1}{N^\gamma}\sum_{i=1}^N a_i\phi_{i\mu} - y_\mu\right)^2 \\ &= \frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P}\left[\frac{1}{N^{2\gamma}}\sum_{i=1}^N\sum_{j=1}^N a_i a_j\,\phi_{i\mu}\phi_{j\mu} - \frac{2}{N^\gamma}\sum_{i=1}^N a_i\,\phi_{i\mu}y_\mu + y_\mu^2\right]. \end{aligned} \tag{16}$$

Introduce the dataset–averaged quantities

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

$$J_{\mathcal{Y}}(\boldsymbol{w}) := \frac{1}{P}\sum_{\mu=1}^{P}\phi(\boldsymbol{w}^\top \boldsymbol{x}_\mu)\,y_\mu. \tag{17}$$

Using $\sum_{i,j} = \sum_{i=j} + \sum_{i\neq j}$ in (16), we obtain

$$-\ln p_{\mathrm{L}} = \frac{1}{2\kappa^2 N^{2\gamma}} \left[\sum_{i=1}^{N} a_i^2 \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu}^2}_{\Sigma(\boldsymbol{w}_i)} + \sum_{i\neq j} a_i a_j \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu}\phi_{j\mu}}_{G(\boldsymbol{w}_i,\boldsymbol{w}_j)} \right] - \frac{1}{\kappa^2 N^{\gamma}} \sum_{i=1}^{N} a_i \underbrace{\frac{1}{P}\sum_{\mu} \phi_{i\mu} y_\mu}_{J_{\mathcal{Y}}(\boldsymbol{w}_i)} + \underbrace{\frac{1}{2\kappa^2 P}\sum_{\mu=1}^{P} y_\mu^2}_{\text{const.}}. \tag{18}$$

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

$$-\ln p_{\text{prior}} = \frac{1}{2\sigma_a^2}\sum_{i=1}^{N} a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^{N} \|\boldsymbol{w}_i\|^2. \tag{19}$$

Combining (18) and (19), and discarding the $\boldsymbol{\theta}$-independent constant, yields Eq. (2):

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

This derivation holds for any nonlinearity ϕ.

C.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.

C.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

$$\widehat{R}_S^{(a)} = \frac{\chi_S^\top K \chi_S}{\mathrm{tr}(K)} = \frac{\sum_{i=1}^N a_i^2\, J_S(\boldsymbol{w}_i)^2}{\sum_{i=1}^N a_i^2\, \Sigma(\boldsymbol{w}_i)}, \tag{20}$$

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.

C.5. The NNGP limit

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 (3) vanishes as well as any m_A-dependent tilt, resulting in $p(\boldsymbol{w})$ collapsing to its prior (see Section F.1 for a *proof*).

NNGP limit (fixed point equations)

$$\begin{aligned} -\ln p_\infty &= \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \\ m_A^\infty &= \frac{\sigma_a^2}{\kappa^2}\Big(\langle J_{\mathcal{Y}}(\boldsymbol{w}) J_A(\boldsymbol{w})\rangle_{p_\infty} \\ &\quad - \sum_B m_B^\infty \langle J_B(\boldsymbol{w}) J_A(\boldsymbol{w})\rangle_{p_\infty}\Big) \end{aligned} \tag{21}$$

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:

$$\begin{aligned} &\boldsymbol{m}^\infty = K\,(K + \tau \mathbb{1})^{-1}\boldsymbol{y}, \text{ where} \\ &\tau = \kappa^2/\sigma_a^2, \quad y_A = \mathbb{E}_{\boldsymbol{x}}[y(\boldsymbol{x})\chi_A(\boldsymbol{x})]. \end{aligned} \tag{22}$$

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 ($w_{ij} \leftrightarrow w_{ij'}$).

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 C.6 how this connects to the usual kernel formulation of the NNGP limit).

C.6. 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

$$K_{AB} = \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_B(\boldsymbol{w})], \tag{23}$$

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:

$$K_{AB} = \mathbb{E}_{\boldsymbol{w}}\left[\mathbb{E}_{\boldsymbol{x}}[\phi(\boldsymbol{w}^\top\boldsymbol{x})\chi_A(\boldsymbol{x})]\cdot\mathbb{E}_{\boldsymbol{x}'}[\phi(\boldsymbol{w}^\top\boldsymbol{x}')\chi_B(\boldsymbol{x}')]\right] \tag{24}$$

$$= \mathbb{E}_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\phi(\boldsymbol{w}^\top\boldsymbol{x})\phi(\boldsymbol{w}^\top\boldsymbol{x}')\chi_A(\boldsymbol{x})\chi_B(\boldsymbol{x}')\right] \tag{25}$$

$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\left[\chi_A(\boldsymbol{x})\chi_B(\boldsymbol{x}')\cdot\underbrace{\sigma_a^2\mathbb{E}_{\boldsymbol{w}}[\phi(\boldsymbol{w}^\top\boldsymbol{x})\phi(\boldsymbol{w}^\top\boldsymbol{x}')]}_{=:K(\boldsymbol{x},\boldsymbol{x}')}\right], \tag{26}$$

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

$$m = K(K+\tau I)^{-1}y, \qquad \tau = \kappa^2/\sigma_a^2. \tag{27}$$

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.

C.7. Derivation of the ARD precision fixed point

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

$$p(\boldsymbol{w}\mid\boldsymbol{\rho}) = \prod_{j=1}^d \mathcal{N}\big(w_j\mid 0,\rho_j^{-1}\big), \qquad p(\boldsymbol{\rho}) = \prod_{j=1}^d \Gamma(\rho_j\mid\alpha_0,\beta_0), \tag{28}$$

and the single-neuron MF–ARD action (Equation (6) in the main text)

$$-\ln p_{\text{ARD}}(\boldsymbol{w},a\mid\{m_A\},\boldsymbol{\rho},\mathcal{D}) = \frac{a^2}{2\sigma_a^2} + \frac{1}{2}\sum_{j=1}^d \rho_j\, w_j^2 + \frac{a^2}{2\kappa^2N^{2\gamma}}\Sigma(\boldsymbol{w}) - \frac{a}{\kappa^2N^\gamma}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w})\right). \tag{29}$$

For a width-N two-layer network, the joint action is the sum over neurons $i = 1,\ldots,N$ of Equation (29), and the (negative) log-evidence (free energy) for fixed $\{m_A\}$ is

$$\mathcal{F}(\boldsymbol{\rho};\{m_A\}) = -\ln Z(\boldsymbol{\rho};\{m_A\}) - \ln p(\boldsymbol{\rho}) \quad \text{with} \quad Z(\boldsymbol{\rho};\{m_A\}) = \int\prod_{i=1}^N d\boldsymbol{w}_i\, da_i\; e^{-\sum_{i=1}^N S_{\text{ARD}}(\boldsymbol{w}_i,a_i)}. \tag{30}$$

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

$$0 = \frac{\partial \mathcal{F}}{\partial \rho_j} = \underbrace{\Big\langle \frac{\partial}{\partial \rho_j} \sum_{i=1}^{N} S_{\mathrm{ARD}}(\boldsymbol{w}_i, a_i) \Big\rangle_{p_{\mathrm{ARD}}}}_{\text{energy term}} \underbrace{- \frac{N}{2}\frac{1}{\rho_j}}_{\text{Gaussian normalizer}} \underbrace{- \frac{\partial \ln p(\boldsymbol{\rho})}{\partial \rho_j}}_{\text{prior term}} = \frac{1}{2}\sum_{i=1}^{N} \langle w_{ij}^2 \rangle_{p_{\mathrm{ARD}}} - \frac{N}{2}\frac{1}{\rho_j} - \frac{\alpha_0 - 1}{\rho_j} + \beta_0, \tag{31}$$

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 (31) yields the closed-form FP update

$$\rho_j^\star \;=\; \frac{\alpha_0 - 1 + \frac{N}{2}}{\beta_0 + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}} \;\overset{\text{MAP corr.}}{\approx}\; \frac{\alpha_0 + \frac{N}{2}}{\beta_0 + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}}, \tag{32}$$

where the MAP conjugacy correction (absorbing the -1 in α_0) gives the form used in the main text (Equation (7)). With the scale-matching choice $\beta_0 = \alpha_0/d$, Equation (32) becomes

$$\rho_j^\star \;=\; \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\langle w_j^2 \rangle_{p_{\mathrm{ARD}}}}. \tag{33}$$

C.8. NNGP limit of the MF-ARD model

NNGP-limit of MF-ARD

$$\mathcal{S}_\infty^{\mathrm{FL}}(\mathbf{w}|\rho) = \frac{1}{2}\sum_{j=1}^{d} \rho_j w_j^2, \qquad p_{\infty,\rho}(\mathbf{w}) = \frac{1}{\mathcal{Z}} \exp\big(-\mathcal{S}_\infty^{\mathrm{FL}}(\mathbf{w}|\rho)\big), \tag{34}$$

$$m_A^\infty(\rho) = \frac{\sigma_a^2}{\kappa^2}\Big\langle \big(J\mathcal{Y}(\mathbf{w}) - \sum_B m_B^\infty(\rho) J_B(\mathbf{w})\big) J_A(\mathbf{w}) \Big\rangle_{p_{\infty,\rho}}, \quad \forall A. \tag{35}$$

$$0 = \frac{1}{2}\mathrm{tr}\big[(A - (Ay)(Ay)^\top)\partial_{\rho_j} K_\rho\big] + \beta_0 - \frac{\alpha_0 - 1}{\rho_j}, \qquad A = (K_\rho + \tau I)^{-1} \tag{36}$$

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

$$m^\infty(\rho) = K_\rho (K_\rho + \tau I)^{-1} y, \quad \tau = \kappa^2/\sigma_a^2 \tag{37}$$

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

$$m^\infty(\rho) = K_\rho (K_\rho + \tau I)^{-1} y, \quad K_\rho \chi_A = \lambda_A(\rho)\chi_A \;\Rightarrow\; m_A^\infty(\rho) = \frac{\lambda_A(\rho)}{\lambda_A(\rho) + \tau}\, y_A. \tag{38}$$

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.

C.9. 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 7 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 [28], the fundamental mechanism of discovering the relevant input subspace via IFS remains the same.

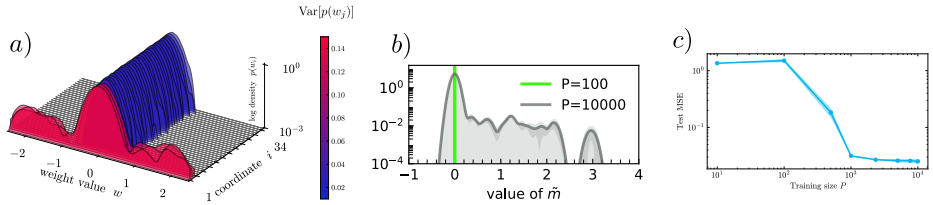

Figure 7: Similar to the ReLU results in the main text, SGLD training with Sigmoid activations leads to strong coordinate-wise symmetry breaking (high variance on relevant features $j \in S$). This demonstrates that IFS is a fundamental mechanism for subspace discovery, distinct from the neuron-level coding schemes discussed in [28].

Appendix D. Additional figures

D.1. Empirical evidence for IFS

An ensemble of trained NNs In Figure 8, 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.

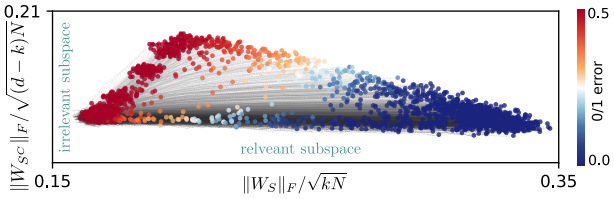

Figure 8: Each point corresponds to a one-hidden-layer $N = 512$ network trained with SGLD on k-sparse parity, with training-set size $P \in [10, 20000]$ and noise level $\kappa_0 \in [5 \cdot 10^{-5}, 5 \cdot 10^{-2}]$. The x-axis shows the normalized Frobenius norm of the first-layer weights restricted 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.

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: $\ell_{\text{iso}}(\boldsymbol{W}) = \lambda_h \sum_{i=1}^{N} \text{Var}_{j\in[d]}\left(|w_{ji}|\right)$. 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., $|w_{1i}| \approx |w_{2i}| \approx \dots$), effectively "switching off" the IFS mechanism while allowing other forms of coordination between neurons (including those discussed in [28]). Figure 9 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.

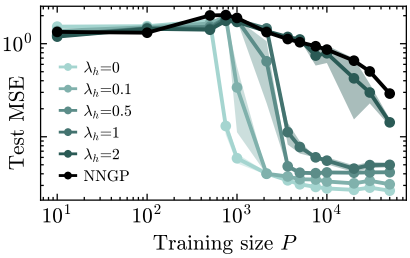

Figure 9: Test loss for k-sparse parity as a function of the regularisation λ_h that forces coordinates within a neuron to be homogeneous (preventing IFS). Increasing λ_h reduces performance till the NN performance degrades back to the Kernel (NNGP) limit. This confirms that coordinate-wise anisotropy is the necessary condition for sample-efficient learning in this setting.

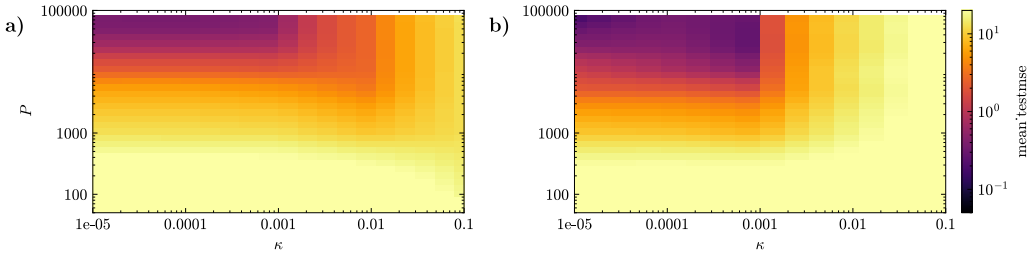

Figure 10: Same plot as Figure 4 **a)** SGLD, **b)** MF-ARD, but for a single-index model with Gaussian inputs $\boldsymbol{x} \sim \mathcal{N}(0, \mathbb{1})$ and teacher $y = \text{He}_p(\boldsymbol{w}^\top \boldsymbol{x})$, where $\boldsymbol{w}_j = 1/\sqrt{k}$ on a k-sized support and 0 otherwise, here $d = 18$, $k = 2$, $p = 4$.

Appendix E. 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 and Sompolinsky [28]. 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.

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

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

- **(Neuron-wise Sparsification):** [28] 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 *inter-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 [28]) typically assume that the weight distribution is isotropic over input coordinates unless the prior enforces otherwise. Consequently, while [28] 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.

E.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:

$$\ell_{\text{total}} = \ell_{\text{MSE}} + \lambda_h \sum_{j=1}^{N} \left[\frac{1}{d} \sum_{i=1}^{d} \left(|w_{ij}| - \left(\frac{1}{d} \sum_{k=1}^{d} |w_{kj}| \right) \right)^2 \right]. \tag{39}$$

Result: As λ_h increases, the learning curve of the neural network reverts to the NNGP learning curve (see Figure 9). 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 [28].

Appendix F. Proofs

F.1. Kernel limit

Theorem 5 *With $\gamma = 1/2$ the infinite width limit has the following FP equations:*

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2}\left(\mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_{\mathcal{Y}}(\boldsymbol{w})] - \sum_B \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w})J_B(\boldsymbol{w})]m_B^\infty\right). \tag{40}$$

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:

$$\mathcal{F}(\{m_A\}) = \frac{1}{2\kappa^2}\sum_A (y_A - m_A)^2 - N \ln \mathcal{Z}_1(\{m_A\}) \tag{41}$$

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:

$$\bar{S}_{\text{eff}}^{(\infty)}(\boldsymbol{w};\{m_A\}) = \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\ln\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right) - \frac{\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_A m_A J_A(\boldsymbol{w})\right)^2}{2\kappa^4 N^{2\gamma}\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right)} \tag{42}$$

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

$$m_A = N^{1-\gamma}\left\langle \mu(\boldsymbol{w})J_A(\boldsymbol{w})\right\rangle_{p(\boldsymbol{w}|\{m_B\})} \tag{43}$$

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:

$$\mu(\boldsymbol{w}) = \frac{\frac{1}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B J_B(\boldsymbol{w})\right)}{\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}} \tag{44}$$

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:

$$\frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} = \frac{\Sigma(\boldsymbol{w})}{N\kappa^2} \xrightarrow{N\to\infty} 0 \tag{45}$$

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:

$$p(\boldsymbol{w}|\{m_A\}) \xrightarrow{N\to\infty} p_\infty(\boldsymbol{w}) \propto \exp\left(-\frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2\right) \tag{46}$$

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

$$\mu(\boldsymbol{w}) \xrightarrow{N\to\infty} \frac{\sigma_a^2}{\kappa^2 N^{1/2}} \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B J_B(\boldsymbol{w}) \right) \tag{47}$$

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:

$$m_A^\infty = N^{1-1/2} \left\langle \left[\frac{\sigma_a^2}{\kappa^2 N^{1/2}} \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B^\infty J_B(\boldsymbol{w}) \right) \right] J_A(\boldsymbol{w}) \right\rangle_{p_\infty(\boldsymbol{w})} \tag{48}$$

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2} \left\langle \left(J_{\mathcal{Y}}(\boldsymbol{w}) - \sum_B m_B^\infty J_B(\boldsymbol{w}) \right) J_A(\boldsymbol{w}) \right\rangle_{p_\infty(\boldsymbol{w})} \tag{49}$$

$$m_A^\infty = \frac{\sigma_a^2}{\kappa^2} \left(\mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_{\mathcal{Y}}(\boldsymbol{w})] - \sum_B \mathbb{E}_{\boldsymbol{w}\sim p_\infty}[J_A(\boldsymbol{w}) J_B(\boldsymbol{w})] m_B^\infty \right) \tag{50}$$

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^∞:

$$m^\infty = \frac{\sigma_a^2}{\kappa^2}(\Xi - K m^\infty) = \frac{\sigma_a^2}{\kappa^2}(Ky - K m^\infty) \tag{51}$$

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

$$\left(I + \frac{\sigma_a^2}{\kappa^2} K \right) m^\infty = \frac{\sigma_a^2}{\kappa^2} K y \tag{52}$$

$$\left(\frac{\kappa^2}{2\sigma_a^2} I + \frac{1}{2} K \right) m^\infty = \frac{1}{2} K y \tag{53}$$

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:

$$m^\infty = K(K + \tau I)^{-1} y \tag{54}$$

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. ■

F.2. Integrating out a

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

Theorem 6 *The distribution is given by*

$$S_{\text{MF}}(\boldsymbol{w}, \{m_A\}) = const. + \frac{1}{2} \ln \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right) - \frac{\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w}) \right)^2}{2N^{2\gamma}\kappa^4 \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2} \right)} \tag{55}$$

$$+ \frac{d}{2\sigma_w^2} \|\boldsymbol{w}\|^2 \tag{56}$$

$$p(a|\boldsymbol{w}) = \mathcal{N}(\mu(\boldsymbol{w}), \sigma^2(\boldsymbol{w})) \tag{57}$$

$$\sigma(\boldsymbol{w})^2 = \frac{1}{2\alpha} = \left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right)^{-1} \tag{58}$$

$$\mu(\boldsymbol{w}) = \sigma^2\beta = \frac{\beta}{2\alpha} = \frac{\frac{1}{N^\gamma\kappa^2}\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w})\right)}{\left(\frac{1}{\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{N^{2\gamma}\kappa^2}\right)} \tag{59}$$

Proof We start with the standard Gaussian integral

$$\int da d\boldsymbol{w} e^{-[\alpha\cdot a^2 - \beta\cdot a + c]} = \int d\boldsymbol{w}\sqrt{\frac{\pi}{\alpha}} e^{\frac{\beta^2}{4\alpha} - c} \tag{60}$$

We have

$$\mathcal{S}_{\text{MF}}(\boldsymbol{w}, a, \{m_A\}) = \frac{1}{2\sigma_a^2}\sum_{i=1}^N a_i^2 + \frac{d}{2\sigma_w^2}\sum_{i=1}^d \|\boldsymbol{w}_i\|^2 + \frac{a^2}{2\kappa^2 N^{2\gamma}}\Sigma(\boldsymbol{w}) \tag{61}$$

$$- \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w})\right) \tag{62}$$

For 1 neuron we get

$$\mathcal{S}_{\text{MF}}(\boldsymbol{w}, a, \{m_A\}) = a^2\left(\frac{1}{2\sigma_a^2} + \frac{\Sigma(\boldsymbol{w})}{2\kappa^2 N^{2\gamma}}\right) - \frac{a}{\kappa^2 N^\gamma}\left(J_{\mathcal{Y}}(\mathbf{w}) - \sum_A m_A J_A(\mathbf{w})\right) + \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \tag{63}$$

with

$$\alpha := \frac{1}{2\sigma_a^2} + \frac{\Sigma(w)}{2N^{2\gamma}\kappa^2}, \tag{64}$$

$$\beta := \frac{J_{\mathcal{Y}}(w) - \sum_A m_A J_A(w)}{N^\gamma\kappa^2}. \tag{65}$$

$$c = \frac{d}{2\sigma_w^2}\|\boldsymbol{w}\|^2 \tag{66}$$

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

$$\int d\boldsymbol{w}\sqrt{\frac{\pi}{\alpha}} e^{\frac{\beta^2}{4\alpha} - c} = \int d\boldsymbol{w} e^{-[\frac{1}{2}\ln(\pi) - \frac{1}{2}\ln(\alpha) + \frac{\beta^2}{4\alpha} - c]} \tag{67}$$

where the exponent is identified as the effective action. ■

F.3. MF- FP equation

Lemma 7 *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*

$$m_S \;=\; N^{1-\gamma}\Big\langle \mu(\boldsymbol{w})\, J_S(\boldsymbol{w})\Big\rangle \;=\; \frac{N^{1-2\gamma}}{\kappa^2}\,(1-m_S)\,\left\langle \frac{J_S(\boldsymbol{w})^2}{\sigma_a^{-2}+\dfrac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}}\right\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)}. \tag{68}$$

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

$$m_S \;\approx\; (1-m_S)\,N\,\Big\langle \frac{J_S(\boldsymbol{w})^2}{\Sigma(\boldsymbol{w})}\Big\rangle_{\boldsymbol{w}\sim p(\boldsymbol{w}|m_S)} \tag{69}$$

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

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

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

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

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 $\boldsymbol{w}$. ■

F.4. Solution to the fixed point equation

Lemma 8 *Consider $\phi =$ 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*

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} R_k(\boldsymbol{w}) \tag{70}$$

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 9 *Consider $\phi =$ 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 8. 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 $\mathbb{E}[\phi(\mathbf{w}^\top \mathbf{x}) \chi_S(\mathbf{x})]$.

$$J_S(\mathbf{w}) = \mathbb{E}\left[\alpha \cdot [k - 2r]_+ \cdot (-1)^r\right] \tag{71}$$

$$= \alpha \sum_{r=0}^{k} P(r) \cdot [k - 2r]_+ \cdot (-1)^r \tag{72}$$

$$= \alpha \cdot 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)(-1)^r \tag{73}$$

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 $\mathbb{E}[\phi(\mathbf{w}^\top \mathbf{x})^2]$.

$$\Sigma(\mathbf{w}) = \mathbb{E}\left[(\alpha \cdot [k - 2r]_+)^2\right] \tag{74}$$

$$= \alpha^2 \sum_{r=0}^{k} P(r) \cdot [k - 2r]_+^2 \tag{75}$$

$$= \alpha^2 \cdot 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)^2 \tag{76}$$

We define the scale-independent constants D_k and C_k:

$$D_k := 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)(-1)^r \tag{77}$$

$$C_k := 2^{-k} \sum_{r=0}^{\lfloor (k-1)/2 \rfloor} \binom{k}{r} (k - 2r)^2 \tag{78}$$

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.

■

F.5. Exact solution of the FP equation

To evaluate this expectation analytically, we employ a saddle-point (Laplace) approximation, a standard technique in the statistical mechanics of learning. Specifically, we assume that in the relevant noise regime, the posterior distribution of the weights concentrates sharply around its optimal direction $\boldsymbol{w}^*$ (the mode). While a rigorous proof of posterior contraction in this specific high-dimensional, non-convex setting remains a challenging open problem for future work, this ansatz allows us to extract the leading-order macroscopic behavior. Under this saddle-point approximation, the expectation $\langle\cdot\rangle_{\boldsymbol{w}|m_S}$ reduces to an evaluation at the optimal weight scale $\alpha^\star$.

Theorem 10 *Using the setup from Theorem 9, espeically the proposed $\boldsymbol{w}^*$, the FP is given by*

$$m_S = (1-m_S)NR_k\left(1-\frac{\sigma_a^{-2}}{A^\star(m_S)}\right), \quad A^\star(m_S) = \frac{-C_k + \sqrt{C_k^2 + 4\left(\frac{dk}{\sigma_w^2}\right)N^{2\gamma}(1-m_S)^2 D_k^2\,\sigma_a^{-2}}}{2\left(\frac{dk}{\sigma_w^2}\right)\kappa^2 N^{2\gamma}} \tag{79}$$

Proof From Theorem 7, 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 9, 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 $A^\star(m_S)$:

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

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

$$m_S = (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{J_S(\alpha^\star)^2}{A^\star} \tag{80}$$

$$= (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{(\alpha^\star)^2 D_k^2}{A^\star} \tag{81}$$

$$= (1-m_S)\frac{N^{1-2\gamma}}{\kappa^2}\frac{1}{A^\star}\left[\frac{\kappa^2 N^{2\gamma}}{C_k}(A^\star - \sigma_a^{-2})\right] D_k^2 \tag{82}$$

$$= (1-m_S)N\frac{D_k^2}{C_k}\frac{A^\star - \sigma_a^{-2}}{A^\star} \tag{83}$$

$$= (1-m_S)NR_k\left(1 - \frac{\sigma_a^{-2}}{A^\star(m_S)}\right) \tag{84}$$

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

Theorem 11 *Consider the fixed–point equation in the infinite P-limit*

$$m_S \;=\; (1-m_S)\,NR_k\left(1 - \frac{\sigma_a^{-2}}{A^\star(m_S)}\right), \tag{85}$$

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

$$\left(\tfrac{dk}{\sigma_w^2}\right)\kappa^2 N^{2\gamma}\left(A^\star\right)^2 \;+\; C_k\,A^\star \;-\; (1-m_S)^2 D_k^2\,\sigma_a^{-2} \;=\; 0, \tag{86}$$

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

$$\kappa_c^2 \;=\; \frac{\sqrt{C_k^2 + 4\,C\,N^{2\gamma}D_k^2\sigma_a^{-2}} \;-\; C_k}{2\,C\,\sigma_a^{-2}\,N^{2\gamma}} \tag{87}$$

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 (86) to get

$$C\,\kappa^2 N^{2\gamma}\,(A^\star)^2 + C_k A^\star - D_k^2\sigma_a^{-2} = 0. \tag{88}$$

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

$$NR_k\left(1 - \frac{\sigma_a^{-2}}{A^\star(0)}\right) = 0 \quad\Longleftrightarrow\quad A^\star(0) = \sigma_a^{-2}. \tag{89}$$

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

$$C\,\kappa_c^2 N^{2\gamma}\sigma_a^{-4} + C_k\sigma_a^{-2} - D_k^2\sigma_a^{-2} = 0, \tag{90}$$

which, after rearrangement, yields

$$\kappa_c^2 = \frac{\sqrt{C_k^2 + 4\,C\,N^{2\gamma}D_k^2\sigma_a^{-2}} \;-\; C_k}{2\,C\,\sigma_a^{-2}\,N^{2\gamma}}, \tag{91}$$

the stated expression.

For $\kappa^2 > \kappa_c^2$ the quadratic gives $A^\star(0) \le \sigma_a^{-2}$, hence $1 - \sigma_a^{-2}/A^\star(0) \le 0$ 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. ■

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

$$v_j := \langle w_j^2\rangle_p, \qquad v_j'(\rho) := \mathbb{E}_{p_\rho}[w_j^2] \tag{92}$$

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

$$\rho_j(v_j) \;=\; \frac{\alpha_0 + \frac{N}{2}}{\frac{\alpha_0}{d} + \frac{N}{2}\,v_j} \;=:\; \frac{A}{B(v_j)}, \qquad A := \alpha_0 + \frac{N}{2},\;\; B(v) := \frac{\alpha_0}{d} + \frac{N}{2}v. \tag{93}$$

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

$$\min_{j\in S} v_j^{t_0} \;-\; \max_{j\notin S} v_j^{t_0} \;\ge\; c\,\varepsilon_0, \qquad v_j^t := \langle w_j^2\rangle_p \text{ at outer time } t. \tag{94}$$

We need to establish the following global bound.

Lemma 12 *Fix $j \in [d]$ and w_{-j}. Assume $\|x\|_\infty \le 1$ (e.g. $x \in \{\pm 1\}^d$) and $\phi(z) = \max(0, z)$, as well as $m_S \le 1$ Let*

$$g(\boldsymbol{w}) = \frac{1}{2}\ln\Big(\sigma_a^{-2} + \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}\Big) - \frac{\big(J\mathcal{Y}(\boldsymbol{w}) - m_S J_S(\boldsymbol{w})\big)^2}{2\kappa^4 N^{2\gamma}\big(\sigma_a^{-2} + \frac{\Sigma(\boldsymbol{w})}{\kappa^2 N^{2\gamma}}\big)}. \tag{95}$$

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

$$\big|\partial_t^2 g(w_{-j}, t)\big| \;\le\; L_\star \qquad \textit{for a.e. } t \in \mathbb{R}. \tag{96}$$

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

$$g(w_{-j}, t) \;\ge\; g(w_{-j}, 0) - \frac{L_\star}{2}\,t^2. \tag{97}$$

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 $\|x\|_\infty \le 1$.

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,

$$\partial_t \Sigma \;=\; 2\,\mathbb{E}_x\big[z\mathbf{1}_{\{z>0\}}\,x_j\big], \qquad \partial_t^2\Sigma \;=\; 2\,\mathbb{E}_x\big[x_j^2\,\mathbf{1}_{\{z>0\}}\big] \;\le\; 2, \tag{98}$$

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

$$|\partial_t\Sigma| \;\le\; 2\big(\mathbb{E}_x[z^2\mathbf{1}_{\{z>0\}}]\big)^{1/2}\big(\mathbb{E}_x[x_j^2\mathbf{1}_{\{z>0\}}]\big)^{1/2} \;\le\; 2\sqrt{\Sigma}. \tag{99}$$

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

$$\partial_t J_A \;=\; \mathbb{E}_x[\mathbf{1}_{\{z>0\}}\,x_j\,\chi_A(x)]\,, \quad \partial_t^2 J_A \;=\; 0 \text{ (a.e.)}, \qquad |\partial_t J_A| \;\le\; \mathbb{E}|x_j| \;\le\; 1, \tag{100}$$

and by Cauchy–Schwarz again

$$|J_A| \;\le\; \big(\mathbb{E}_x[\phi(z)^2]\big)^{1/2}\big(\mathbb{E}_x[\chi_A(x)^2]\big)^{1/2} \;=\; \sqrt{\Sigma}. \tag{101}$$

2. Decompose g and bound each second derivative

Let

$$a := \sigma_a^{-2}, \qquad c := \frac{1}{\kappa^2 N^{2\gamma}}, \qquad J_\Delta := J_{\mathcal{Y}} - m_S J_S, \qquad q := J_\Delta^2, \tag{102}$$

so

$$g(\boldsymbol{w}) \;=\; \underbrace{\tfrac{1}{2}\log\big(a + c\Sigma\big)}_{=:g_1(\Sigma)} \;-\; \underbrace{\frac{q}{2\kappa^4 N^{2\gamma}\,(a+c\Sigma)}}_{=:g_2(\Sigma,J_\Delta)}\,. \tag{103}$$

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,

$$\partial_t^2 g_1 \;=\; h_1''(\Sigma)\,(\partial_t\Sigma)^2 \;+\; h_1'(\Sigma)\,\partial_t^2\Sigma. \tag{104}$$

Using (99) and $\partial_t^2\Sigma \le 2$,

$$\Big|h_1''(\Sigma)\,(\partial_t\Sigma)^2\Big| \;\le\; \frac{c^2}{2(a+c\Sigma)^2}\cdot 4\Sigma \;=\; 2c^2\,\frac{\Sigma}{(a+c\Sigma)^2} \;\le\; \frac{c\,\sigma_a^2}{2}, \tag{105}$$

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

$$\big|\partial_t^2 g_1\big| \;\le\; \frac{3}{2}\,c\,\sigma_a^2\,. \tag{106}$$

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 (101) 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

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

- $|q''|$: 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

$$\left|\tfrac{q''}{a+c\Sigma}\right| \le \frac{2C_\Delta^2}{a}, \tag{107}$$

$$\left|\tfrac{2\,q'\,c\,\Sigma'}{(a+c\Sigma)^2}\right| \le \frac{2\cdot 2C_\Delta^2\sqrt{\Sigma}\cdot c\cdot 2\sqrt{\Sigma}}{(a+c\Sigma)^2} = 8C_\Delta^2 c\,\frac{\Sigma}{(a+c\Sigma)^2} \;\le\; \frac{2C_\Delta^2}{a}, \tag{108}$$

$$\left|\tfrac{q\,c\,\Sigma''}{(a+c\Sigma)^2}\right| \le \frac{C_\Delta^2\Sigma\cdot c\cdot 2}{(a+c\Sigma)^2} \;\le\; \frac{C_\Delta^2}{a}, \tag{109}$$

$$\left|\tfrac{2\,q\,(c\Sigma')^2}{(a+c\Sigma)^3}\right| \le \frac{2\,C_\Delta^2\Sigma\cdot c^2\cdot 4\Sigma}{(a+c\Sigma)^3} = 8C_\Delta^2 c^2\,\frac{\Sigma^2}{(a+c\Sigma)^3} \;\le\; \frac{32C_\Delta^2}{27\,a}, \tag{110}$$

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 $\left|\partial_t^2 g_2\right| \le \alpha\,K_2$ 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 (106) and the bound for $\partial_t^2 g_2$, there exists

$$L_\star \;:=\; \frac{3}{2}\,c\,\sigma_a^2 \;+\; \alpha\,K_2 \tag{111}$$

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

4: Standard taylor expansion

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

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

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

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

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 $g(w_{-j}, t) \ge g(w_{-j}, 0) - \frac{L_\star}{2}t^2$. ■

F.6. Proof of Theorem 2

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

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

with g as in Lemma 12. 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:

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

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 **??** therefore yields

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

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

On-support. For $j \in S$, the same curvature bound $|\partial_t^2 g(w_{-j}, t)| \le L_\star$ implies the 1D conditional potential $U_j(t) = \frac{\rho}{2}t^2 + g(w_{-j}, t)$ is $(\rho - L_\star)$-strongly convex, hence $(w_j \mid w_{-j}) \le (\rho - L_\star)^{-1}$. Moreover $|\partial_t g(w_{-j}, 0)|$ is $O(1)$ uniformly in w_{-j} (by the same bounds used in Lemma 12), so the conditional mean satisfies $|\mathbb{E}[w_j \mid w_{-j}]| = O(\rho^{-1})$, and therefore

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

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

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

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

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

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

■

F.7. 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 $\rho_j \equiv \rho_{\mathrm{eff},j}$.

■

Appendix G. Algorithms

G.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,

$$f(\boldsymbol{x}) = s_f \sum_{b=1}^{B} a_b\, \phi(\boldsymbol{w}_b^\top \boldsymbol{x}), \qquad s_f = \frac{N^{1-\gamma}}{B}. \tag{112}$$

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 C.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 $z_{pb} = \boldsymbol{x}_p^\top \boldsymbol{w}_b$:

$$C_{1,b} = \sum_{p=1}^{P} \phi(z_{pb})\, r_p, \qquad C_{2,b} = \sum_{p=1}^{P} \phi(z_{pb})^2, \tag{113}$$

$$G_b^{\text{data}} = -\frac{2}{P} \sum_{p=1}^{P} \Big(r_p - s_f a_b\, \phi(z_{pb})\Big)\, \phi'(z_{pb})\, \big(s_f a_b\big)\, x_p \in \mathbb{R}^d. \tag{114}$$

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:

$$E_{\text{prior}} = \tfrac{1}{2} \sum_{b=1}^{B} \rho^\top (\boldsymbol{w}_b \odot \boldsymbol{w}_b) + \tfrac{1}{2\sigma_a^2} \sum_{b=1}^{B} a_b^2, \quad \mathcal{L}_{\text{data}} = \frac{1}{P} \sum_{p=1}^{P} (y_p - f_p)^2, \tag{115}$$

and update parameters by Langevin dynamics on the potential

$$U(W, a) \;=\; T\, E_{\text{prior}} + \mathcal{L}_{\text{data}}. \tag{116}$$

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

$$\nabla_{a_b} U = \frac{T}{\sigma_a^2}\, a_b \;-\; \frac{2\, s_f}{P}\, C_{1,b} \;+\; \frac{2\, s_f^2}{P}\, C_{2,b}\, a_b, \qquad \nabla_{w_b} U = G_b^{\text{data}} \;+\; T\,(\rho \odot \boldsymbol{w}_b). \tag{117}$$

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

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

$$w_b \leftarrow w_b - \eta\, \nabla_{w_b} U + \sqrt{2T\eta}\, \xi_{w_b}, \qquad a_b \leftarrow a_b - \eta\, \nabla_{a_b} U + \sqrt{2T\eta}\, \xi_{a_b}, \tag{118}$$

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.

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 $f(x) = s_f \sum_b a_b \phi(w_b^\top x)$

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 $T \leftarrow 2\kappa^2$ **for** $t = 1..T_{out}$ **do**

end

compute $f_p = s_f \sum_b a_b \phi(x_p^\top w_b)$ and residuals $r_p = y_p - f_p$ **for** $k = 1..K$ **do**

end

compute $\{C_{1,b}, C_{2,b}, G_b^{\text{data}}\}$ via formulas above

Form $\nabla_{a_b} U, \nabla_{w_b} U$; update $w_b \leftarrow w_b - \eta \nabla_{w_b} U + \sqrt{2T\eta}\, \xi_{w_b}$, $a_b \leftarrow a_b - \eta \nabla_{a_b} U + \sqrt{2T\eta}\, \xi_{a_b}$

refresh f_p, r_p after the last inner step

ARD update ρ: $\alpha_{\text{post}} = \alpha_0 + \frac{B}{2}$, $\beta_{\text{post}} = \beta_0 + \frac{1}{2} \sum_b \|w_b\|^2$, $\rho \leftarrow (1-\lambda)\rho + \lambda\, \alpha_{\text{post}}/\beta_{\text{post}}$

ARD update The ARD update is:

$$\alpha_{\text{post}} = \alpha_0 + \tfrac{B}{2}, \qquad \beta_{\text{post}} = \beta_0 + \tfrac{1}{2} \sum_{b=1}^{B} \|w_b\|_2^2, \qquad \rho \leftarrow (1-\lambda)\, \rho + \lambda\, \frac{\alpha_{\text{post}}}{\beta_{\text{post}}}. \tag{119}$$

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.

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

$$g \;=\; \frac{1}{P_{\text{eval}}}\, c^\top c, \tag{120}$$

and the (scalar) empirical overlap

$$m_S \;=\; \frac{1}{P_{\text{eval}}} \sum_{\mu=1}^{P_{\text{eval}}} \chi_S(\boldsymbol{x}_\mu)\, f(x_\mu) \;=\; \frac{1}{P_{\text{eval}}}\, c^\top f. \tag{121}$$

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

$$\widehat{\mathcal{E}}_{\text{test}} = \frac{1}{2P_{\text{eval}}} \sum_{\mu=1}^{P_{\text{eval}}} \big(f(\boldsymbol{x}_\mu) - \chi_S(\boldsymbol{x}_\mu)\big)^2 = \underbrace{\tfrac{1}{2}(1 - m_S)^2}_{\text{mode term}} + \underbrace{\tfrac{1}{2}\big(\bar{f}^2 - 2\, m_S^2 + g\, m_S^2\big)}_{\text{noise / orthogonal term}} \tag{122}$$

$$= \frac{1}{2}\Big(\bar{f}^2 - 2\, m_S + g\Big), \tag{123}$$

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

Appendix H. Training details (hyperparameters)

Here, we present the hyperparameters for SGLD and MF-ARD for the different figures. The hyperparameters for Figure 4 are specified below.

- Data in Figure 1 is a slice of Figure 4 for $\kappa = 5 \cdot 10^{-3}$.
- Data in Figure 6 is a slice of Figure 4 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	$\{5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 7.5\times10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}\}$
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)	5×10^{-4}
start LR η_{start}	5×10^{-3}
LR scheduler	polynomial (power 2): $\eta_{\text{start}} \to \eta$ over 2×10^6 steps

Table 2: Algorithm outlined in Section C.1. Hyperparameters for Figure 4 **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: $5 \times 10^{-3} \to 5{\times}10^{-4}$ 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 3: Algorithm outlined in Section G.1. Hyperparameters for Figure 4 **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 4: Algorithm outlined in Section C.1. Hyperparameters for Figure 10 **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: $2 \times 10^{-3} \to 5 \times 10^{-4}$ 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 5: Algorithm outlined in Section G.1. Hyperparameters for Figure 10 **b)** with ARD enabled.